ISSN : 2165-4069(Online)
ISSN : 2165-4050(Print)

IJARAI

International Journal of Advanced Research in Artificial Intelligence

Volume 5 Issue 6

www.ijarai.thesai.org

A Publication of
The Science and Information Organization

Editorial Preface

From the Desk of Managing Editor…

Artificial Intelligence is hardly a new idea. Human likenesses, with the ability to act as human, dates back to Geek mythology with Pygmalion's ivory statue or the bronze robot of Hephaestus. However, with innovations in the technological world, AI is undergoing a renaissance that is giving way to new channels of creativity.

The study and pursuit of creating artificial intelligence is more than designing a system that can beat grand masters at chess or win endless rounds of Jeopardy!. Instead, the journey of discovery has more real-life applications than could be expected. While it may seem like it is out of a science fiction novel, work in the field of AI can be used to perfect face recognition software or be used to design a fully functioning neural network.

At the International Journal of Advanced Research in Artificial Intelligence, we strive to disseminate proposals for new ways of looking at problems related to AI. This includes being able to provide demonstrations of effectiveness in this field. We also look for papers that have real-life applications complete with descriptions of scenarios, solutions, and in-depth evaluations of the techniques being utilized.

Our mission is to be one of the most respected publications in the field and engage in the ubiquitous spread of knowledge with effectiveness to a wide audience. It is why all of articles are open access and available view at any time.

IJARAI strives to include articles of both research and innovative applications of AI from all over the world. It is our goal to bring together researchers, professors, and students to share ideas, problems, and solution relating to artificial intelligence and application with its convergence strategies. We would like to express our gratitude to all authors, whose research results have been published in our journal, as well as our referees for their in-depth evaluations.

We hope that this journal will inspire and educate. For those who may be enticed to submit papers, thank you for sharing your wisdom.

Editor-in-Chief
IJARAI
Volume 5 Issue 6 June 2016
ISSN: 2165-4069(Online)
ISSN: 2165-4050(Print)

Editorial Board

Reviewer Board Members

CONTENTS

A New Technique to Manage Big Bioinformatics Data Using Genetic Algorithms

Huda Jalil Dikhil
Dept. of Comoputer Sciecne
Applied Science Private University
Amman, Jordan

Mohammad Shkoukani
Dept. of Comoputer Sciecne
Applied Science Private University
Amman, Jordan

Suhail Sami Owais
Dept. of Comoputer Sciecne
Applied Science Private University
Amman, Jordan

***Abstract*—The continuous growth of data, mainly the medical data at laboratories becomes very complex to use and to manage by using traditional ways. So, the researchers started studying genetic information field in bioinformatics domain (the computer science field, genetic biology field, and DNA) which has increased in past thirty years. This growth of data is known as big bioinformatics data. Thus, efficient algorithms such as Genetic Algorithms are needed to deal with this big and vast amount of bioinformatics data in genetic laboratories. So the researchers proposed two models to manage the big bioinformatics data in addition to the traditional model. The first model by applying Genetic Algorithms before MapReduce, the second model by applying Genetic Algorithms after the MapReduce, and the original or the traditional model by applying only MapReduce without using Genetic Algorithms. The three models were implemented and evaluated using big bioinformatics data collected from the Duchenne Muscular Dystrophy (DMD) disorder. The researchers conclude that the second model is the best one among the three models in reducing the size of the data, in execution time, and in addition to the ability to manage and summarize big bioinformatics data. Finally by comparing the percentage errors of the second model with the first model and the traditional model, the researchers obtained the following results 1.136%, 10.227%, and 11.363%, respectively. So the second model is the most accurate model with less percentage error.**

***Keywords*—Bioinformatics; Big Data; Genetic Algorithms; Hadoop MapReduce**

I. INTRODUCTION

The important evaluation of the Bioinformatics and genetics field in the recent years has helped scientists and doctors to understand illnesses and diagnose it the better way and discover the reasons behind many diseases and genetic mutations, including muscular degeneration, which causes disability of many children around the world. To diagnose genetic diseases at medical laboratories, it requires a comparison procedure between the defective genes with the natural ones by alignment and matching sequence of Nucleated (nitrogenous bases) in the genes through National Center for Biotechnology Information (NCBI), which consider as the largest database and repository of genes.

Processing medical data due to the large size of bioinformatics data is hard to manage and it is not easy to reduce the size of needed data. For this and other reasons, it becomes important to develop such models and algorithms that can manage big bioinformatics data that are produced by genetic laboratories, and have the ability to find the defective gene in less time with less error because medical application requires high accuracy.

So, for managing big bioinformatics data, the authors proposed two new models. The original model used only the Hadoop MapReduce. Since Genetic Algorithms GAs have many benefits especially in optimization problems, the authors tried to propose two new models by applying Genetic Algorithms before and after MapReduce. So, the first model was by applying Genetic Algorithms before MapReduce, and the second model was by implementing Genetic Algorithms after the MapReduce.

The paper consists of eight sections. The first section is an introduction. The second Section discusses the Big Data, its characteristics, and the architectures. The third section demonstrates the Bioinformatics. The fourth section explains the Genetic Algorithms and its features. The fifth section presents the problem statement of the research. The sixth part discussed the two proposed models. Section seven explains the data description. Part eight explored the results of the proposed models.

II. BIG DATA

Big Data is a term used to describe the enormous and massive amounts of data that could not be handled and processed using traditional methods. Big Data size has increased conspicuously in various fields over the past twenty years, where the volume of the generated and duplicated data has grown more than ten times in the over years, which cannot be predicted because data continuously increased to be double every two years. Data in Big Data are structured and unstructured. Thus, it needs more complicated tools rather than the traditional ones to be analyzed and managed. Managing this data brings more challenges and requires more efficient methods [1, 2].

Managing Big Data is one of the main challenges that faces large corporations, and has attracted the interest of researchers in the past years [1, 2]. Big Data has several characteristics known by *nV's* characteristics, and it has several type of architectures for Big Data analysis.

A. Big Data Characteristics

There are several characteristics (5Vs) that distinguish Big Data from standard set of data: Volume, Variety, and Velocity,

Value, and Veracity [14, 15]. The 5Vs characteristics of the Big Data were illustrated in Fig. 1.

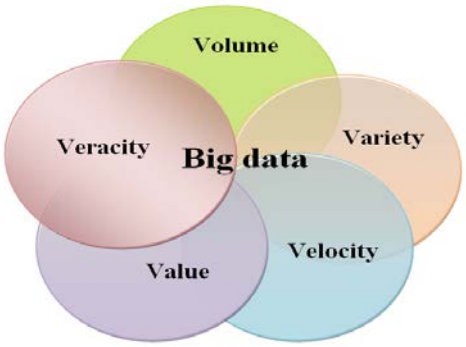


Fig. 1. The 5Vs Characteristics of the Big Data

Some other researchers have different views and consider only 3Vs (Volume, Velocity, and Variety) as fundamental features. And others add Veracity as 4Vs [1, 2, 3, 4].

B. Architectures of Big Data Analytics

On the different type of sources and different structures. There are three main types of architectures for Big Data analysis, MapReduce architectures, fault-tolerant graph architectures and streaming graph architectures. Some of the characteristics of these styles as shown in Table 1 in terms of the used memory type: if the used memory is local memory or shared global memory, in addition to the fault tolerance [2].

III. BIOINFORMATICS

Bioinformatics is relatively an old field, it started before more than a century and introduced by the Austrian scientist Gregor Mendel, who known as the "father of genetics." Since then, the understanding of genetic information has increased, especially in past thirty years. The researches and studies in the domain of bioinformatics led to the creation of the largest international organization (HUGO), the first international organization that published the first complete map of the genome of sustainable in bacteria. Bioinformatics is the relationship between computer science and biology [5, 6, 7].

TABLE I. CHARACTERISTICS OF DIFFERENT TYPES OF ARCHITECTURES FOR BIG DATA ANALYTICS

	Architectures		
Characteristics	**MapReduce architecture**	**Fault tolerant graph architecture**	**Streaming graph architecture**
Memory	Local memory Global memory	Global Memory	Data not need to be stored into disks
Fault Tolerance Allow	Allow	Allow	Not
Operations Synchronization	Synchronization	Synchronization	Asynchronous

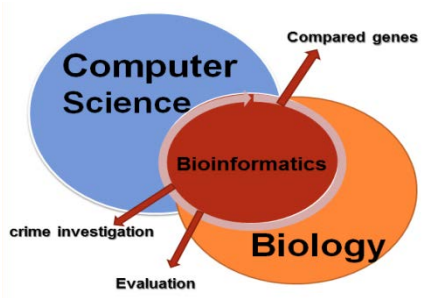


Fig. 2. Bioinformatics in general [6, 7, 8]

Fig. 2 shows that the bioinformatics in general which is an intersection between the biology and the computer science. It can be used in different fields such as crime scene investigation, comparing genes, and evaluation.

Bioinformatics characteristic debate the collaborative resources that work together for such task [8].

The theory “structure prediction” as a technique to recruit computer tools and algorithms is the most important objective of bioinformatics (Molecular Bioinformatics) in addition to being an alternative method and attractive [8]. Molecular bioinformatics logic and dealing with the concepts of biology regarding molecules and the application of the "information" to understand the technology and organization associated with these biomolecules in cells and organisms information. New genes discovered by searching for systematic data available to genome sequence, so through the sequence identity algorithms are appointed the supposed new genes function [8].

IV. GENETIC ALGORITHMS

Genetic algorithm GAs is one of the most powerful computer algorithms that based on natural living genes combining, producing and inheritance acts; it has vital importance in Computer Sciences branches like Artificial Intelligent and Computer Vision techniques. Genetic algorithms are useful tools for search and optimization problems. [9, 10].

The most important characteristic of genetic algorithms is solving hard problems with an optimal solution. Fig. 3 presents the simple Genetic Algorithm flowchart which it briefly describes the four basic operators to resolve a problem as follows: fitness function, selection operator, crossover operator, and mutation operator [9, 10, 11, 12, 13, 14].

V. PROBLEM STATEMENT

There is a tight relationship between big data and bioinformatics since there is a vast data in bioinformatics especially the DNA, which each human genome sequence approximately 200 gigabytes [2].

The development of Computer Science (CS) helped other scientific fields and became a key and essential part in most biological and medical experimentations. With the continuous growth of data, especially the medical data at laboratories (lab), it becomes very hard to use this data and manage it using the traditional ways, so efficient algorithms are needed to deal with this large and vast amount of bioinformatics data in genetic laboratories, which includes a gene and protein sequence. Thus, the researchers used one of the evolutionary algorithms which are genetic algorithms.

VI. Proposed Models

Data management is a very arduous task, especially when you have an enormous amount of data such as DNA. The proposed model based on genetic algorithm and Hadoop MapReduce. The researchers presented two models, the first one (GAHMap) by applying Genetic Algorithm before Hadoop MapReduce. The second one (HMapGA) by executing Genetic Algorithm after Hadoop MapReduce.

C. Model 1 (GAHMap): GAs before Hadoop MapReduce

Fig. 4 demonstrates the stages of the first proposed model (GAHMap) as follows:

1) The input of prototype 1 is the big bioinformatics data which is denoted by (M).

2) Applying Genetic Algorithms on (M), which will produce an optimized data which is indicated by (M′).

3) Carrying out Hadoop MapReduce on (M′), the result will be the reduced data which is denoted by (M″).

D. Model 2 (HMapGA): GAs after Hadoop MapReduce.

The second paradigm (HMapGA) presented in Fig. 5 and it has the following stages:

1) The input of prototype II is the big bioinformatics data which is denoted by (M).

2) Applying Hadoop MapReduce on (M), the result will be the reduced data which is denoted by (M′).

3) Executing Genetic Algorithms (M′), which will produce an optimized data which is indicated by (M″).

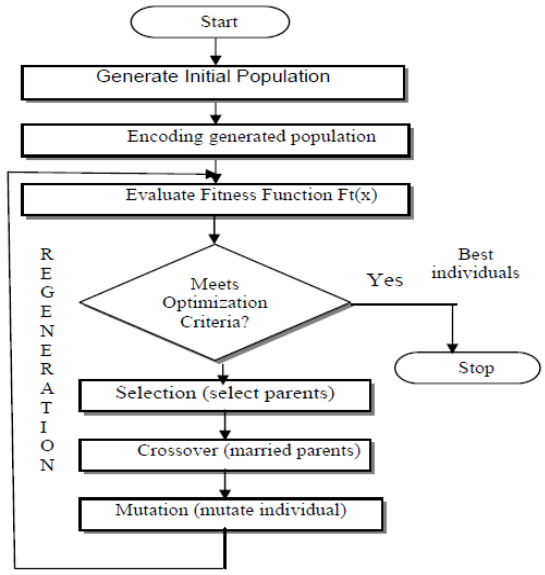


Fig. 3. Flowchart of Genetic Algorithms [9]

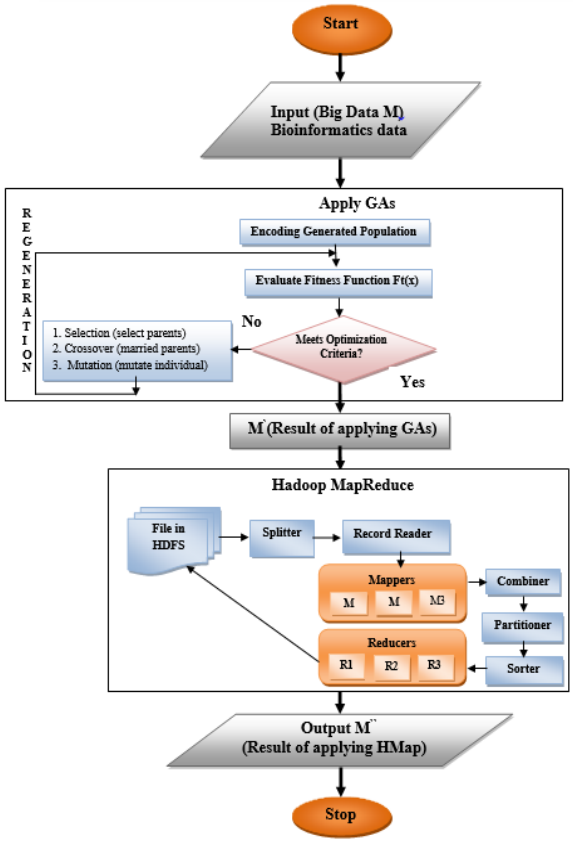


Fig. 4. Proposed Model I GAHMap stages

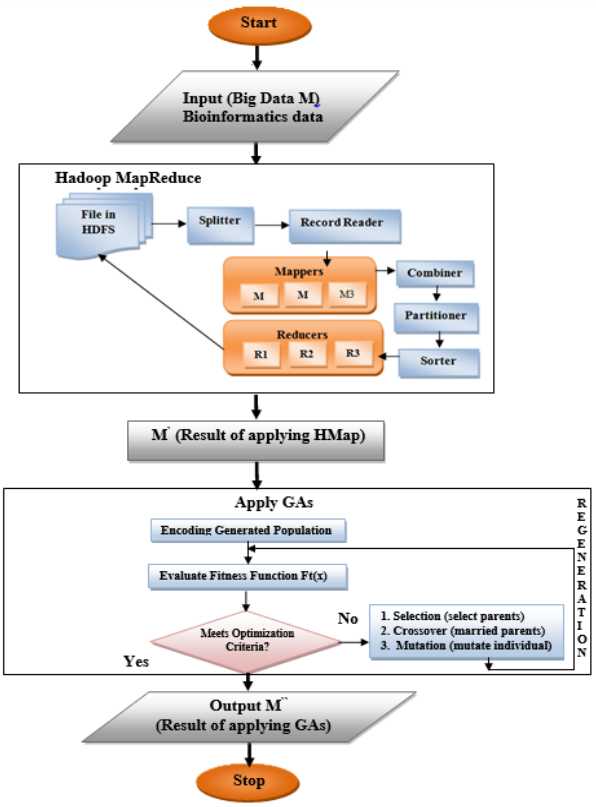


Fig. 5. Proposed Model II HMapGA Stages

The two models were fully implemented using MATLAB and applied as graphical user interface (GUI) system as shown in Fig. 6. The system mainly consists of two parts: the first part is the Create Data Server that reads the big bioinformatics data as input dataset (M). The second part consists of three options to manage the Big Data: the first alternative to execute the first model GAHMap, the second option to perform the second model HMapGA, and the third option to proceed the original model which uses Hadoop MapReduce without genetic algorithms (HMap).

The results from the first and second models will be compared to identify the model that can give the best result which reduces the size of the data with better accuracy. After that, the outcome of the chosen model will be compared with the result from the original Hadoop MapReduce. Finally, the outcomes will determine the best model among them.

VII. Data Description

The dataset of genes used in this research acquired from the Genetics Center at Specialty Hospital–Amman, Jordan. The dataset is related to Duchene Muscular Dystrophy (DMD), which is a popular and widespread genetic disease in the country as well as all over the world. It was an 88 sample from 88 individuals (genes). Each gene in the dataset represents 108 gigabytes of DNA tape; gene number 19 was obtained for this research and saved in text file format (txt file). Each gene and file contain 2,220,388 nucleotides; the nucleotides consist of a base (one of four chemicals and amino acids: cytosine, guanine, adenine, and thymine). The dataset with the 88 genes has been already diagnosed and alignment using the global location of the genes NCBI website by the genetic center to find the defective and normal genes. The results showed that 48 of the genes were defective and suffer from DMD disease, and the other 40 genes were normal. A sample of one of the genes from the DNA sequence was saved in a text file as shown in Fig. 7.

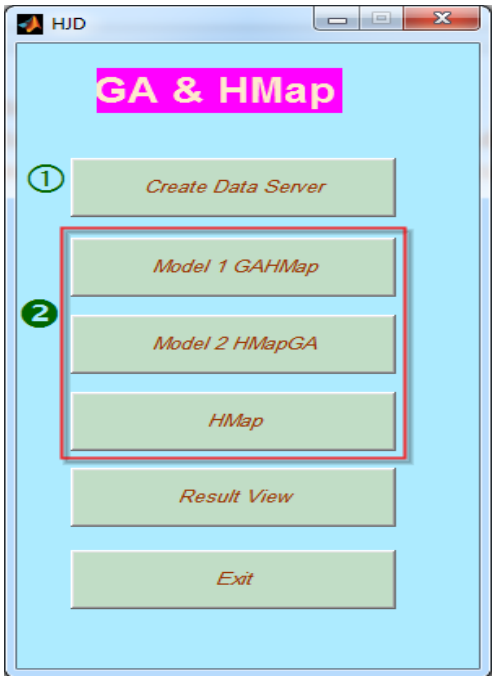


Fig. 6. Graphical User Interface of the Proposed Method

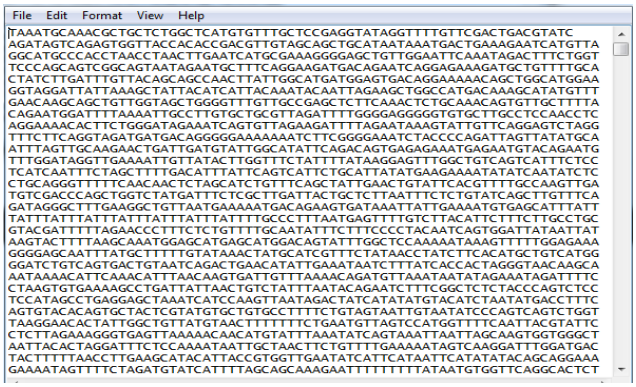


Fig. 7. Gene text file with nucleotides sequence

VIII. System Test Result

The system was tested using dataset within 88 genes (40 normal genes and 48 defective genes) as an input of Bioinformatics' Big Data (M) to find which model will be better. So, after data server reads the dataset, it was tested by applying the three models on the same dataset. During execution, each model will display the result with the following information:

- TP: True Positive, which means the number of standard genes.
- TN: True Negative, which means the number of genes which is defective.

- FP: False Positive, which means the number of genes which is regular and detected as defective.
- FN: False Negative, which implies the number of genes which is defective and detected as deranged.

Fig. 8 shows the outcome of GAHMap implementation, which concludes that there are 40 TP and zero FP which implies that there are no normal genes revealed as defective genes, and there are 39 TN and 9 FN which means that there are 9 defective genes detected as standard genes which listed in FN sequence.

Fig. 9 displays the result of second model HMapGA execution, which concludes that there are 40 TP and zero FP which means that there are no natural genes revealed as defective genes, and there are 47 TN and 1 FN which signifies that there is only one deficient gene exposed as standard genes.

Fig. 10 presents the consequence of the original model HMap enforcement, which determines that there are 40 TP and zero FP which implies that there is no normal genes appeared as defective genes, and there are 38 TN and 10 FN which means that there are 10 deficient genes detected as normal genes.

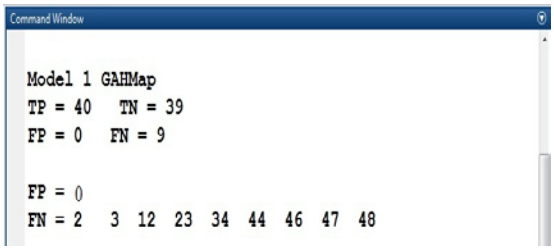


Fig. 8. GAHMap Model 1 Result

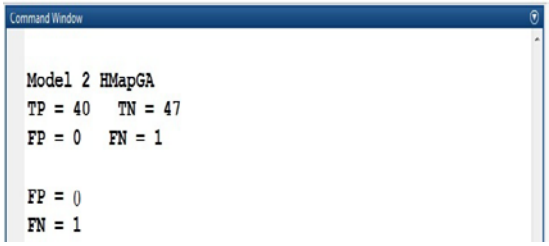


Fig. 9. HMapGA Model Results

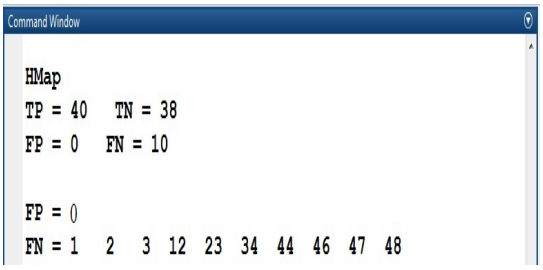


Fig. 10. HMap Model Results

Table 2 summarizes the results of the three models GAHMap, HMapGA, and HMap.

TABLE II. THREE MODELS SUMMARY RESULTS

Model	TP	TN	FP	FN
GAHMap	40	39	0	9
HMapGA	40	47	0	1
HMap	40	38	0	10

For more accuracy of the results, the percentage error (%Error) was calculated for each model by using the mathematic formula which is the difference between the experimental value and theoretical value divided by theoretical value as shown in equation (1) [15]:

$$\%Error = \frac{\left|\text{TheoreticalValue} - ExperimentalValue\right|}{\left|TheoreticalValue\right|} \times 100\% \quad (1)$$

Where, the *TheoreticalValue* means the total number of genes used in the research (natural + defective), and the *ExperimentalValue* means the total number of correctly detected genes (natural + defective) by the system.

As shown in Table 3 by applying the percentage error equation, it was found that the percentage error of the first model GAHMap =|(Total number of genes used in the research (normal + defective)) - (total number of correctly detected genes (normal + defective)|/ (total number of genes used in the thesis (normal + defective))*100

=| (40+48) – (40+39)| / (40+48) * 100%
= |88-79| / 88 * 100
= 10.227 %
For the second model HMapGA % Error =| (40+48) – (40+47)| / (40+48) * 100
= |88-87| / 88 * 100
= 1.136%
For the original model HMap % Error =| (40+48) – (40+38)| / (40+48) * 100
= |88-77| / 88 * 100
= 11.363%.

TABLE III. THREE MODELS RESULTS OF PERCENTAGE ERROR

Model	GAHMap	HMapGA	HMap
% Error	10.227%	1.136%	11.363%

According to the results in Table 3 the researchers conclude that the HMapGA is better than the GAHMap, and if the HMapGA compared with the original model HMap it found that the HMapGA is also better than the original one HMap. So the HMapGA proved to be the most accurate model with less percentage error and succeed in achieving the objectives of this research which includes organizing big bioinformatics data by matching and finding normal and defective genes with less time and less percentage error.

IX. CONCLUSION

This paper proposed two models to manage big bioinformatics data of DMD disorder. In the first model, GAHMap was implemented genetic algorithms before Hadoop MapReduce. In the second model, HMapGA has executed

genetic algorithms after Hadoop MapReduce. The proposed models in addition to the original model were tested using the real dataset of 88 genes related to DMD disorder.

By comparing the results of the three paradigms, the researchers found that the number of genes which is natural and revealed as defective (FP) was zero for all models, but the number of genes which is faulty and detected as normal (FN) were 9, 1, and 10 defective genes for the first, second, and original models respectively. The researchers conclude that the HMapGA detected less number of defective genes as natural ones.

Also, when comparing the percentage error for the three models, the second model has 1.136 % which is the most accurate model with the less percentage error.

Finally, the researchers conclude that the second model HMapGA is the best model since it succeeds in matching and finding normal and defective genes in less time and less percentage error. So HMapGA provides an efficient technique to manage and reduce the size of big bioinformatics data in the laboratory.

REFERENCE

[1] M. Chen, S. Mao, Y. Zhang, and V. C. Leung, Big data: related technologies, challenges and future prospects. Springer, 2014.

[2] H. Kashyap, H. A. Ahmed, N. Hoque, S. Roy, and D. K. Bhattacharyya, Big data analytics in bioinformatics: a machine learning perspective. Arxiv preprint arxiv:1506.05101, 2015.

[3] M. Moorthy, R. Baby, and S. Senthamaraiselvi, "An analysis for big data and its technologies", International Journal of Computer Science Engineering and Technology (IJCSET), Vol. 4, no 12, pp. 412-418, December 2014.

[4] J. Hurwitz, A. Nugent, F. Halper , M. Kaufman, Big Data For Dummies, 2013.

[5] S. M. Thampi, Introduction to Bioinformatics, LBS College of Engineering, 2009.

[6] Ralf Hofestädt, Bioinformatics: german conference on bioinformatics, GCB'96, Leipzig, Germany, Springer Science & Business Media, vol. 1278, September 30-October 2, 1996.

[7] D. C. Rubinsztein, Annual review of genomics and human genetics, 2001.

[8] P. Narayanan, Bioinformatics: A Primer. New Age International, 2006.

[9] S. S. Owais, P. Krömer, and V. Snásel, Query optimization by Genetic Algorithms. In Proceedings of the Dateso 2005 Annual International Workshop on databases, pp. 125-137, April 2005.

[10] R. Kaur, and S. Kinger, Enhanced genetic algorithm based task scheduling in cloud computing. International Journal of Computer Applications, vol. 101, no 14, pp. 1-6, 2014.

[11] S. N. Sivanandam, and S. N. Deepa, Introduction to genetic algorithms, Springer Science & Business Media, 2007.

[12] A. E. Eiben, and J. E. Smith, Introduction to evolutionary computing. Springer Science & Business Media, 2003.

[13] C. Y. Jiao, and D. G. Li, Microarray image converted database-genetic algorithm application in bioinformatics. In biomedical Engineering and Informatics, International Conference, vol. 1, pp. 302-305, May 2008.

[14] M. Chen, S. Mao, Y. Zhang V. C. Leung. Big data related technologies, challenges and future prospects. SpringerBriefs in Computer Science, New York Dordrecht London, 2014.

[15] Suhail Sami Owais, And Nada Sael Hussein. Extract Five Categories CPIVW from the 9V's Characteristics of the Big Data. International Journal of Advanced Computer Science and Applications (IJACSA), vol. 7, no. 3, pp. 254-258, 2016.

Improved Fuzzy C-Mean Algorithm for Image Segmentation

Hind Rustum Mohammed
CS dept. Faculty of Computer Science and Mathematics
University of Kufa
Najaf, Iraq

Husein Hadi Alnoamani
MS dept. Faculty of Computer Science and Mathematics
University of Kufa
Najaf, Iraq

Ali AbdulZahraa Jalil / M.Sc. Student
CS dept. Faculty of Computer Science and Mathematics
University of Kufa, Najaf, Iraq

***Abstract*—The segmentation of image is considered as a significant level in image processing system, in order to increase image processing system speed, so each stage in it must be speed reasonably. Fuzzy c-mean clustering is an iterative algorithm to find final groups of large data set such as image so that is will take more time to implementation. This paper produces an improved fuzzy c-mean algorithm that takes less time in finding cluster and used in image segmentation.**

Keywords—pattern recognition; image segmentation; fuzzy c-mean; improved fuzzy c-mean; algorithms

I. INTRODUCTION

To recognize pattern and analysis an image the main process is segmentation of image[1-3]. Is an operation of dividing an image into parts that have same features and the collection of these parts form the original image[4]. Fig.1. illustrate variant levels of processing of image and technique of analyzing [5], and it shows clearly segmentation stage.

There are many types of image's pattern recognition and segmentation, but there are two mainly types of classification which are used: Supervised classification and unsupervised classification, in the first one the classes are defined in advance and in the second they are not defined in advance which known as clustering. There are two types of clustering: hard clustering and fuzzy clustering, in hard clustering, the data item is belong exactly to one cluster but in fuzzy clustering, the data item belong by the degree of membership to each cluster of clusters, and the summation of all memberships values to one of data items is equal to one.

Fuzzy c-mean clustering is one of unsupervised clustering algorithms that is widely used in image processing and computer vision because it easy to implement and clustering performance[6], [7]. It's used to segment an image by grouping pixels that have similar or nearly similar values into a cluster, where each group of pixel's values that belong to one cluster are similar to each other and different from pixel's values that belong to other clusters, and then these clusters represent the segments of the segmented image. The traditional fuzzy c-mean suffers from some limitations, it's not accurate in the segmentation of noisy image and time consuming because it's iterative nature. Our proposed algorithm which named Improved fuzzy c-mean algorithm offers an overcoming of one limitation of traditional fuzzy c-mean which is time-consuming.

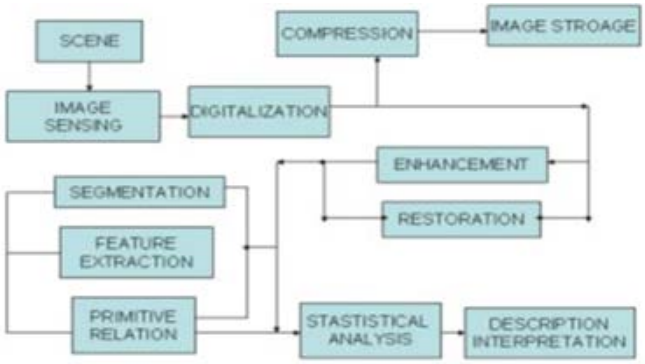


Fig. 1. Schema of variant levels of processing of image and technique of analyzing

In our proposed algorithm we use frequency of each data item of image and processing these frequencies instead of processing whole data items of the image. That is reduce processing time in the great form. This paper contains five parts and arranged as follow: Section 2 talking about time complexity, section 3 the traditional fuzzy c-mean, section 4 proposed an algorithm, section 5 Experimental Results, and in section 6 the conclusion.

II. TIME COMPLEXITY

It's time that required to run or execute an algorithm[11].the notation big O is used to express time complexity. it's proportional to the size of input data. If the input size is n, then the time complexity is the time required by the algorithm to process these input. Each algorithm has a primitive operation(s), so the time of the algorithm is determined by computing the summation of times that required to run each of these operations. It's always expressed by the prevalent term, which is the term have exponent with the highest value. It also ignores constant of multiplication and constant of the division. if the time required to accomplish an algorithm of n input size is $10n^3 + 6n$, then the expression of its time complexity is $O(n^3)$ also if the time required to accomplish an algorithm is $c*n^2$ or n^2/c where c is constant then the time complexity is $O(n^2)$. If the algorithm processed all inputted data to get the desired solution, then the time complexity called the worst-case of time complexity, which is show, the algorithm take maximum time to achieve the required process. There are many types of time complexities which depend on algorithm's function nature. Some common types of time complexities are constant time

O(1), linear time O(n), quadratic time $O(n^2)$, exponential time $O(c^n)$ where $c \geq n > 1$. In our proposed algorithm we suggest an algorithm that consumes so little time amount compared with the traditional fuzzy C-mean algorithm.

III. TRADITIONAL FUZZY C-MEAN

The fuzzy c-mean algorithm is one of the common algorithms that used to image segmentation by dividing the space of image into various cluster regions with similar image's pixels values. For medical images segmentation, the suitable clustering type is fuzzy clustering. The Fuzzy c-means (FCM) can be seen as the fuzzified version of the k-means algorithm. It is a clustering algorithm which enables data item to have a degree of belonging to each cluster by degree of membership. It's developed by Dunn [9] and changed by Bezdek [10]. The algorithm is an iterative clustering method that produces an optimal c partition by minimizing the weighted within group sum of squared error objective function [10]. Is widely used in image segmentation and pattern recognition. Following are steps of traditional fuzzy c-mean:

Step1:Choose random centroid at least 2 and put values to them randomly.

Step2:Compute membership matrix:

$$U_{ij} = \frac{1}{\sum_{k=1}^{c}\left[\frac{|x_i - c_j|}{x_i - c_k}\right]^{\frac{2}{m-1}}} \quad \text{, where } m > 1, \text{ c cluster's No.} \qquad (1)$$

Step3: calculate the clusters centers:

$$C = \frac{\sum_{i=1}^{n} U^{m}{}_{ij} * x_i}{\sum_{i=1}^{n} U^{m}{}_{ij}} \qquad (2)$$

Step4: if $C^{(k-1)} - C^{k} < \varepsilon$ then Stop else go to Step2.

This traditional algorithm is an iterative algorithm that suffers from time and memory consuming because it computes membership value for each item in the data.

IV. PROPOSED ALGORITHM

In the following section we provide the improved fuzzy c-mean algorithm:

Step1: Let H represent the frequency of each item in Data.

Step2:create vector $I = \min(Data) : \max(Data)$

Step3:Choose random centroid at least 2.

Step4:Compute membership matrix:

$$U_{ij} = \frac{1}{\sum_{k=1}^{c}\left[\frac{|I_i - c_j|}{|I_i - c_k|}\right]^{\frac{2}{m-1}}} \qquad (3)$$

Step5: calculate the cluster center:

$$C = \frac{\sum_{i=1}^{n} U^{m} * H * I}{\sum_{i=1}^{n} U^{m} * H} \qquad (4)$$

Step6: if $C^{(k-1)} - C^{k} < \varepsilon$ then Stop else go to Step4.

The improved fuzzy c-mean use values that represent the frequency of items instead of actual values, in gray images the number of values of it may be reached to 256*256=65,536 and that is will take more time in processing, but in improved algorithm will take, at worst case, 256 item to process it. The proposed algorithm does not depend on whole data of image, it actually depends on data that represent the frequency of each data item in original image's data. A number of frequencies at most is 256.

V. EXPERIMENTS RESULTS

We tested Improved fuzzy c-mean by implemented by using MATLAB and compared it with implementation of fuzzy c-mean algorithm that used by MATLAB by calling command fcm, we try algorithm in database of images contains 100 images, in the following we provide a sample from tested images, in this testing sample we use C=3:

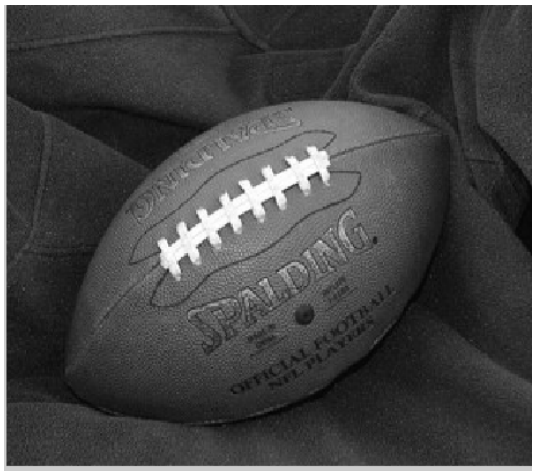


Fig. 2. Orginal image, "football"

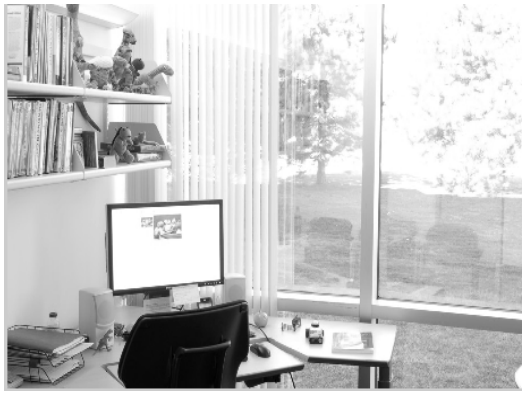


Fig. 3. Orginal image, "office"

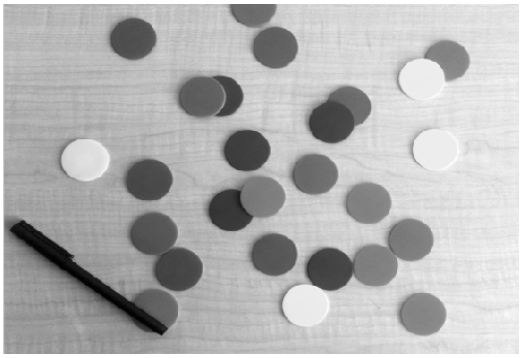


Fig. 4. Orginal image, “coloredChips”

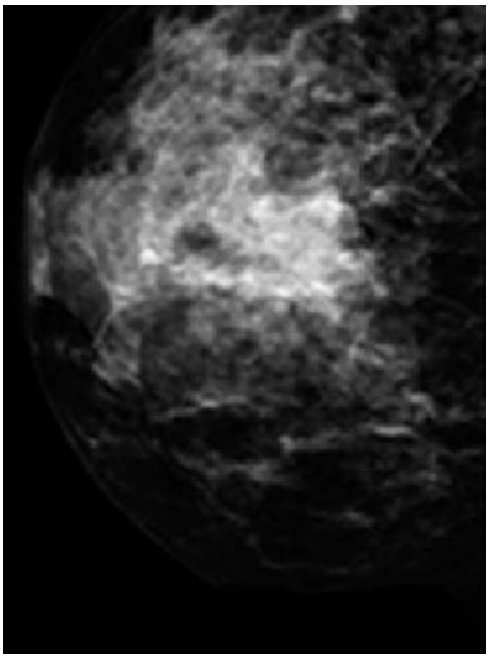


Fig. 5. Orginal image, “breast”

Fig. 6. Orginal image, “house”

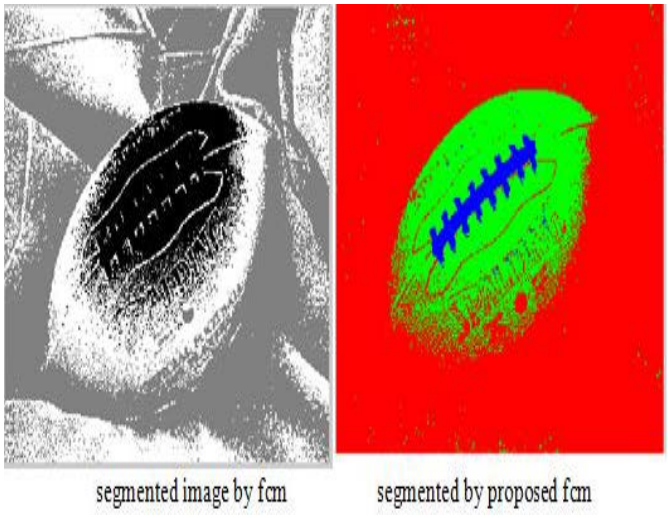


Fig. 7. Comparison between fcm and proposed fcm on “football” image

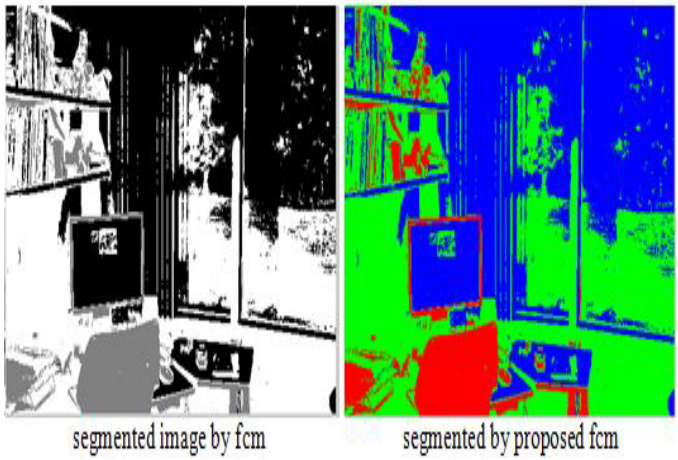


Fig. 8. Comparison between fcm and proposed fcm on “office” image

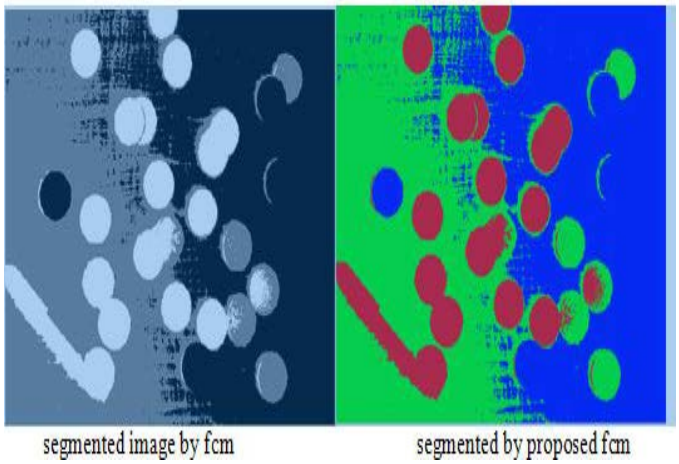


Fig. 9. Comparison between fcm and proposed fcm on “coloredChips” image

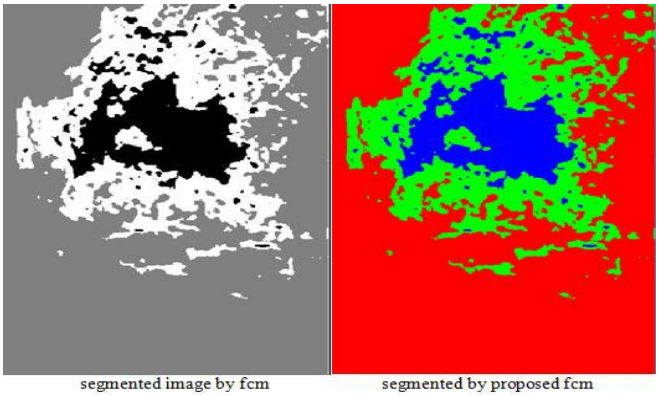


Fig. 10. Comparison between fcm and proposed fcm on "breast" image

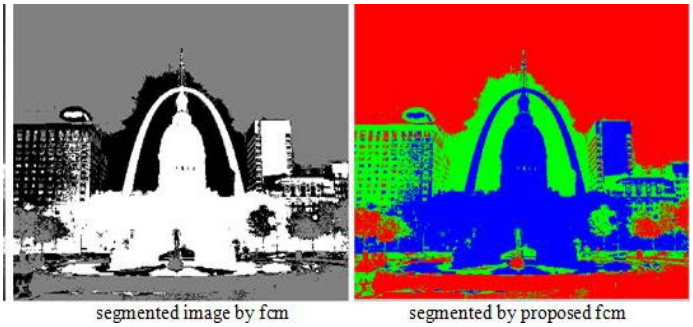


Fig. 11. Comparison between fcm and proposed fcm on "house" image

TABLE I. TIME COMPARISON BETWEEN TRADITIONAL FCM AND PROPOSED FCM

Image name	Segmentation time by fcm (sec)	Segmentation time by proposed fcm (sec)
football	3.828125	0.015625
office	22.796875	0.140625
coloredChips	8.296875	0.031250
breast	14.00000	0.062500
house	24.828125	0.078125

VI. CONCLUSION

From above results in accuracy and speed of proposed fuzzy c-mean algorithm compared with the traditional fuzzy c-mean algorithm, we conclude this algorithm is a great enhancement in implementation and performance of traditional fuzzy c-mean.

REFERENCES

[1] K.S. Deshmukh, G.N. Shinde, An adaptive color image segmentation, Electron.Lett. Comput. Vis. Image Anal. 5 (4) (2005) 12–23.

[2] Y. Zhang, A survey on evaluation methods for image segmentation, Pattern Recognition 29 (8) (1996) 1335–1346.

[3] V. Boskovitz, H. Guterman, An adaptive neuro-fuzzy system for automaticy image segmentation and edge detection, IEEE Trans. Fuzzy Syst. 10 (2) (2002) 247–262.

[4] C.Harris and M.Stephens, "A Combined Corner and Edge Detection," Proc.Fourth Alvey Vision Conf., pp.147-151, 1988.

[5] Cahoon, T.C .Sutton, M .A. Bezdek "Brest cancer detection using image processing techniques", J.C.Dept.of Comp.Sci.Univ.of West Florida, Pensacola, FL Fuzzy IEEE 2000.The ninth IEEE conference.

[6] D. L. Pham and J. L. Prince, "An adaptive fuzzy c-means algorithm for image segmentation in the presence of intensity in homogeneities," Pattern Recognition. Lett. vol. 20, pp 57-68, 1999.

[7] W. J. Chen, M. L. Giger, and U. Bick, "A fuzzy c-means (FCM)-based approach for computerized segmentation of breast lesions in dynamic contrast enhanced MRI images," Acad. Radiol, vol. 13, pp. 63-72, 2006.

[8] J. M. Gorriz, J. Ramirez, E. W. Lang, and C. G. Puntonet, "Hard c-means clustering for voice activity detection," Speech Commun, vol. 48, pp. 1638-1649, 2006

[9] J. C. Dunn, "A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters," J. Cybernetics, vol. 3, no. 3,pp. 32-57, 1973.

[10] J. C. Bezdek, "Pattern recognition with fuzzy objective function algorithms," New York, Plenum, 1981.

[11] Sipser, Michael, "Introduction to the Theory of Computation". Course Technology Inc. ISBN 0-619-21764-2, 2006.

Overview on the Using Rough Set Theory on GIS Spatial Relationships Constraint

Li Jing
[1]College of Mobile Telecommunications, Chongqing University of Posts and Telecommunications
Chongqing, China

Zhou Wenwen
[2]Chongqing University of Posts and Telecommunications
Chongqing, China

***Abstract*—To explore the constraint range of geographic video space, is the key points and difficulties to video GIS research. Reflecting by spatial constraints in the geographic range, sports entity and its space environment between complicated constraint and relationship of video play a significant role in semantic understanding. However, how to position this precision to meet the characteristic behavior extraction demand that becomes research this kind of problem in advance. Taking Rough set theory reference involved, that make measuring space constraint accuracy possible. And in the past, many GIS rough applications are based on the equivalence partition pawlak rough set. This paper analyzes the basic math in recent years in the research of rough set theory and related nature, discusses the GIS uncertainty covering approximation space, covering rough sets, analysis of it in the geographic space constraint the adjustment range.**

Keywords—space constraint; GIS; rough set; fuzzy geographic; spatial relationship

I. INTRODUCTION

GIS is a spatial information system that with the capacity of the collection, storage, management, analysis and description of the earth surface and space distribution [1, 2]. And the powerful spatial analysis ability provides advantages for understanding GIS semantics, besides it becomes the crucial premise of intelligent security monitoring prototype system by using geographical boundaries. The scale is an important concept in geographical spatial cognition, also geographic process and how to extract behavior features to satisfy movement elements have a high degree of dependence to scale, therefore, to study on such problems can be converted into discussion and explore on the positioning constraints space accuracy. But in the traditional processing method of GIS data can appear error or uncertainty, which causes the result is not entirely reliable, even error, will eventually lead to the decision-making mistakes or failure. However, rough Set as a new mathematical tool of knowledge as Deal with uncertainty and fuzzy information recently has widely put into use in GIS [3, 4]. Thus it becomes a new means to measure space constraints accuracy problem.

Studies on spatial data uncertainty at home and abroad mainly focused on space position data [5] and attribute data [6], and the unsureness of spatial relationships [7], etc. The point is the study of the uncertainty modeling, precision analysis, transmission method and visual expression [8-10]. Due to the geographical phenomenon of changing over time, the spatial data can only express one geographical phenomenon in one particular time, by that among of these studies, the considered space targets are mainly to explicitly recognize space entities (such as roads, rivers, etc.), meanwhile the uncertainty is mostly caused by measurement, digital data acquisition, and subsequent spatial analysis [11, 12]. In the approaches, to a large extent to use the methods of observational error processing in geodesy [13]. The vagueness refers to the thing can be clear described by the predefined attributes, but the boundary of the adjacent target is hard to differentiate clearly. Thus, the rough set theory has been widely used in the field of image processing, which includes system simplification, remote sensing image segmentation and remote sensing image recognition, etc. [14, 15].

This paper mainly probes into the uncertainty of this kind of spatial constraint, which includes introducing rough set into the geographical border uncertainty research, besides its basic concepts and essence (in part 2), the expression methods and rules of spatial relationship, and the applied analysis in GIS (in part 3), finally this paper discusses the current research progress, future development and difficulties (in part 4 and part 5).

II. ROUGH EXPRESSION AND REASONING OF SPATIAL RELATIONSHIP

A. Basic conception of rough set

Rough set theory is a mathematical tool proposed by professor Pawlak in 1982 which can quantitative analyze and deal with imprecise, inconsistent, incomplete information and knowledge [16]. Rough set theory, the initial prototype from the relatively simple information model, its basic idea is formed by classified concepts and rules of the relational database, through the classification of the equivalence relation and classification for the target approximate knowledge discovery. And the following are four kinds of rough set basic definitions methods [17, 18]:

1) Equivalence relation and the indiscernibility relationship

Set R is the equivalence relation limited on the U field (Meet the reflexivity, symmetry, and transitivity), recorded as $R \subseteq V \times U$. In U field, all $x \in v$ has the equivalent relation with the collection of R, recorded as $[x]\ R= \{y \in U | (x, y) \in R\}$. U/R indicates all the sets of R constitute of the equivalence classes that is quotient set.

And set R is as the equivalence relation clan, in which P⊆R, and P≠∅. All the intersection of equivalence relations in P is called the indistinguishable relationship of P, which can be recorded as ind(P), and that is [x]ind(p) = ⋂(R∈P)[X]R.

2) The upper approximation, lower approximation and rough sets

The uncertainty of rough set theory is based on the concept of upper and lower approximation. Set R is the equivalence relation limited on the U field, considered to the set X∈U, coupling (RX，$\overline{R}$X) is called a rough approximation of X in the approximate space (U, R), and

$$RX = \{x \in U | [x]\ R \subseteq X\}, \overline{R} = \{x \in U | [x]\ R \cap X \neq \emptyset\} \quad (1)$$

In which R X, $\overline{R}$ X indicate the R upper and lower approximation of X respectively. Also the set bnR(X) = RX-$\overline{R}$X refers to the R boundary region of X. The rough set defines that when bnR(X) = ∅, means RX=$\overline{R}$X, it refers to that X is R accurate sets and when bnR(X) ≠∅, means RX≠$\overline{R}$X, it calls X is R rough sets.

3) Variable precision rough set definition

The variable precision rough set definition is the extension of Pawlak rough set, it introduces β (0≤ β< 0.5) to the basic rough set, therefore it should define the majority incorporate coefficient β, before the variable precision rough set.

Definition: Set X and Y are the non-void subset of the limited discourse domain U. If to each e (e∈ X), there is e∈Y, and called it Y includes X that is recorded as X⊆ Y. And also make

$$C(x, y) = \begin{cases} 1 - |X \cap Y| / |x| & |x| > 0 \\ 0 & |x| = 0' \end{cases} \quad (2)$$

In which |x| is the cardinal number of set X and C (x, y) is the relative error resolution of set X regarded to set Y.

Definition: Based on the majority incorporate coefficient relationship, set (U, R) as approximation spaces, in which discourse domain U is non-empty limited set, and R is the equivalence relation on U, U/R = {E_1, E_2,…, E_n } is the set constituted of equivalence class or basic set of R. According to X⊆U define the β lower approximation $\underline{R}_\beta$ and the β upper approximation $\overline{R}_\beta$of X, and it approaches to variable precision rough set definition as:

$$\underline{R}_\beta X = \bigcup \{E \in U/R \mid c(E, X) \leq \beta\} \quad (3)$$

$$\overline{R}_\beta X = \bigcup \{E \in U/R \mid c(E, X) < 1 - \beta\} \quad (4)$$

4) Reduction

Set U as a discourse domain, Q and P are defined as two equivalence relation clusters on U besides Q⊆ P, if Q is independent and ind(P)=ind(Q), then call Q is the absolutely reduction of the equivalence relation cluster P and recorded it as red(P). All the absolutely relationship sets in P are the core of equivalence relation cluster P, recorded as core (P).

The most significant difference of rough set theory and other theories of dealing with uncertainty and imprecise problems is about processing not provide any prior information except data collection, so the description of the uncertainty or processing can be said to be more objective [19], therefore it provides preferential conditions for the study of spatial direction relationship.

B. The Rough thoughts of spatial direction relationship

If it is necessary to integrate the rough set theory to the GIS, must start from the basic data model to establish uncertain, fuzzy geographic data model, so as to solve the two kinds of inaccuracy and fuzzy problems between the direction relationships of the objects, and then solve the uncertainty caused by fuzzy object fuzzy boundary problems. In the spatial relationships [20], there are fuzzy and precise objects in the space object, therefore, the spatial direction relation can be mainly divided into four types: fuzzy objects and the direction relationship between fuzzy objects, the fuzzy objects and the direction relationship between precise objects, the direction relationship between the fuzzy and precise objects, and also precise objects and the direction relationship between the precise objects. Because the space objects can approximately be expressed by the rough set, the spatial direction relationship between fuzzy and precise objects can be solved by researching on the relationships between its upper and lower rough approximations sets.

Set the fuzzy objects A and B, the upper and lower rough approximations sets of them are RA, $\overline{RA}$, RB, $\overline{RB}$ separately, and adopt 048 to represent the direction relationships of them. When the grading of the direction relationships is as 8, the direction relationship knowledge base is regarded as {N, NE, E, SE, S, W, SW, NW, O}, when the grading is four direction relationships, the knowledge base is regarded as {N, E, S, W，0}. The rough expression of the spatial objects direction relationships is to represent the concept of complex direction relations as a collection of basic knowledge of the knowledge base. In which the classification knowledge of the discourse domain space is known, and the key is to confirm the relational functions between the concepts and the basic knowledge. There are two kinds of accuracy and fuzziness issues about direction relationships of the objects. The first one is caused by the fuzzy boundaries of the fuzzy objects, and the second one is by the improper methods which are caused by adopting the basic direction relationships of the knowledge base to represent the object direction relationships. The former is inherent, and the latter is the issue of methods and cognition.

The rough expression of spatial direction relationships. Set the extension cord of the outside rectangular of the object A divide the space region into O_i (1≤i≤ n), and n is as the resolution ratio of the direction. The membership function of objects B and Oi is as:

$$U(B \in O_i) = \begin{cases} 1, & B \cap O_i \neq 0 \\ 0, & B \cap O_i = 0 \end{cases} \quad (5)$$

The direction relation of objects A and B is O_{AB} = {O_i|u(B ∈O_i)=1}. Another definition is: the upper and lower rough approximations sets of the fuzzy objects direction relationships O_{AB} are $\underline{O}_{AB}$ and $\overline{O}_{AB}$， $\underline{O}_{AB}$= {$O_{\underline{AB}}$}, $\overline{O}_{AB}$= {$O_{\overline{AB}}$， $O_{A\overline{B}}$,$O_{\overline{A}\overline{B}}$}.

Because the fuzzy and precise objects can be unified expressed by the upper and lower rough approximations sets,

therefore by adopting the rough set is with the ability to unify the approximately express of the direction relationships of the fuzzy and precise objects in the frame, in order to process and reason. The upper and lower rough approximations sets of the fuzzy and precise objects direction relationships are equal, so the boundary is empty, it is consistent with the traditional expression method which eight direction relationship is based on projection, the approximation precision is 1. The boundary of the rough expression of spatial direction relationships is mainly caused by the boundary of the fuzzy objects, so the method can describe the fuzziness direction relationships of the which are caused by the fuzziness of the fuzzy objects boundary, it mainly is to approximately express the fuzzy direction relationship created by the fuzzy objects boundary, but it fail in solving the second kind of inaccuracy and fuzzy issues. [21]

III. Geography Spatial Relation Rules Extraction Based on Rough Set

A. Rough set expression of spatial relationship

To quantitatively express all kinds of the geological phenomenon spatial relationship, and then effectively converted into the format of the rough set data processing method, is the necessary conditions to use rough set rules to extract major spatial relationship of geological phenomena. Because of the reason that the rough set need to represent the data into the form of a two-dimensional table as processing the data, accordingly it requires to various kinds of geological phenomenon spatial relationship of two-dimensional form.

1) choose spatial relationship:

To aim at the specific issues of the geography, it chooses the specific spatial relationships of the geological phenomenon as the research objects according to the prior knowledge. Such as the respective features of different geological phenomenon: the water cycle, atmospheric circulation, ocean vortex, land usage and coverage, select the major effected spatial relationship factors such as the distance, topology and etc.

2) quantitative expression of spatial relationship:

To aim at the various spatial geological phenomenon, it adopts the appropriate description methods to quantitatively describe the spatial relationships, for example by employing the Euclidean distance to quantitatively describe the distance.

3) Construct spatial relationship decision table:

To convert quantitative description of geological phenomena spatial relationships into the form of a decision table. And the rows of the decision table say the research objects of geological phenomena, on the other side the columns of the decision table represent two parts: The former part known as condition attributes, on behalf of all kinds of geological phenomenon spatial relationships, the latter part of the decision table is decision attributes, the values of them are specific geological results. The values of each row are the quantitative descriptions of spatial relationships approached by various description methods of spatial relationship (except decision attributes). By using this two-dimensional table to express the spatial relationship of geological phenomena, we can employ the method of the rough set to analyze and extract the main spatial relationship rules of the geological phenomenon.

B. The spatial relationship rules extraction

Using the rough set method to process the geological phenomenon of intrinsic spatial relation rules extraction, it is mainly divided into the following steps:

1) The spatial relationship of rough sets expression: aim to the study of geological problems, by the method in Ⅲ (A), do the processing of expressing the geological phenomena of spatial relationship to the data processing format of the rough set -- the form of a decision table.

2) Using the discretization method of the rough set theory to get the decision table then to discretize. As the rough set to process the decision table, it requests the values in the decision table expressed by discrete data (such as integer, string type, enumeration type), therefore, before processing the data it must do the decision table discretization.

3) Using the attribute reduction algorithm of the rough set to do the processing of spatial relationship reduction on the space relationship the discrete decision table of the geological phenomenon space relationship, and finally form the space relationship decision rule table. The spatial relationship decision table after reducing then become the space relationship decision rule table. Because the results of the space relationship reduction are not unique, and each reduction result of the space relationship decision table will become one space relationship decision rule table, so the finally, space relationship decision rule table is the “and” of all the space relationship decision rule tables that came from each reduction result. To the final spatial relationship rule, that asks for calculating the coverage and confidence of the spatial relationship decision rules.

IV. Rough Set Theory in the Application of GIS Spatial Relationships

A. GIS data

Data analysis is an important part of GIS data processing. Rough set theory has some unique opinions such as knowledge granularity, new membership, which makes rough set particularly suitable for data analysis, therefore, there are some successful applications by using the rough set theory in GIS data analysis, for example, to adopt Worboys to handle the inaccuracy caused by multi-space or multi-semantic resolution ratio [22, 23] models like Theresa based on rough set and Egg-Yolk model study on the fuzzy and uncertainty problems of spatial data [24]; Du introduces the rough set theory into the expression of direction relationships, and present the direction relationship rough expression method, variable precision rough representation methods and rough reasoning method of direction relationship, which leads to enhance the processing and analysis ability to handle accuracy and fuzziness, and also can unify the direction relationship between the fuzzy objects and the precise objects into a framework [25]; Shi has already discussed on the rough set theory in the application of GIS uncertainty problems, which shows the rough set theory is

valuable in GIS uncertainty, but recent researches have not get deeply [26].

B. Spatial data mining

GIS is the main part of the spatial database development and contains a large number of spatial and attribute data, which has more rich and complex semantic information than the general database, and hides abundant information, all of these are very necessary for data mining. Spatial data mining means to extract the information users interested in which includes common relationships of spatial patterns and features, or spatial and non-spatial data, and some other general data characteristics hidden in the database data. Accordingly spatial data shows increasing important in the found and remake nature projects of people activity, the research and application of spatial data mining also increasingly aroused concerns, and the rough set theory is one of the important methods introduced to the data mining, in 1995 Theresa Beaubouef tried to describe a database model based on the original rough sets theory, and introduced some rough relational database models which include systems involving ambiguous, imprecise, or uncertain data [27], and Wang used GIS attribute mining as an example to analyze the application of rough set in GIS data mining [28].

C. Fuzzy geographical object modeling

The fuzzy object modeling have a wide range of meanings. The real world is complicated and full of all kinds of uncertainty, however in GIS, traditional geographic object modeling only consider the clear objects cannot reflect the complexity and uncertainty of the real world well. That causes the poor decision ability of GIS based on these kind models, which leads to hinder the development of GIS intellectualization. Using rough set to describe fuzzy object, is with the ability to fully represent the fuzziness of fuzzy objects, therefore abundant researches and applications of geographic object modeling based on the rough set theory have emerged, such as the research of Liao [29] is based on the rough set theory to transfer method to consider the polygon boundary of data fuzziness, and to employ the membership function to determine the uncertainty of the polygon boundary. Besides Du [30] combined the advantages of the rough sets and fuzzy sets dealing the fuzziness and uncertainty of the spatial data to express the fuzzy objects, and leads to expand the space data model expressing ability of fuzzy data.

D. The combination of rough set and other soft computing methods

The rough set theory is one kind of soft computing method, and the purpose of the soft computing method is to adapt to the inaccuracy of the real world around, to explore the tolerance to the accuracy, the uncertainty and partial real, and in order to achieve hand lability, robustness, and better contact with reality, whose function model is the mind of human. The main methods of calculating software are with rough sets and fuzzy sets, neural networks, genetic, and the theory of transport, etc. As solving practical problem, to adopt several computing technologies collaboratively rather than mutually exclusively has superiority compared with using one kind of computing technology. And also, it can combine the various sources of knowledge, technology and methods which solving complicated practical problems ask for. Due to the rough set has certain shortcomings as processing the data, it is necessary to combine the rough set method with other uncertain methods. At present, there are some applications of GIS data processing that combined the rough set with other soft computing methods, and the more commonly used is the combination of rough set and neural network or fuzzy sets [31, 32].

V. CONCLUSION

Rough set theory is a data analysis tool, which provides a powerful tool for the expression of GIS uncertainty information and processing, and offers favorable conditions to solve uncertain boundary space constraints. In which the fuzzy set and probability statistics method are also the commonly used methods dealing with uncertain information, but these methods need some additional information or data prior knowledge, such as fuzzy membership function and probability distribution, however sometimes it is not easy to get the information. On the other hand the rough set theory just use the information provided by the data itself, without any prior knowledge, at the same time has great advantage to reveal and express multi-level spatial knowledge.

To make a better use Rough set theory in GIS, there are still many problems to be solved. Mainly displays in: rough set can only be used for discrete space, and must be qualitative, therefore only apply to raster data, the application of vector data is difficult to determine; Rough set theory to study the expression of uncertainty in spatial analysis: recently rough set is used in attribute data, involved little in the location data uncertainty. Combining rough set and other uncertain methods, it although has made some achievements, but still there is a lot of unsolved problems ask for further research. With the further increasing of GIS data processing requirements, rough set theory is widely used to spatial data processing, at the same time, it will promote the development of the future GIS data processing technology, especially the spatial decision support system.

REFERENCES

[1] Zhang Xiao-Xiang,Yao Jing, Li Man-Chun, A Review of Fuzzy Sets on Spatial Data Handling[J], Remote Sensing Information, 2005(2).

[2] Robinson V B. Some implications of fuzzy set theory applied to geographic databases [J]. Computers Environment and Urban Systems, 1988, (12):89-97.doi:10.1016/0198-9715 (88) 90012-9.

[3] Liu Wenbao, DengMin, Analyzing Spatial Uncertainty of Geographical Region in GIS [J], Journal Of Remote Sensing, 2002, 6(1).

[4] LIAO Wei-hua, Method Study of GIS Data Transformation Based on Fuzzy Rough Set[J], Remote Sensing Technology And Application, 2007, 22(6)

[5] (Zhang Jingxiong Du Daosheng, Field-based Models for Positional and Attribute Uncertainty [J], Acta Geodaetica Et Cartographica Sinica, 1999, 28(3).

[6] Shi Wenzhong, Wang Shuliang, State of the Art of Research on the Attribute Uncertainty in GIS Data[J], Journal Of Image And Graphics, 2001, 6(9)

[7] CHENG Jicheng, JIN Jiangjun, The Uncertainty of Geographic Data[J], Geo-Information Science,2007,9(4)

[8] WANG Xiaoming, LIUYu, ZHANGJing, Geo -Spatial Cognition : An Overview [J], Geography and geographic information science,2005,21(6),pp.1-10.

[9] Roy AJ, Stell JG. Spatial relations between indeterminate regions. International Journal of Approximate Reasoning, 2001,27(3),pp.205-234.

[10] Leung Y, Ma J H , Goo dchild M F. A General Framework for Error Analysis in Measurement based GIS [C]. The 2nd International Symposium on Spatial Data Quality, Hong Kong, 2003.

[11] Hu Shengwu, Wang Hongtao, Representation and Properties Researches about Fuzzy Geographic Entities[J], Geomatics & Spatial Information Technology, 2006,29(2).

[12] Jonathan Lee, Introduction: Extending Fuzzy Theory to Object-Oriented Modeling[J], International journal of intelligent systems, 2001, 16(7).

[13] Cheng Tao, Deng Min, LI Zhilin, Representation Methods of Spatial Objects with Uncertainty and Their Application in GIS[J], Geomatics and Information Science of Wuhan University, 2007, 32(5), pp.389-393.

[14] Sun Lixin, Gao Wen, Selecting The Optimal Classification Bands Based On Rough Sets[J], Pattern recognition and artificial intelligence, 2000, 13(2).

[15] XU Yi, LI Longshu, Image Segmentation Based on Rough Entropy and K-Means Clustering Algorithm[J], Journal Of East China University Of Science And Technology(Natural Science Edition), 2007, 33(2).

[16] Pawlak Z. Rough set. International Journal of Computer and Information Sciences, 1982(11), pp.341-356.

[17] Han Zhenxiang, Zhang Qi, Wen Fushuan, A Survey on Rough Set Theory and Its Application[J], CONTROL THEORY & APPLICATIONS,1999, 16(2).

[18] Pawlak Z. Rough set-theoretical aspects of reasoning about data[M],.Dordrecht:Kluwer Academic Publishers,1991

[19] Wang Guoyin, Yao Yiyu, YuHong, A survey on rough set theory and applications[J], Chinese journal of computer, 2009, 32(7), pp. 1229-1246.

[20] Liao Weihua. GIS uncertainty analysis based on covering rough set [J], Science of Surveying and Mapping. 2012, 37(4), pp.154-156.

[21] Burrough P A, Frank A U. Geographic Objects with Indeterminate Boundaries[M].Basingstoke:Taylor and Francis,1996.

[22] WORBOYS M. computation with imprecise Geospatial Data[J].Computers, Environment and Urban Systems,1998.85-106.doi:10.1016/S0198-9715(98)00023-4.

[23] WORBOYS M. Imprecision in Finite Resolution Spatial Data [J].Geo lnformatica,1998,(03):257-279.).

[24] THERESA B, FEDERICK E P. Vague regions and spatial relationships: A rough set approach [A].2001.pp.313-318.

[25] Du Shihong, WangQiao, Spatial Orientational Relations Rough Reasoning[J], Acta Geodaetica Et Cartographica Sinica, 2003, 32(4) : 334—338.

[26] Wang Shuliang, Li Deren, Theory and Application of Geo-rough Space[J], Geomatics And Information Science Of Wuhan University, 2002, 27(3), pp. 274—282.

[27] Theresa Beaubouef, Frederick E. Petry, Bill P. Buckles, Extension Of The Relational Database And Its Algebra With Rough Set Techniques[J], Computational Intelligence, 1995, 11(2),pp.233-245.

[28] Deng Xueqing, Dong Guangjun, Gis Attribute Data Mining Based On Rough Set Theory[J], SURVEYING AND MAPPING OF SICHUAN, 2003, 26(4) .

[29] LIAO Wei-hua, Method Study of GIS Data Transformation Based on Fuzzy Rough Set[J], Remote Sensing Technology And Application, 2007, 22(6)

[30] Du Shihong, WangQiao, The Reserch of Rough Expression of Fuzzy Objects and their Spatial Relations[J], Journal Of Remote Sensing, 2004, 8(1)

[31] Greco, S., Matarazzo, B.S., S lowi'nski, R., "Rough membership and Bayesian confirmation measures for parameterized rough sets", Rough sets, Fuzzy Sets, Data Mining, and Granular Computing, Proceedings of RSFDGrC'05, LANI 364 1, PP. 314 - 324, 2005

[32] Ali Azadeha, Morteza Saberi, An integrated Data Envelopment Analysis–Artificial Neural Network–Rough Set Algorithm for assessment of personnel efficiency[J], Expert Systems with Applications, 2011, 38(3), pp. 1364–1373.

Students' Weakness Detective in Traditional Class

Artificial Intelligence

Fatimah Altuhaifa
College of Computer Engineer & Science
Prince Mohammed bin Fahd University
Alkhobar, Saudi Arabia

***Abstract*—In Artificial Intelligent in Education in learning contexts and domains, the traditional classroom is tough to find students' weakness during lecture due to the student's number and because the instruction is busy with explaining the lesson. According to that, choosing teaching style that can improve student talent or skills to performs better in their classes or professional life would not be an easy task. This system is going to detect the average of students' weakness and find either a solution for this or instruct a style that can increase students' ability and skills by filtering the collected data and understanding the problem. After that, it provides a teaching style.**

Keywords—emotional learner prediction; voice identifier and verifier; weakness detecting; artificial intelligent in education

I. INTRODUCTION

Students' weakness is one of the most important factors that prevent students' improvement. For that, researchers built a lot of e-learning that detects student's difficulty and provides the students with suitable learning style. Not all students have the motivation to use e-learning while most of them care to attend their classes. This paper aims to find a solution for students' weakness in an environment such as a traditional classroom. Understanding students' character is the primary factor for detecting difficulty. Analyzing students' emotion and activities can obtain the nature of that student.

II. SELECT A PROBLEM

This paper purposed to invent a solution for a system that helps in improving and enhancing student's skills and weakness. It will find the average of students' weakness in the class by collecting learners' emotion, reasoning the data and representing solution depending on the database or knowledge-based that the system has. The aim of this software will be reached by collecting information about each student at class and touching student's voice and analyzing it to find the suitable teaching style.

III. THE SOLUTION

As each student has his/her weakness that can affect his/her performance in the professional life, this system detects the average of students' vulnerability for every class. Then it provides a teaching style that helps in improving students' strengths.

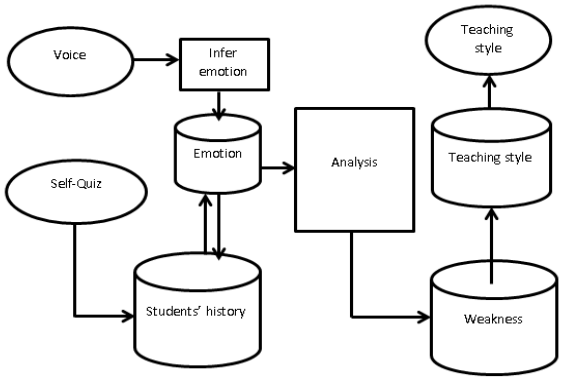


Fig. 1. Clarifying the solution

The system has one output and two inputs; one is students' voice, and the other is self-quiz for each student.

The system will touch the Students' voice for one week. Then the tone of speech is sent to software which will infer student's emotion. The system will have a database for storing students' emotion. Each student will have an ID, which is used to save assumed feelings with the unique voice. The ID with the unique voice will prevent the system from keeping the same emotion for the same student. Also, the system will add new colon if it is needed. Then the average for emotion will be taken and saved.

TABLE I. THE SAVED EMOTIONS OF STUDENTS

ID	**Hash1**	**Hash2**	**Hash3**	**Hash4**
01	Shy	Lack of self-confidence		
02	Sad	Fear	Anxiety	Despair
03	Angry	Anxiety		
04	Lack of self-confidence			
05	shy			

TABLE II. SHOWS THE AVERAGE OF EMOTIONS

Average	Shy	Anxiety	Lack of self-confidence

The system has several emotions such as self-esteem, shy, fear, cooperative, sadness and exciting. And these feelings can be changed as the user want.

The second input is a self-quiz which each student is going to take at every beginning of the semester. This self-quiz will focus on student's character to enable the system inferring the weakness. Every student has to use his/her ID for accessing to this self-quiz, and this will prevent the student from doing the quiz more than once. After taking the information, the information will be sent to the database that has the history for each student. The most characters that system will focus on are self-confidence, self-awareness, work in the team and the ability for improving.

After the system stores the information in the students' history and stores the inferred emotions in the emotion database, the system will send the information and the inferred emotions sent to investigate block. The analyzing block will examine the self-quiz for each student, and it will put the result in the table, then it will take the average of students' characters. And because the result of self-quiz can be 50% for one character and 50% for another character, the system can record more than one or two characters for one ID.

TABLE III. DETECTING MORE THAN ONE CHARACTERISTIC FOR THE SAME PERSON

ID	Character1	Character2	Character3	
01	Lack of Self-awareness	Lack of work in team		
02	self-confidence			
03	self-confidence			
04	Lack of improving			

TABLE IV. THE AVERAGE FROM THE FIRST TABLE

Average	self-confidence

Then the analysis block will relate the character and the emotion to each other to grip an appropriate weakness to the class. The system will send the gripped vulnerability to weakness the database which has a table for vulnerability and a proper teaching style for each weak point. After the system chooses the style of education, the system will send the result to the screen as a report which has the method of teaching.

Also, there is another problem that system can face. This problematic is one pair voice come from outside the class to the software. For solving this problem, the system will have an application that can be work on and work off by the instructor. This application is identifying students' voice at the beginning of the semester, or when any new student attends the class, who has done late registration for the course, the application can identify the students' voice when each student identify himself/herself to everyone at the first class of the semester.

1) *How can the voice be touched?*
2) *How will the emotion extract from the tone of speech?*
3) *How can the system identify and verify the vocal sound?*
4) *What is the important of identifying and confirming the voice?*
5) *How are the data going to be analyzed?*
6) *Which is the type of teaching style going the system has?*
7) *How can the system collect the students' data?*
8) *How can the system deal with overlapping voice?*

IV. METHODOLOGY AND ANSWERING SOME OF THE QUESTIONS

Most of the work will depend on how to handle the voice so the answer will be depending on understanding the natural of the sound and its signal. And more on the signal, because each type of speech has different vibration, frequency and signal which can help in understanding students' natural.

A. Catching the Voice

As known the voice travels through the medium (air, water, etc.) as a vibration. The system will have a Neumann microphone that one of the employees in the school has to install it on the ceiling. The mounted instrument will allow each sound to be easy to catch. This kind of tool is upright for using in studio, TV and room because it can detect the voice from a long distance.

B. Extracting Voice's Feature

When the microphone receives the speech, it will send it to the system. The system will extract at least one feature of the voice signal then it will determine the emotion that is related to the voice signal **[1].** Each sound has different vibration, for example, the noise has high waves, and these waves are close while the beneath sound has vibrations that have small distant from each other.

Fig. 2. Neumann microphone [2]

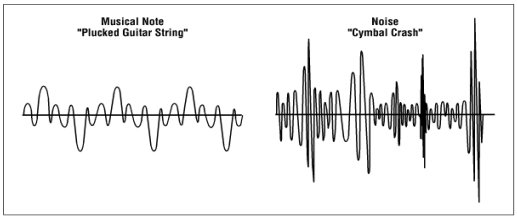


Fig. 3. Diversity in voice's frequency [1]

The Sound has three composed loudness, pitch, and timbre. Loudness is a strength or weakness of sound due to air pressure. The application can know the pitch by the frequency of the voice and rate repetition. Timber is the tone, color, or texture [1].

C. The Ability of Extraction Emotion From The Voice

A.Valery works in a system that can detect emotion in the voice signal. They claim that each emotion has different frequency and vibration. The system uses Object Oriented Programming (OOP). This programming helps in developing computer software and doing an analysis. This system receives the voice and does an analysis to it **[3].**

D. Identifying Verifying Voice

One of the inputs to the system is students' speech which will be touched in the class by the microphone for one week every month. Then, the vibrations of voice will be analyzed to infer the emotion of student. The system will need to check student's awareness every month to check students' improvement. Also, avoiding duplicate in emotion for the same student will be required for getting a perfect result. The system will need to recognize the voice and assigns it to specific ID, and it will need to verify the voice to ensure that voice has an ID to prevent the system from saving the duplicate voice of the same student. For this purpose, the system will have a database for storing voice's feature in it. The Hidden Markov Models with the Gaussian Mixtures (HMM-GM) is the used method [3]. With this approach, there are three ways for extraction voice's feature. These are filters, Speaker Dependent Frequency Filter Bank (SDFFB) and Speaker Dependent Frequency Cepstrum Coefficients (SDFCC). The filters work to find the domain of frequency by defining the discrete frequency domain. According to the different in vocal sound, the SDFFB's method is Linear Prediction Coefficients, while the system cannot use the SDFCC in some types of voice recognition because it "emphasizes the speaker influence too much" [4].

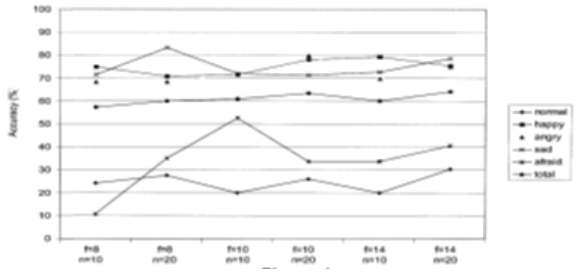


Fig. 4. It shows the different in sound frequency for diverse emotion [1]

E. Ability of Distinguishing Voices in Overlapping

T.WEI-OH and L.SHI-JIE wrote about a system that can distinguish between overlapping sounds and non-overlapping. And this system can determine who speak each part of overlapping sounds. The system will take the signal and checks if it is overlapping or not overlap then it will start analyzing the voice to determine who is speaking?[5].

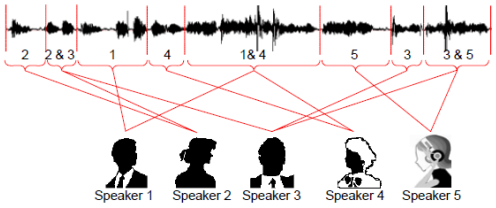


Fig. 5. Extracting the simulating voices [5]

F. Avoiding Duplicating Emotion For the Same Student

The system will need some algorithm or coding for achieving this goal. After verifying the voice and being sure if it is existing or not, then it will check the identification of the voice. It is not important to know exactly for which student this voice but it is important to assign each tone with unique ID. After recognition the voice, the system has to record the extracting emotion with its accent which means with its ID. But before recording the emotion, the system must ensure that this ID does not have this feeling. It the ID has the emotion the system will not record it while if the sense does not exist for that ID, the system has to record it.

```
If (newVoice == OldVoice){
If ( newEmotion == OldEmtoin)
Add emotion;
Else
Exit 0; }
Else{
If( saving == True )
Add newVoic;
Else
Exit 0;}
```

V. Related Works

This paper aimed to detect the student's knowledge state to help the system to find an appropriate learning plan for the student that improving the student and the learning effectively. It is a web-based educational system which provides the student with a quiz and observes his/her interaction on the web and his/her moving to the mouse and choosing the answer with observing the time for his/her action. Choosing appropriate learning plan depends on student diagnose. For achieving this purpose, the system will have student profile, student model, content model and learning plan. The application uses fuzzy logic supported modeling in analyzing student. The using of fuzzy logic supported modeling is depending on the student profile and the quiz **[6].** The way that the system is using for analyzing student and finding the appropriate plan for him/her is the most part that is related to

evaluating students' weakness in the class. The reason that makes it the most important part is to diagnose student individually will be needed to discover the average of weakness in the class.

N.Aghaee and S.Ören discussed in their paper the process of finding the best solution for education style for the e-learning depending on the student's emotions. The primary inputs, at the System, are the student's personality and emotion while the output will be the teaching's style. The aim of this application is to inspire the student and enhance their ability. For reaching their goal, they used personality filter for software agent, emotion filter for software agent and MBTI indicator. The system has eight types of emotions which are fear, anger, sorrow, joy, disgust, acceptance, anticipation and surprise. The MBTI is the responsible for inference the suitable learning style, according to student's cognitive, emotion and character **[7].** This paper related to our paper in the way in finding the appropriate learning style depending on the emotions and character but has different in the environment and how to collect the student's data. In this paper, the emotions will be inferred by facial expression while in our paper, the emotions will be inferred by using an agent that can cause analysis emotions from voice. In our paper, student's character will be assumed by student's history. In their paper, the environment is eLearning while our environment is a traditional classroom.

Another relative document to this article is A Model of the Student Behaviors in a Virtual Educational Environment, which is written by Moisil. This paper has a system that finds a suitable learning style on online for student depending on student's behavior, beliefs, and motivation. The system has different modeling in learning style for various kinds of students which help to determine the type of student performance. The system has four learning style. For achieving the most appropriate style, the designer provides the system with two types of questionnaire. One has 80 items, and the other has 40 items. The four learning style are Activists (Do), Reflectors (Review), Theorists (Conclude) and Pragmatist (Plan) **[8].** This system has three agents that help in improving student's skills. These agents are a personal assistant, tutor, and the mediating agent. This system similar to our system in trying to find learning style depending on student behavior and beliefs while it differs in the environment. This system is e-learning while our system is in the real environment such as the classroom.

VI. CONCLUSION

The purpose of this application is to increase the students' skills and ability. For accomplishing that, the weakness of the student needs to be known which will enable the system to choose an appropriate style of education. This system depends on the analyzed students' emotion and history for getting the result.

This system will need to improve analyzing part in the feature. To see how the system can do the analysis and what is the proper technique for this purpose. Also, it needs work deeply in algorithm part or coding part for being sure that the analyzing done will. In addition to that, the system needs to mention the teaching style and how this refers to each weakness.

VII. SELF-QUIZ

1. **My need to take this course now:**
   - High. I need it straight away for a degree, job, or other important reason.
   - Moderate. I can take it on campus later or substitute another course.
   - Low. It's a personal interest that I can postpone it.
2. **Considering my professional and personal schedule, the amount of time I have to work on a course is:**
   - More than adequate for a campus class.
   - The alike as for a class on campus.
   - Less than adequate for a class on campus.
3. **I can classify myself as someone who:**
   - Often I can do things before the dead time.
   - Needs reminding to get things done on time.
   - Puts things off until the last minute.
4. **Feeling that I am one of a class is:**
   - Not particularly necessary for me.
   - Somewhat important to me.
   - Critical to me.
5. **As a reader, I would classify myself as:**
   - Good. I usually understand the text without help.
   - Average. I sometimes need help to understand the text.
   - Slower than average. **[5]**
6. **You have an assignment that you have to do it in group; you prefer to work with**
   - Friends.
   - No matter with whom.
   - Not with friends.
7. **You are going to present your work you prefer to perform in front of**
   - Only Friends.
   - No matter in front who.
   - Only the class and your instructor without other teachers.

Do you believe in yourself? Yes no

Are you happy in your major? Yes no

Is your major mostly what you talk about? Yes no

Do you feel guilty for doing the things that you want to do?Yes no [**9]**

REFERENCES

[1] J. Beggs and D. Thede. *Designing Web Audi.* Chapter 2,The Science of Sound and Digital Audio. O'Reilly & Associates .2001.

[2] Boré, G., & Peus, S. (1999). *Microphones Methods of Operation and Type Examples*. Berlin: Druck-Centrum Fürst GmbH.

[3] A.Valery.(1999). System Method and Article of Manufacture for Detecting Emotion in Voice Signals through Analysis of a Plurality of Voice Signal Parameters. United States Patent **Publication**. G10L17/00; G10L17/00; (IPC1-7): G10L15/00.

[4] F.Orság. (2010). Speaker Dependent Coefficients for Speaker Recognition. *International Journal of Security & Its Applications*, 4(1), 31-47.

[5] T. WEI-HO, and L. SHIH-JIE. (2010). Speaker Identification in Overlapping Speech. *Journal of Information Science & Engineering*, 26(5), 1891-1903.

[6] D. Xu, H. Wang and.K. Su (2002). Intelligent Student Profiling with Fuzzy Models. *HICSS '02 Proceedings of the 35th Annual Hawaii International Conference on System Sciences (HICSS'02*,3(3), 0-7695-1435-9.

[7] *N.Aghaee and S.Ören (2008). Agents with Personality and Emotional Filters for an E-learning Environment. 2008 Spring Simulation Multiconference (SpringSim'08)- Poster Sessions (SCS-Poster sessions 2008).Ottawa, Canada, April 14 - 17, 2008*

[8] Moisil, (2008). A Model of the Student Behaviour in a Virtual Educational Environment. *International Journal of Computers, Communications & Control*, 3(3), 108-115.

[9] http://www.clt.odu.edu/oso/index.php?src=pe_isdlforme_quiz

Thresholding Based Method for Rainy Cloud Detection with NOAA/AVHRR Data by Means of Jacobi Itteration Method

Kohei Arai[1]
Graduate School of Science and Engineering
Saga University
Saga City, Japan

***Abstract*—Thresholding based method for rainy cloud detection with NOAA/AVHRR data by means of Jacobi iteration method is proposed. Attempts of the proposed method are made through comparisons to truth data which are provided by Japanese Meteorological Agency: JMA which is derived from radar data. Although the experimental results show not so good regressive performance, new trials give some knowledge and are informative. Therefore, the proposed method suggests for creation of new method for rainfall area detection with visible and thermal infrared imagery data.**

Keywords—Jacobi itteration method; Multi-Variiate Regressive Analysis; AVHRR; Rainfall area detection; Rain Radar

I. INTRODUCTION

Rainfall area detection with satellite based visible and thermal infrared sensor data is tough issue because the visible and thermal infrared sensor data represent just cloud surface reflectance and temperature. In general, the rainy clouds which cause rainfall can be divided into two kinds, nimbostratus and cumulonimbus. The reflectance and the temperature at the top of the cumulonimbus are relatively high and extremely cold, respectively because the height of the cumulonimbus is quite high. On the other hand, the reflectance and the temperature of the nimbostratus are comparatively low and relatively warm, respectively because the height of the nimbostratus is comparatively low. Meanwhile, no rainy cloud types show very similar characteristics in terms of cloud top reflectance and temperature. Therefore, it is extremely difficult to discriminate between rain and no rain clouds.

There are 10 types of clouds which include four types of cumulus, cumulonimbus, stratus, stratocumulus in the lower cloud, three types of nimbostratus high-rise clouds and high cumulus clouds in the middle clouds, in the high-rise clouds three types of cirrus, cirrocumulus, and cirrostratus. Moreover, these clouds are overlapped sometime. Therefore, it is tough to discriminate rainy clouds by using only reflectance and temperature at the top of the clouds.

Microwave radiometer data represent cloud liquid in rainy cloud. Therefore, some methods for detecting rainy clouds with microwave radiometer data have been proposed already [1]-[9]. On the other hand, limb sounding data also represent some extent of rainy clouds information. Therefore, some methods for rainy cloud detection based on limb sounding data have also been proposed so far [10]-[16].

The rainy cloud detection method proposed here is based on thresholding of visible and thermal infrared radiometer data by means of Jacobi iteration method. The proposed method is to be compared to the multiple linear regressive analyses by only using visible and thermal infrared data observed from space. In the method, Probability Density Function: PDF of the visible and thermal infrared data are calculated. Then the PDF is approximated with the best fit ideal normal distribution. After that, the visible and thermal infrared data are binarized (0 denotes no rain, and 1 means rain) with the most appropriate threshold determined by means of Jacobi iteration method.

The proposed method is described followed by experiments. The experimental results are validated with the posterior created weather maps by using rainfall radar data and the other meteorological data in the following section followed by conclusion with some discussions.

II. PROPOSED METHOD

A. Discrimination of Cloud Types

Within 10 types of clouds, cumulonimbus and nimbostratus clouds are major concern because I intend to discriminate between rainy clouds and the clouds without rainfall. The cumulonimbus clouds are situated in the lower layer of the atmosphere while the nimbostratus clouds are situated in the middle layer of the atmosphere. Therefore, relatively low cloud top temperature and comparatively low cloud top reflectance of clouds have to be found with visible and thermal infrared radiometer data. By using visible and thermal infrared data, appropriate threshold which allows discriminate nimbostratus / cumulonimbus and the other clouds has to be determined.

B. Jacobi Iteration Method

The proposed method uses Jacobi iteration method. The Jacobi iteration method is expressed as follows,

$$g_k = g_{k-1} + \alpha_k F'_k \quad (1)$$

$$\alpha_k = \frac{1}{2} \quad (2)$$

$$F_k = \frac{1}{\sqrt{2\pi\sigma_{S1}^2}} exp\left(\frac{-(g_{k-1}-m_{S1})^2}{2\sigma_{S1}^2}\right) + \frac{1}{\sqrt{2\pi\sigma_{S2}^2}} exp\left(\frac{-(g_{k-1}-m_{S2})^2}{2\sigma_{S2}^2}\right) \quad (3)$$

where k denotes iteration number while F_k denotes summation of the PDF functions of the approximated normal distribution of S_i (i=1 denotes rainy cloud while i=2 denotes non rainy clouds of visible and thermal infrared radiometer data. Namely, the PDF functions of visible and thermal infrared data are firstly created then the most appropriate approximation normal distribution functions are calculated. After that, cross point between two approximated normal distribution is determined by using Jacobi iteration method. This cross point is used for threshold for discrimination between rainy and non-rainy clouds.

C. Process Flow

Fig.1 shows the process flow of the proposed method for discrimination of rainy and non-rainy clouds with visible and thermal infrared radiometer data.

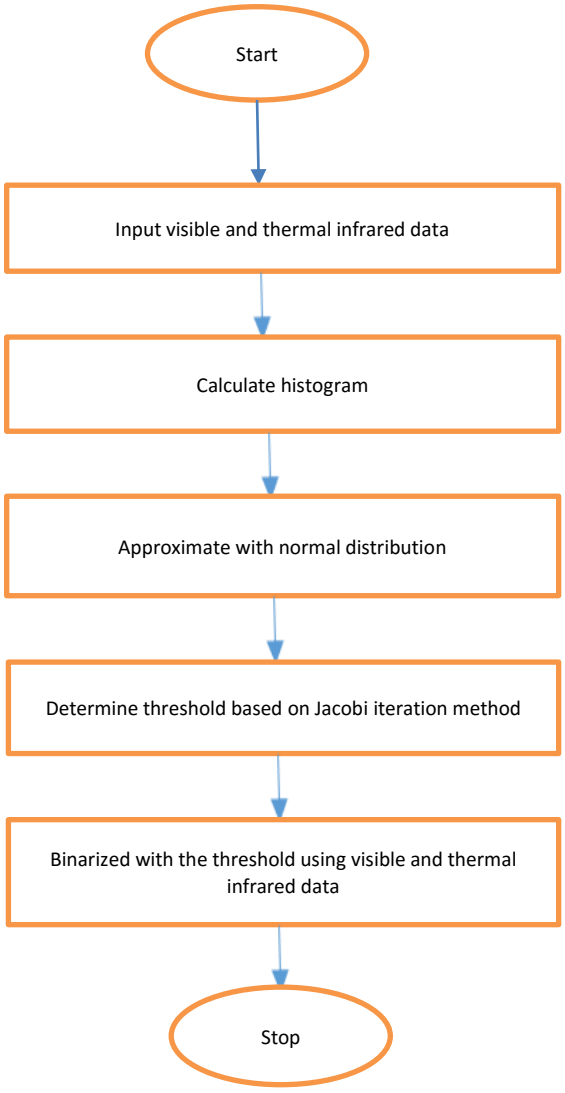
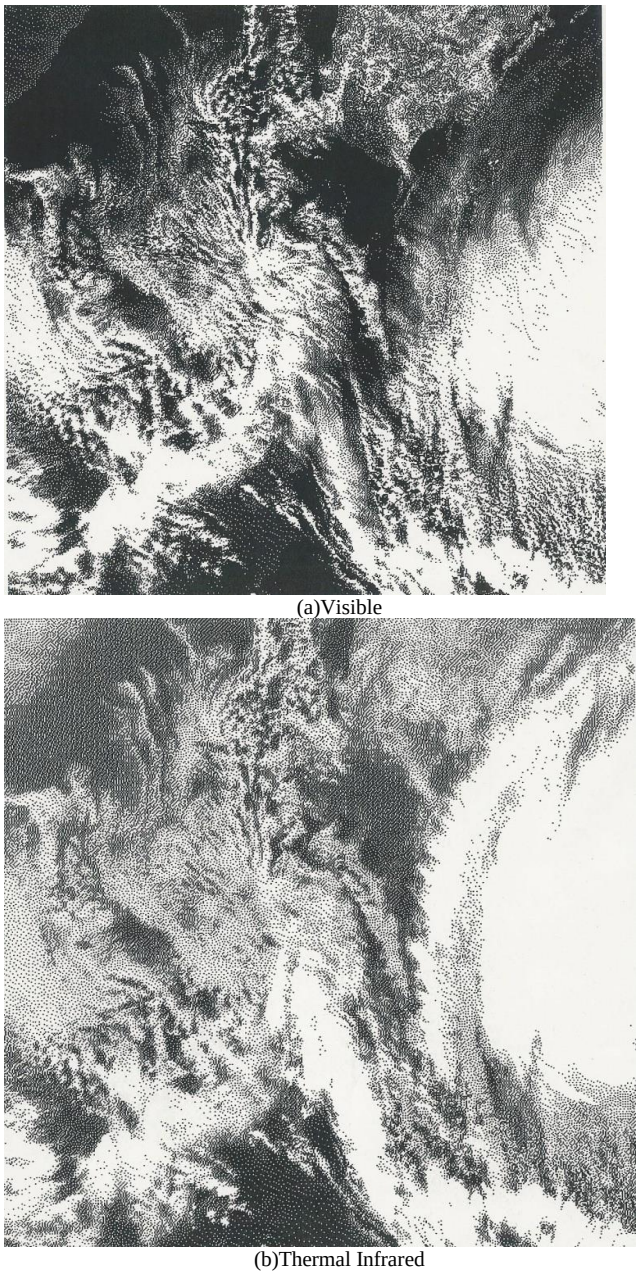


Fig. 1. Process flow of the proposed method for discrimination between rainy and non-rainy clouds with visible and thermal infrared data

III. Experiments

A. Visible and Thermal Infrared Imagery Data Used

NOAA/AVHRR (National Oceanic and Atmospheric Administration / Advanced Very High Resolution of Radiometer) of visible and thermal infrared data of Tohoku, Japan which is acquired on February 12 1997 is used. Visible channel which covers the wavelength ranges from 0.73 to 1.10 μm is used while thermal infrared channel which covers the wavelength ranges from 10.3 to 11.3μm is used.



(a)Visible

(b)Thermal Infrared

Fig. 2. NOAA/AVHRR of visible and thermal infrared data of Tohoku, Japan which is acquired on February 12 1997 is used for the experiments

Fig.2 shows the visible and the thermal infrared imagery data used for the experiments. The images consists of 512 by 512 pixels (the pixel represent 2.2km by 2.2km ground surface areas). Radiometric resolution of visible channel is 0.1% of

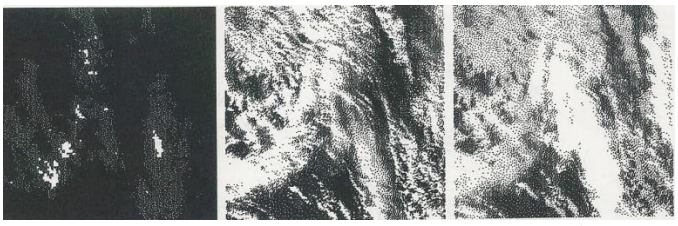
albedo while that of thermal infrared channel is 0.2degree C. These data are represented by 8 bits (256 levels) while minimum and maximum physical values correspond to 0 to 35% for visible channel while 243K to 294K for thermal infrared channel.

B. Truth Data Used

As a truth data of rainfall areas, radar data derived rainfall areas which is provided by Japanese Meteorological Agency: JMA is used. Fig.3 shows the radar data derived rainfall areras of image which consists of 500km by 500km (the pixel consists 2.6km by 2.6km). Black areas show the rainfall areas with 1 to 4mm/hr of rainfall rate while hatched areas shows the rainfall areas with less than 1mm/hr of rainfall rate.

Fig. 3. Radar data derived rainfall areas which is observed at 10:00 in the morning on February 12 1997

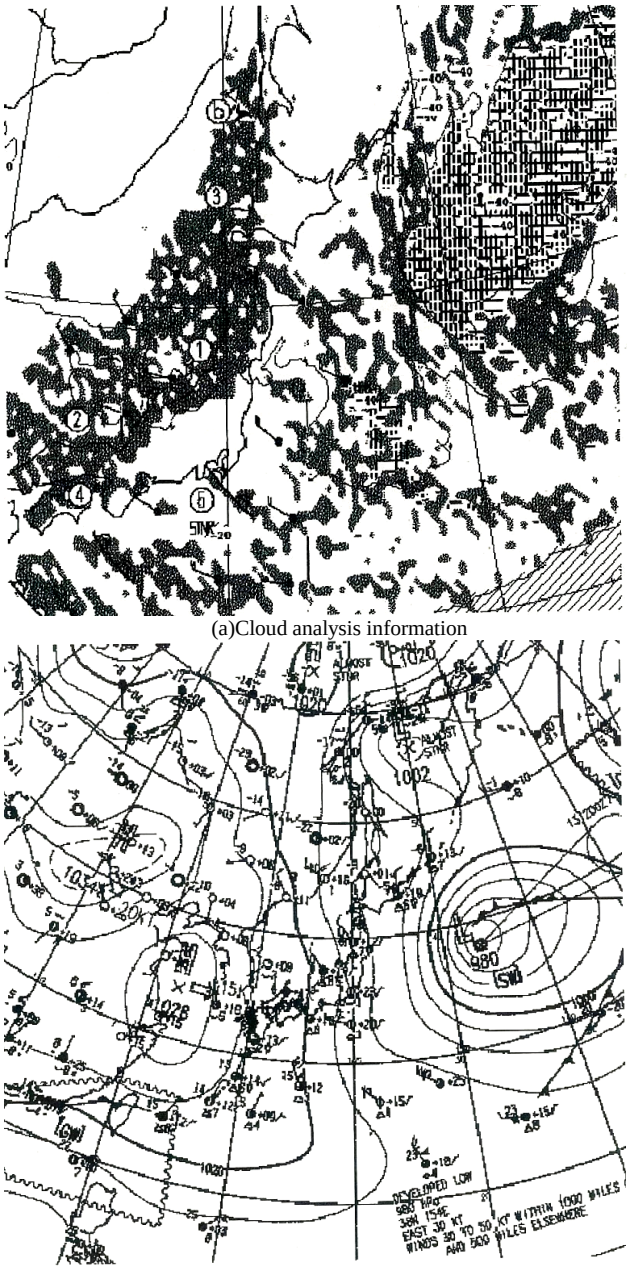
Also, Fig.4 shows the imagery data which are used for multiple linear regressive analysis (Radar data on the right, Visible channel of imagery data in the middle, and Thermal infrared imagery data on the left, respectively) which are acquired on February 12 1997.



(a)Radar (b) Visible (c) Thermal Infrared

Fig. 4. Imagery data which are used for multiple linear regressive analysis (Radar data on the right, Visible channel of imagery data in the middle, and Thermal infrared imagery data on the left, respectively) which are acquired on February 12 1997

These visible and thermal infrared imagery data are extracted from the NOAA/AVHRR imagery data. Also, the cloud analysis information image which is acquired at 18:00 on that day provided by JMA and weather map at 12:00, noon on that day is shown in Fig.5 as reference data for rainfall areas. ① to ④ in Fig.5 (a) denotes cumulus clouds while ⑤ and ⑥ areas denote non rainy areas. The eastern portion of Japanese island, in particular, Hokkaido and Tohoku, there are relatively large cloudy areas. Particularly, ① area shows rainfall areas.



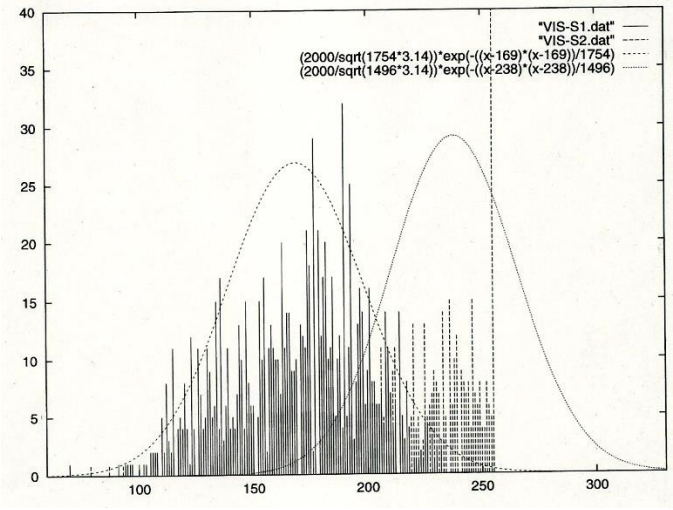
(a)Cloud analysis information

(b)Weather map

Fig. 5. Cloud analysis information image which is acquired at 18:00 on that day and weather map at 12:00, noon on that day are shown in Fig.5 as reference data for rainfall areas

C. Experimental Results

Fig.6 (a) shows histograms of rainy (S1) and non-rainy (S2) cloud areas of visible channel of imagery data and the approximate PDF functions for the histograms for both rainy and non-rainy cloud areas. Meanwhile, Fig.6 (b) shows histograms of rainy (S1) and non-rainy (S2) cloud areas of thermal infrared channel of imagery data and the approximate PDF functions for the histograms for both rainy and non-rainy cloud areas. The mean and variance of S1 of the visible channel of data are 168.874 and 877.135, respectively while those for S2 of visible channel of data are 238.292 and 748.170, respectively.



(a)Visible

(b)Thermal Infrared

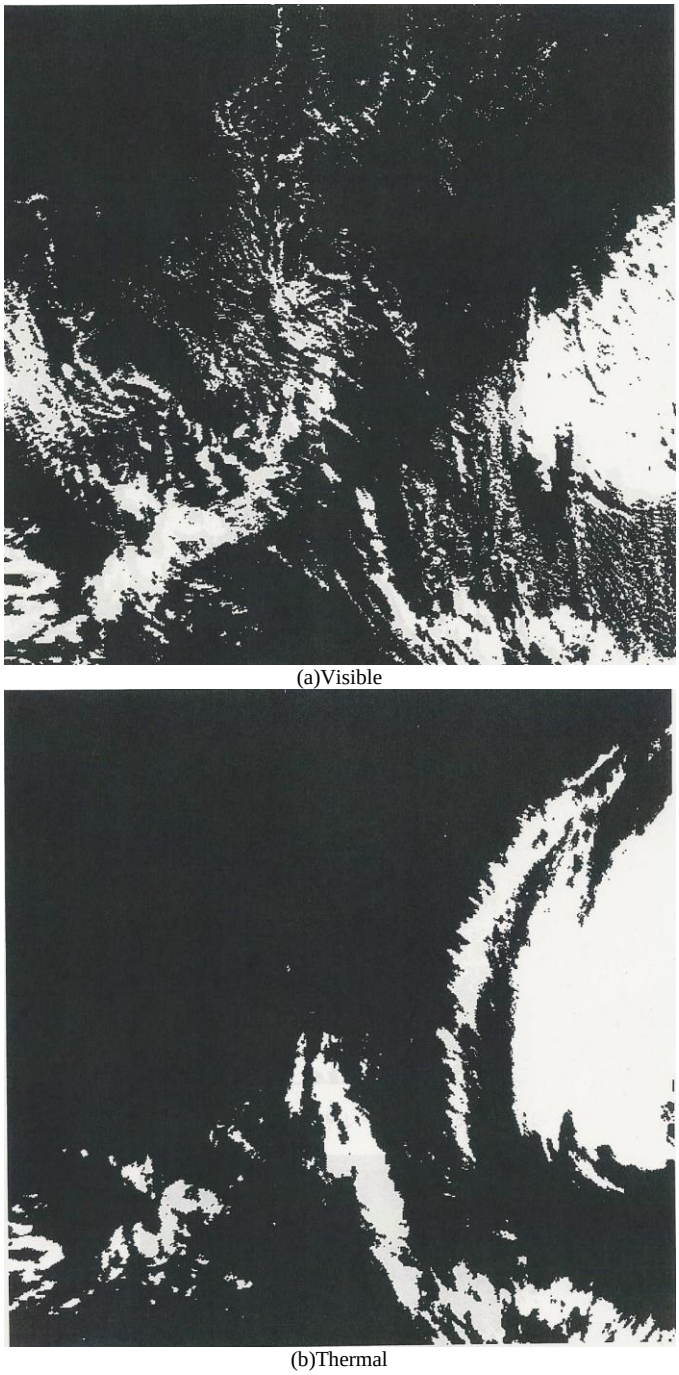
Fig. 6. Histograms of rainy (S1) and non-rainy (S2) cloud areas of thermal infrared channel of imagery data and the approximate PDF functions for the histograms for both rainy and non-rainy cloud areas

It is found that histograms of rainy and non-rainy clouds are very close for the thermal channel of data while those are relatively distinguishable for the visible channel of data.

Fig.7 (a) shows the binarized image of the visible channel of data while Fig.7 (b) shows that of the thermal infrared channel of data with the determined thresholds by the Jacobi iteration method, respectively.

D. Validation of the Proposed Method

Fig.8 (a) shows raw image of Rainfall radar while Fig.8 (b) shows the rainfall rate extracted image with rain fall radar data. On the other hand, Fig.9 shows the extracted rainfall areas with NOAA/AVHRR of visible and thermal infrared imagery data. White square box in the Fig.9 shows the corresponding area of interest with the rain radar derived rainfall areas.



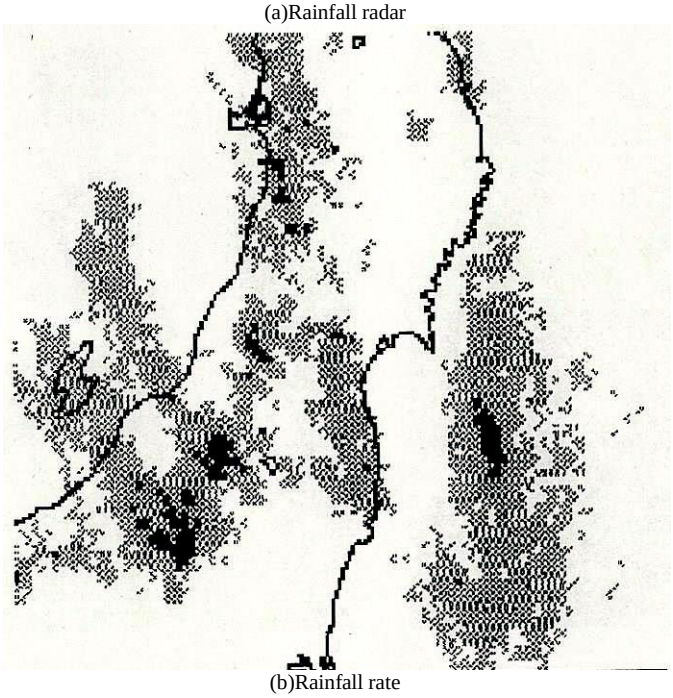
(a)Visible

(b)Thermal

Fig. 7. Binarized images of the visible and the thermal infrared channels of data which are acquired on February 12 1997

Through comparisons between rainfall radar data derived rainfall rate image and the extracted rainfall areas with NOAA/AVHRR data based on the proposed method, it is found that the extracted rainfall areas with NOAA/AVHRR data based on the proposed method shows relatively heavily rainfall areas. The extracted rainfall areas with NOAA/AVHRR data based on the proposed method is corresponding to the rainfall areas with rainfall rate of 2 to 4 mm/hr.

(a)Rainfall radar



(b)Rainfall rate

Fig. 8. Rainfall radar image and the rainfall rate extracted image with rain fall radar data

Fig.10 (a) shows the binarized image of rainfall radar derived rainfall rate while Fig.10 (b) shows the binarized image of NOAA/AVHRR of visible and thermal infrared data derived rainfall areas based on the proposed method. Both images show marginal coincidence in terms of rainfall areas. Root Mean Square Difference: RMSD between the aforementioned two binarized images is 13.727. Therefore, it is marginal accuracy of rainfall area detection.

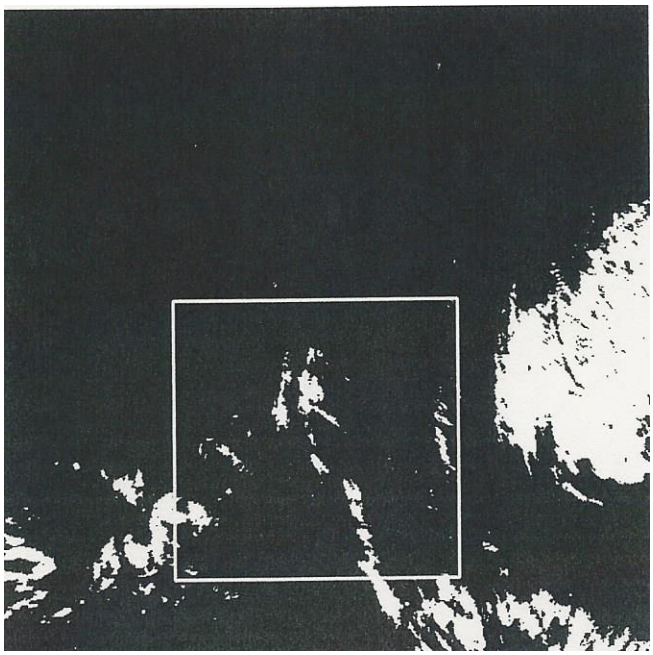
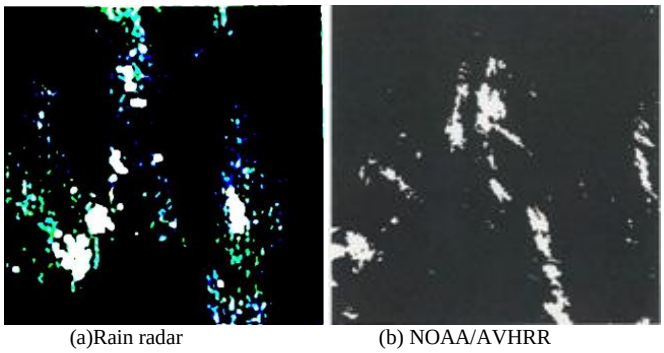


Fig. 9. Extracted rainfall areas with NOAA/AVHRR of visible and thermal infrared imagery data (White square box shows the corresponding area of interest with the rain radar derived rainfall areas)



(a)Rain radar (b) NOAA/AVHRR

Fig. 10. Binarized images of rainfall radar derived rainfall rate and NOAA/AVHRR of visible and thermal infrared data derived rainfall areas based on the proposed method

E. Alternative Method for Rainfall Area Detection

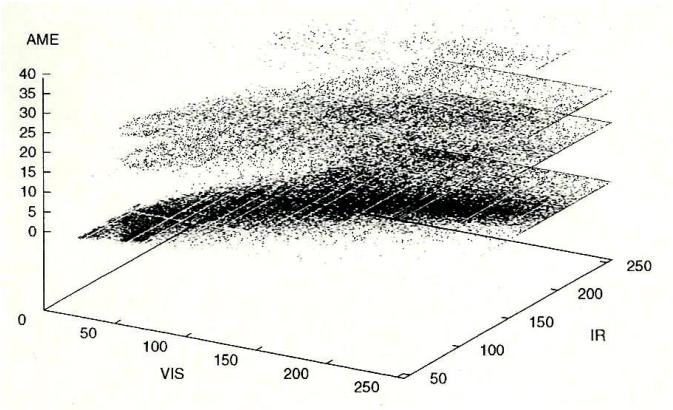
As described before, there is the alternative method of rainfall area detection, multiple linear regressive analyses: MLRA based method. Namely, rainfall radar data derived rainfall rate is approximated with the NOAA/AVHRR of visible and thermal infrared radiometer data through the MLRA. Fig.11 (a) shows the scatter plots of the rainfall rate

(AME) and the visible and the thermal infrared channels of NOAA/AVHRR data while Fig.11 (b) shows MLRA equation which is the result from the MLRA expressing the rainfall rate as the functions of the visible channel and the thermal channel of NOAA/AVHRR data.

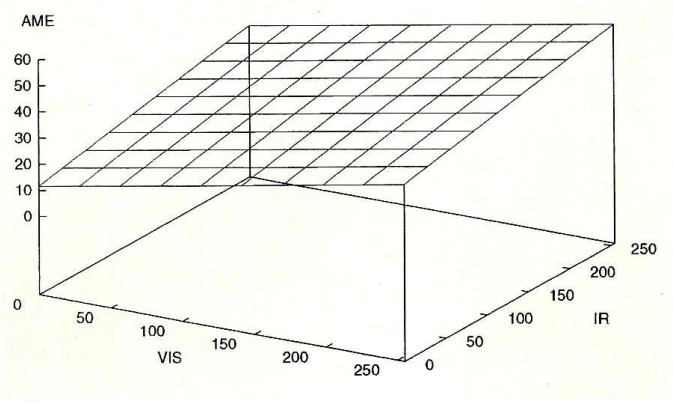
MLRA equation is expressed in equation (4).

$$Z=0.0155x+0.00941y+1.731 \quad (4)$$

where Z denotes rainfall rate while x and y denotes the visible and the thermal infrared channels of NOAA/AVHRR data.



(a)Scatter Plots



(b)MVRA Equation

Fig. 11. Scatter plots of the rainfall rate (AME) and the visible and the thermal infrared channels of NOAA/AVHRR data and MLRA equation which is the result from the MLRA expressing the rainfall rate as the functions of the visible channel and the thermal channel of NOAA/AVHRR data

The coefficient of determination of the MLRA is 0.020, and multiple correlation coeffient is 0.142. Also, degree of freedum corrected coefficient of determination is 0.020 while degree of freedum corrected multiple correlation coefficient is 0.141. Therefore, not so good correlation is found between rainfall rate and the visible and the thermal infrared channels of NOAA/AVHRR data. Consequently, the proposed method is superior to the MLRA based approach.

IV. CONCLUSION

Thresholding based method for rainy cloud detection with NOAA/AVHRR data by means of Jacobi iteration method is proposed. Attempts of the proposed method are made through comparisons to truth data which are provided by Japanese Meteorological Agency: JMA which is derived from radar data. Although the experimental results show not so good regressive performance, new trials give some knowledge and are informative. Root Mean Square Difference: RMSD between two binarized images of the rainfall radar derived rainfall rate and the NOAA/AVHRR derived rainfall area detected resultant image is 13.727. Therefore, it is concluded that the proposed method has a marginal accuracy of rainfall area detection. Therefore, the proposed method suggests for creation of new method for rainfall area detection with visible and thermal infrared imagery data.

Through the comparison between the proposed method and the multiple linear regressive analyses, it is concluded that the proposed method is superior to the Multiple Linear Regressive Analysis: MLRA based approach.

Further investigations are required for new additional information such as collocated microwave radiometer data and limb sounding data.

ACKNOWLEDGMENT

The author would like to thank Mr. Hirokazu Taniguchi of former student of Saga University for his effort to conduct the experiments.

REFERENCES

[1] Chuang, C. and K. V. Beard, A numerical model for the equilib- rium shape of electrified raindrops, J. Atmos. Sci., 47, 1374- 1389, 1990.

[2] Crewell, S., H. Czekala, U. LShnert, and C. Simmer, MICCY- a 22 channel ground-based microwave radiometer for atmopsh- eric research, submitted to Radio Science, 2000a.

[3] Crewell, S., U. LShnert, A. van Lammeren, and C. Simmer, Cloud remote sensing by combining synergetic sensor information, Phys. Chem. Earth (B), 25, No. 10-12, 1043-1048, 2000b.

[4] Czekala, H., S. Crewell, A. Hornbostel, A. Schroth, C. Simmer, and A. Thiele, Validation of microwave radiative transfer cal- culations for nonspherical liquid hydrometeors with ground- based measurements, J. Appl. Meteorol., 2000, (submitted for publication).

[5] Czekala, H. and C. Simmer, Microwave radiative transfer with non- spherical precipitating hydrometeors, J. Quant. $pectros. Radiat. Transfer, 60, 365-374, 1998.

[6] Fox, N., and A. J. Illingworth, The potential of spaceborne cloud radar for the detection of stratocumulus clouds, J. Appl. Me- teorol., 36, 676-687, 1997.

[7] Gfildner, J. and D. Spiinkuch, Results of year-round remotely sensed integrated water vapour by ground-based microwave radiometry, J. Appl. Meteorol., 38, 981-988, 1999.

[8] Mishchenko, M. I., Calculation of the amplitude matrix for a nonspherical particle in a fixed orientation, Appl. Opt., 39, 1026-1031, 2000.

[9] Solheim, F., J. Godwin, E. R. Westwater, Y. Han, S. Keihm, K. Marsh, and R. Ware, Radiometric profiling of temperature, water vapor and cloud liquid water using various inversion methods, Radio Science, 33, 393-404, 1998.

[10] Ameur, Z., Ameur, S., Adane, A., Sauvageot, H., Bara, K., (2004) Cloud classification using the textural features of Meteosat images. International Journal of Remote Sensing, 25, 21, 4491-4503

[11] Feidas H., Giannakos A., (2010) Identifying precipitating clouds in Greece using multispectral infrared Meteosat Second Generation satellite data. Theor Appl Climatol DOI: 10.1007/s00704-010-0316-5

[12] Haralick, R., Shanmugan, K., Dinstein, I., (1973) Texture features for image classification. Transactions on Systems, Man, and Cybernetics, 3, 6, 611-621

[13] Kwon, E., Sohn, B., Schmetz, J., Watts, P., (2010) Intercomparison of height assignment methods for opaque clouds over the tropics. J. Atmos. Sci., 46, 1, 11-19

[14] Strabala, K., Ackerman, S., (1993) Cloud Properties Inferred from 8-12μm Data. J. Appl. Meteor, 33, 212229

[15] Thies, B, Nauss T, Bendix J (2008b) A new technique for detecting precipitation at mid-latitudes during daytime using Meteosat Second Generation SEVIRI. 2008 EUMETSAT Meteorological Satellite Conference, Darmstadt, Germany

[16] Thies, B., Nauss T, Bendix J., (2008a) Discriminating raining from non-raining cloud areas at mid-latitudes using meteosat second generation SEVIRI night-time data. J. Appl. Meteor, 15, 219-230

AUTHORS PROFILE

Kohei Aarai He received BS, MS and PhD degrees in 1972, 1974 and 1982, respectively. He was with The Institute for Industrial Science and Technology of the University of Tokyo from April 1974 to December 1978 and also was with National Space Development Agency of Japan from January, 1979 to March, 1990. During from 1985 to 1987, he was with Canada Centre for Remote Sensing as a Post Doctoral Fellow of National Science and Engineering Research Council of Canada. He moved to Saga University as a Professor in Department of Information Science on April 1990. He was a councilor for the Aeronautics and Space related to the Technology Committee of the Ministry of Science and Technology during from 1998 to 2000. He was a councilor of Saga University for 2002 and 2003. He also was an executive councilor for the Remote Sensing Society of Japan for 2003 to 2005. He is an Adjunct Professor of University of Arizona, USA since 1998. He also is Vice Chairman of the Commission-A of ICSU/COSPAR since 2008. He received Science and Engineering Award of the year 2014 from the minister of the ministry of Science Education of Japan and also received the Bset Paper Award of the year 2012 of IJACSA from Science and Information Organization: SAI. In 2016, he also received Vikram Sarabhai Medal of ICSU/COSPAR and also received 20 awards. He wrote 34 books and published 520 journal papers. He is Editor-in-Chief of International Journal of Advanced Computer Science and Applications as well as International Journal of Intelligent Systsems and Applications. http://teagis.ip.is.saga-u.ac.jp/

Highly Accurate Prediction of Jobs Runtime Classes

Anat Reiner-Benaim
Department of Statistics
University of Haifa
Haifa, Israel

Anna Grabarnick
Department of Statistics
University of Haifa
Haifa, Israel

Edi Shmueli
Intel Corporation
Haifa,
Israel

***Abstract*—Separating the short jobs from the long is a known technique to improve scheduling performance. This paper describes a method developed for accurately predicting the runtimes classes of the jobs to enable the separation. Our method uses the fact that the runtimes can be represented as a mixture of overlapping Gaussian distributions, in order to train a CART classifier to provide the prediction. The threshold that separates the short jobs from the long jobs is determined during the evaluation of the classifier to maximize prediction accuracy. The results indicate overall accuracy of 90% for the data set used in the study, with sensitivity and specificity both above 90%.**

Keywords—Runtime Prediction; Job Scheduler; Server Farms; Classifier; Mixture Distribution

I. INTRODUCTION

Supplying job schedulers with information on how long the jobs are expected to run enabled the development of the backfilling algorithms, which leverage the information to pack the jobs more efficiently and improve system utilization [1]. The backfilling algorithms, however, were designed for parallel systems, in which the jobs require many processors in order to execute, and processor fragmentation (idleness) is a big concern. Thus in parallel system environments the scheduler needs to know the actual runtimes of the jobs (use numeric predictions) to be able to optimize the schedule and improve performance [10].

Our work targets systems in which most jobs are serial, like server farms that are used for software testing. In serial system environments sophisticated scheduling algorithms are not required, and in order to improve performance it is enough to simply separate the short jobs from the long ones, and assign them to different queues in the system [12]. The separation reduces the likelihood that short jobs will be delayed after long ones, improves the average turn-around times of the jobs and overall system throughput (Figure 1).

Respectively, to implement such a system it is enough to only predict the runtime classes of the jobs – whether they will be short or long, in order to assign them to the right queue. On the other hand, any misclassification of the jobs can severely impact performance. For example, mistakenly assigning long jobs to the short jobs queue will cause many of the short jobs to be delayed, average turnaround time to increase, and the overall throughput to decrease as a result.

Motivated by the later usage model (server farms), a method that allows predicting the runtime classes of the jobs with high accuracy was developed.

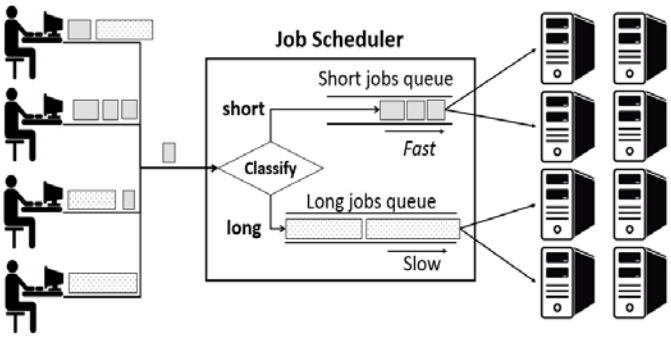


Fig. 1. Separating the short jobs from the long reduces the likelihood that short jobs will be delayed after long ones and improves system performance

The method is based on applying a log transformation on the runtimes of the jobs (historical records), revealing a mixture of two overlapping Gaussian distributions that represent the short and long jobs. We use the mixture model to determine the distribution parameters and to set the initial separation threshold between the short and long runtime populations.

A key design aspect for the proposed method is to be able to predict the classes with high accuracy. In order to achieve high accuracy, the threshold that separates the short jobs from the long is not determined in advance (which can lead to an eventual high misclassification rate). Instead, the threshold is determined as part of the evaluation of the classifier: a subset of the data that is close to the means of the distributions is used for training the classifier, and then the full dataset is used to select the threshold that optimizes a desired target function.

For class prediction for newly incoming jobs, the CART classifier is used [18]. CART is suitable for binary classification and can account for both continuous and categorical classifying variables, and is based on a tree optimizing algorithm that minimizes classification error while reducing overfitting by branch pruning.

The proposed method was applied on a job trace obtained from one of Intel's server farms, and which contained more than one million job records. Setting the target on achieving the best trade-off between misclassifications of short jobs and misclassifications of long jobs resulted in prediction accuracy of 90% (total misclassification rate of 10%) on the independent validation set. The predictions were based on estimated distribution means of 140 and 3,500 seconds for the "short" and "long" classes, respectively, and a separating threshold of 608 seconds.

This paper is organized as follows. Section 2 describes the data that is used for training, testing and validating the model. Section 3 describes the initial class labeling based on the mixture model analysis. Section 4 reviews the CART model. Section 5 describes the learning algorithm along with the optimal threshold determination procedure for best accuracy. Section 6 describes the results of the study. Section 7 surveys related work and Section 8 concludes the paper.

II. THE DATA

Our data is based on two traces obtained from one of Intel's server farms. The first trace, which was used to train the model, contained a sample of around one million job records that executed in the farm during a period of ten consecutive days. The second trace, which was used to validate the model, contained a sample of additional 755,000 records (approximately) of jobs that executed during a period of seven consecutive days. The validation on independent data is important for establishing the robustness of the model obtained in the training stage.

Each record in the traces contained 13 fields pertaining to a particular job. The continuous variables: "Submitime", "Starttime" and "Finishtime" indicated when the job was submitted, when the job started, and when the job finished executing, respectively. In order not to reveal any information about the workload, the traces did not contain any descriptive information about the jobs. Instead, the values in the fields were transformed into discrete values (categorical variables) that can be used for the analysis. In addition, the names were also transformed in order not to reveal information about the possible meanings of the values.

Table I groups the 9 categorical variables and roughly explains the meaning of each group. Table II outlines basic statistics on each of the variables.

TABLE I. ROUGH GROUPING OF THE 9 CATEGORICAL VARIABLES

Group	tab	Relates to	Example
A	3	Scheduling information	Resources requested by the job
B	2	Execution-specific information	Command line and arguments
C	4	Association information	Project and component

TABLE II. STATISTICS REGARDING THE CATEGORICAL VARIABLES

Variable	# of categories	# of missing (in training data)
A1	9	0
A2	7	0
A3	5	0
B1	44	173
B2	22	184
C1	2	0
C2	5	239
C3	6	184
C4	32	0

In addition to the above, two additional categorical variables were defined, day and hour, based on the three continuous variables in the trace. These variables indicate the day of the week (1 for Sunday to 7 for Saturday) and the hour of the day (0 to 23), the job was submitted, started, and finished executing, respectively. Figure 2 shows the distribution of the respective temporal categorical variables along the timeline axis. As can be seen, during weekdays longer jobs are typically submitted during the morning hours, with occasional peaks in runtime during evening hours. During weekends, peaks in runtime also occur during the afternoon and evening hours.

Figure 3 shows the distribution of the jobs runtime "as-is", and after applying a log base 2 transformation on the runtime. As can be seen in Figure 3a, the vast majority of the jobs are short (the shortest job ran 3.5 seconds), and there are few long ones (the longest job ran for nearly 9 days). This well corresponds to previous observations made on the runtime, describing a phenomenon that characterizes many production workloads [11].

Transforming the runtime to the log scale (Figure 3b) reveals a mixture pattern of two main Gaussian-like distributions (some-times referred to as a "hyper lognormal distribution), with a stretching right tail.

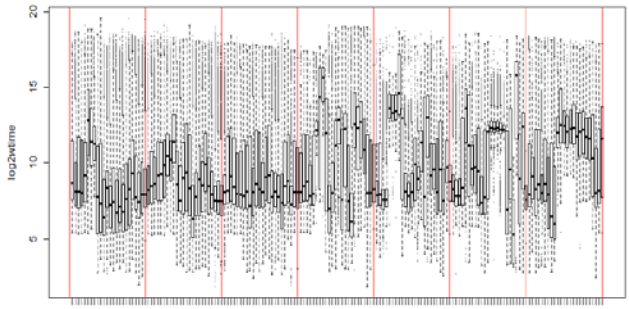


Fig. 2. Runtime boxplots (in log base 2 scale) along time (Sunday through Saturday). The days are separated by vertical red lines. Each tick mark along the time axis marks an hour of the day

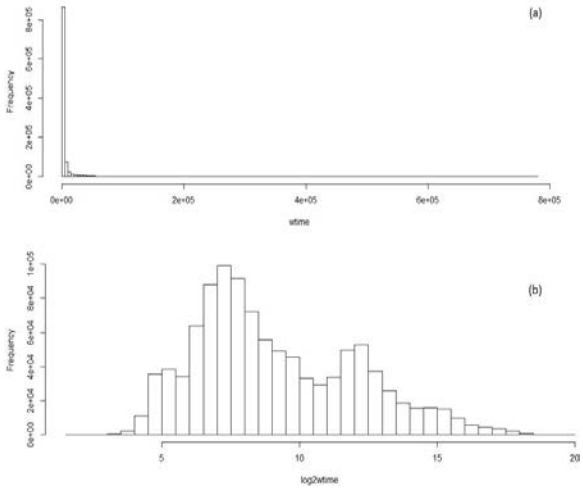


Fig. 3. Histogram of runtime. (a) Raw data (b) After log base 2 transformation

III. Class Construction by Mixture Distribution Analysis

The first component in the proposed analysis sets the base for defining the two runtime classes by estimating the mixture distribution parameters, and then labeling each job as "short" or "long". Once the predictor variables are selected through a training-and-testing algorithm (see Section 5.1), the model is optimized by selecting the mixture threshold which provides the best performance, namely minimizes the prediction error or approaches the desired sensitivity-specificity combination (see Section 5.2).

The Gaussian (normal) mixture model has the form

$$f(x) = \sum_{m=1}^{M} \alpha_m \phi(x;\ \mu_m, \Sigma_m)$$

with mixing proportions α_m, $\sum_m \alpha_m = 1$, and each Gaussian density has a mean μ_m and covariance matrix Σ_m. The parameters are usually fit by the maximum likelihood approach using the EM algorithm, which is a popular tool for simplifying difficult maximum likelihood problems.

Based on the mixture observed in Figure 2, one to four mixture components may be defined. However, for the purpose of the current study it was decided to focus only on two classes, namely "short jobs" and "long jobs". Thus the runtime Y is modeled as a mixture of the two normal variables

$$Y_1 \sim N(\mu_1, \sigma_1^2), \qquad Y_2 \sim N(\mu_2, \sigma_2^2).$$

Y can be defined by

$$Y = (1 - \Delta) \cdot Y_1 + \Delta \cdot Y_2,$$

where $\Delta \in \{0, 1\}$ with $\mathbb{P}(\Delta = 1) = \pi$. This generative representation is explicit: generate a $\Delta \in \{0, 1\}$ with probability π, and then depending on the outcome, deliver either Y_1 or Y_2. Let $\phi_\theta(x)$ denote the normal density with parameters $\theta = (\mu, \sigma^2)$. Then the density of Y is

$$g_Y(y) = (1 - \pi)\phi_{\theta_1}(y) + \pi\phi_{\theta_2}(y).$$

Suppose the model is fit to the data by maximum likelihood. The parameters are

$$\theta = (\pi, \theta_1, \theta_2) = (\pi, \mu_1, \sigma_1^2, \mu_2, \sigma_2^2).$$

The log-likelihood based on N training cases is

$$l(\theta;\ \mathrm{Z}) = \sum_{i=1}^{N} \log\left[(1 - \pi)\phi_{\theta_1}(y_i) + \pi\phi_{\theta_2}(y_i)\right].$$

Direct maximization of $l(\theta;\ \mathrm{Z})$ is quite difficult numerically due to the sum of terms inside the logarithm. There is, however, a simpler approach. We consider unobserved latent variables Δ_i taking values 0 or 1 as earlier: if $\Delta_i = 1$ then Y_i comes from distribution 2, otherwise Y_i comes from distribution 1. Suppose the values of the Δ_i's are known. Then the log-likelihood would be

$$l(\theta;\ \mathrm{Z}, \Delta) = \sum_{i=1}^{N} \left[(1 - \Delta_i)\log\phi_{\theta_1}(y_i) + \Delta_i \log\phi_{\theta_2}(y_i)\right] + \sum_{i=1}^{N} \left[(1 - \Delta_i)\log\pi + \Delta_i \log(1 - \pi)\right]$$

and the maximum likelihood estimates of μ_1 and σ_1^2 would be the sample mean and the sample variance of the observations with $\Delta_i = 0$. Similarly, the estimates for μ_2 and σ_2^2 would be the sample mean and the sample variance of the observations with $\Delta_i = 1$.

Since the Δ_i values are actually unknown, the procedure continues in an iterative fashion, substituting for each Δ_i in the previous equation its expected value

$$\gamma_i(\theta) = \mathbb{E}(\Delta_i|\theta, \mathrm{Z}) = \mathbb{P}(\Delta_i = 1|\theta, \mathrm{Z}),$$

which is also called the responsibility of model 2 for observation i.

We use the following procedure, known as the EM algorithm, for the two-component Gaussian mixture:

1) Take initial guesses for the parameters $\hat{\pi}, \hat{\mu}_1, \hat{\sigma}_1^2, \hat{\mu}_2, \hat{\sigma}_2^2$ (see below).

2) Expectation step: compute the responsibilities

$$\hat{\gamma}_i = \frac{\hat{\pi}\phi_{\hat{\theta}_2}(y_i)}{(1 - \hat{\pi})\phi_{\hat{\theta}_1}(y_i) + \hat{\pi}\phi_{\theta_2}(y_i)}, \qquad (1)$$

$$i = 1, 2, \ldots, N.$$

3) Maximization step: compute the weighted means and variances,

$$\hat{\mu}_1 = \frac{\sum_{i=1}^{N}(1 - \hat{\gamma}_i)y_i}{\sum_{i=1}^{N}(1 - \hat{\gamma}_i)}, \qquad \hat{\sigma}_1^2 = \frac{\sum_{i=1}^{N}(1 - \hat{\gamma}_i)(y_i - \hat{\mu}_1)^2}{\sum_{i=1}^{N}(1 - \hat{\gamma}_i)},$$

$$\hat{\mu}_2 = \frac{\sum_{i=1}^{N}\hat{\gamma}_i y_i}{\sum_{i=1}^{N}\hat{\gamma}_i}, \qquad \hat{\sigma}_2^2 = \frac{\sum_{i=1}^{N}\hat{\gamma}_i(y_i - \hat{\mu}_1)^2}{\sum_{i=1}^{N}\hat{\gamma}_i},$$

and the mixing probability,

$$\hat{\pi} = \frac{\sum_{i=1}^{N}\hat{\gamma}_i}{N}.$$

4) Iterate steps 2 and 3 until convergence.

In the expectation step, a "soft" assignment of each observation to each model is done: the current estimates of the parameters are used to assign responsibilities according to the relative density of the training points under each model. In the maximization step, the responsibilities are used within weighted maximum-likelihood fits to update the estimates of the parameters.

A simple choice for initial guesses for $\hat{\mu}_1$ and $\hat{\mu}_2$ is two randomly selected observations y_i. The overall sample variance $\sum_{i=1}^{N}\frac{(y_i - \bar{y})^2}{N}$ can be used as an initial guess for both $\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$. The initial mixing proportion $\hat{\pi}$ can be set to 0.5.

The "mixtools" R package [15, 16] was used for the mixture analysis, with the function "normalmixEM" for parameter and posterior probability (responsibility) estimation.

IV. The CART Model

The CART (Classification and Regression Trees) model, also named the decision tree model, is an approach for making either quantitative or class prediction. The CART model is non-parametric, thus no assumptions are made regarding the underlying distribution of the predictor variables, enabling CART to handle numerical data that are skewed or multi-

modal. Both continuous and categorical predictors can be considered, including ordinal ones.

CART identifies classifying, or "splitting", variables based on an exhaustive search of all classifying possibilities with the available variables. Useful CART trees can be generated even when there are missing values for some variables, by using "surrogate" variables, which contain information similar to the missing variables.

CART analysis consists of the following steps:

- Tree building, during which a tree is built using recursive splitting of nodes. This process stops when a maximal tree has been produced. The higher the splitter variable in the tree, the higher its importance in the prediction process.
- Tree "pruning", which is a simplification of the tree by cutting nodes off from the maximal tree.
- Optimal tree selection, which selects one tree from the set of pruned trees with the least evidence of over fit.

Each path from the root of a decision tree to one of its leaves can be transformed into a rule. Less complex decision trees are preferred, since they are easier for interpretation and may be more accurate.

V. Model Learning and Optimization Procedure

Once labeled data is obtained, a supervised learning technique is used for the purpose of generating a classification rule. In the first stage a set of variables that will be included in the model are selected, while evaluating the performance of each model, and in the second stage a final model is obtained by using the selected variables on the full dataset and evaluate the model based on the performance target function.

This stage of the analysis is done on a subset of the training data (containing the ten days period), which is extracted as follows. Since the two observed runtime distributions overlap, observations that are within 0.5 standard deviations off the two means are selected, such that they will be distant from the overlapping region and will belong to the corresponding classes with high certainty (Figure 4). A total of 257,467 observations labeled short=1 (belonging to the Gaussian population with the smaller mean) and 192,205 observations with short=0 (belonging to the Gaussian population with the larger mean) were selected. Together they made around 43% of the data.

A five-fold cross-validation procedure is then performed in order to select variables and evaluate each model, by iteratively dividing the data in random into a training set (80% of the learning data) and an evaluation set (20% of the learning data) and implementing the CART model with the mixture threshold of 0.5. The importance measure, which considers how high the splitting variable is in the tree, is averaged across all iterations for each variable, and the variable having an importance score above the baseline level is selected.

Once a set of variables is selected for each model type, a model is fit to the full training data. We account for the two types of misclassification error, the false "positive" classification and the false "negative" classification. Defining classification into "short" as "positive", the former refers to the erroneous classification of a long job into the "short" class, and the latter refers to the erroneous classification of a short job into the "long" class.

In the job runtime context, sensitivity is defined as the proportion of short jobs classified as short, while the specificity is defined as the proportion of long jobs classified as long. Subtracting the specificity from 1 will give the proportion of long jobs erroneously classified as short. For the CART classifier, the mixture threshold that yields the best tradeoff between the two errors is chosen. The full set of sensitivity-specificity combinations can be summarized in a pseudo-ROC (Receiver Operating Characteristic) curve, in which the sensitivity is plotted against 1-specificity, for each threshold of the probability obtained in the mixture model (the final value obtained for equation 1).

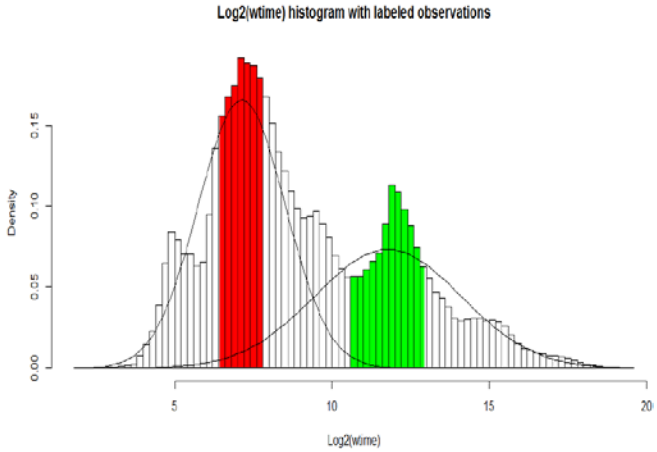


Fig. 4. Runtime (in log base 2 scale) density. The red and green colored regions mark the observations selected for the learning process

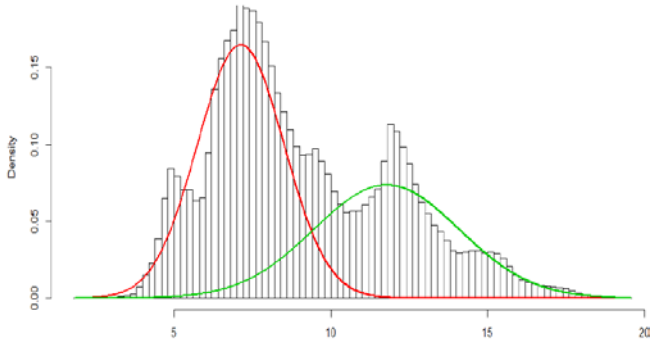


Fig. 5. Density estimates obtained by mixture analysis for the two families underlying the runtime distribution (on the log base 2 scale). The red line marks the estimated density for the “short” class, while the green line marks the estimated density for the “long” class

A model performing a perfect discrimination has an ROC curve that passes through the upper left corner (100% sensitivity, 100% specificity). Therefore the closer the ROC curve to the upper left corner, the higher the overall accuracy of the test. Yet, the consequences, or costs, of each type of error may vary among applications and among policy makers. Thus the optimal threshold may allow higher weight to one of the errors on account of the other.

VI. RESULTS

A. Class definition

Implementing the mixture model clustering approach, two Gaussian families underlying the runtime distribution (on the log base 2 scale) were defined. The density estimates are super imposed on the runtime density in Figure 5. The parameters for each family and the mixing proportions are detailed in Table III.

The mixture analysis also yields the posterior probability, as defined by equation (1), for each observation to belong to the "short" class. For a probability threshold of 0.5, 631,059 observations (nearly 60%) are classified into the "short" class, while 411,053 observations are classified into the "long" class. Once a classifier is found (see the next subsection), the threshold is refined to optimize the sensitivity-specificity tradeoff.

B. Classifying by CART

Applying the CART classifier on the training data through the cross-validation procedure, six variables obtained high importance scores (Figure 6). The classifier achieved a total misclassification error of 3.5%.

TABLE III. DENSITY PARAMETER ESTIMATES OBTAINED BY THE MIXTURE ANALYSIS

	First Gaussian Family		Second Gaussian Family	
Mixing Proportion	0.57		0.43	
Mean	7.13	$2^{7.13}$	11.78	$2^{11.78}$
Standard Deviation	1.38	$2^{1.38} = 2.60\ sec$	2.32	$2^{2.32} = 4.99\ sec$

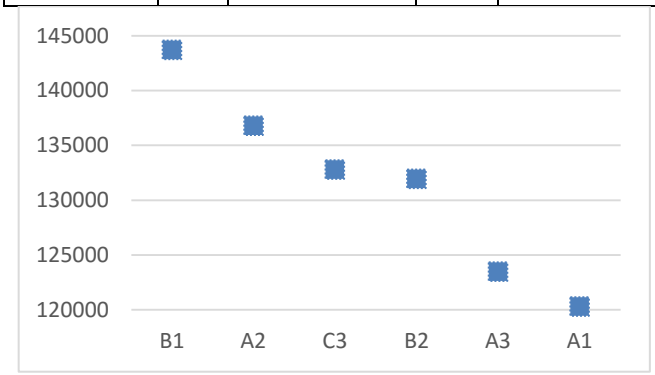


Fig. 6. Top ranking importance scores for the CART model, averaged across 150 cross-validation iterations

A model containing the six selected variables was fit to the full dataset for a series of class mixture threshold. The pseudo-ROC in Figure 7 presents the sensitivity-specificity combinations obtained for a set of threshold probability used for the mixture distribution. The best performing model is the one using the threshold of 0.45, which achieves sensitivity of 92.5% and specificity of 91.1%, with a total misclassification error of 8.9%. This threshold corresponds to a runtime of 9.25 on the log scale, or 608 seconds on the original scale.

The selected model yielded a tree containing four of the six variables that were tried (Figure 8). The total misclassification rate was 8.08%. Implementing the obtained tree on the validation data resulted in a total misclassification error of 9.17%, with specificity of 91.5% and sensitivity slightly beyond 90%.

VII. RELATED WORK

Supplying the scheduler with information on how long the jobs are expected to run has always been a challenging task. In general, two approaches were used estimating runtime. The first is to ask the users to supply the information, and the other is to try and predict the runtimes automatically using historical data on jobs that have already completed.

Asking the users to estimate the runtimes has been shown to be highly inaccurate, as users tend of overestimate the runtimes in order to prevent the scheduler from killing their jobs [1]. Furthermore, Tsafrir et al. [2] has observed that the users further tend to "round" the estimates, thereby limiting the scheduler's ability to optimize the schedule. Bailey et al. [7] have shown that users are quite confident of their estimates, and that most likely they will not be able to improve much the accuracy of their estimates.

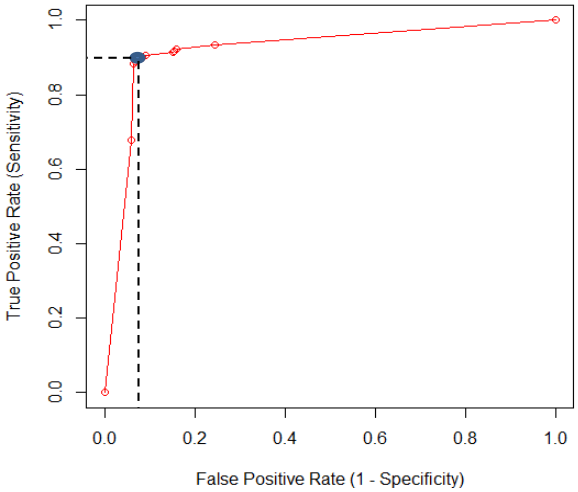


Fig. 7. The pseudo-ROC curve obtained by the CART classifier for the full training data. The blue circle marks the optimal tradeoff between sensitivity and specificity (enhanced by the dashed lines), obtained for mixture probability threshold of 0.45

Predicting the runtimes automatically is therefore the default alternative. This is usually composed of two steps:

1) Identifying classes of similar jobs within the historical jobs records, and

2) Using the aforementioned classes to predict the runtimes for newly submitted jobs.

Gibbons [4] and Downey [5] classified the jobs using a statically defined set of attributes, e.g. user, executable, queue, etc. For newly submitted jobs, Gibbons used the 95th percentile of the runtimes in the respective class, while Downey used a statistical model that was based on a log-

uniform distribution of the runtimes in order to provide the prediction.

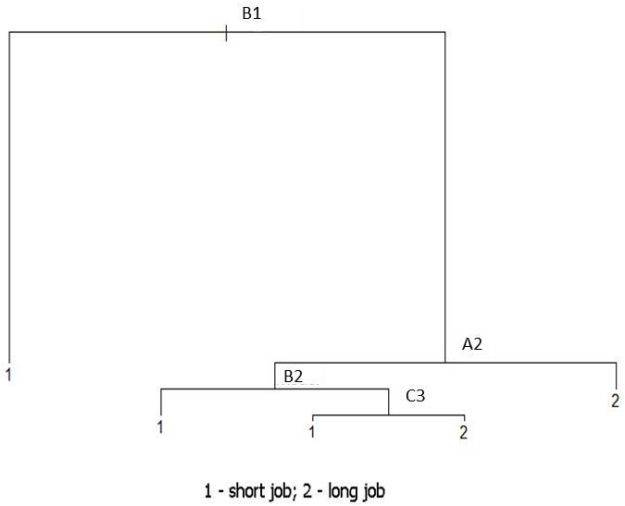


Fig. 8. The final tree obtained on the full data

Smith et al. [6] suggested the use of genetic algorithms to refine the selection of attributes used for the classification, and achieved up to 60% improvement in accuracy compared to the static approaches. Respectively, Kapadia et al. [8] used instance-based learning and Krishnaswamy et al. [9] applied rough-set theory. Finally, Tsafrir et al. [10] showed that complicated prediction techniques may not be required if the scheduling algorithm itself can be modified, and suggested to average the last two jobs by the same user in the history.

These techniques, however, were mainly designed for parallel systems, in which the scheduler needs to the actual runtimes of the jobs (use numeric predictions) to be able to optimize the schedule. For the server farm usage model which is targeted in the current work, sophisticated scheduling algorithms are not required, and the paper shows that it is enough to simply separate the jobs into short and long in order to improve performance [12]. Our work provides the facility to perform the separation and hence forms the basis for enabling such systems.

VIII. Discussion and Conclusions

Predicting the runtimes of jobs using actual numeric values is of high importance for parallel systems. Here, fragmentation is a big concern, and in order to minimize it (namely to fill the holes in the schedule), it is important for the scheduler to know the exact runtime of the jobs. For other types of systems like server farms used for software testing, it is enough to only predict the runtime classes of the jobs e.g., short or long, in order to send the jobs to the right queue and improve performance.

Motivated by the later usage model, a method that facilitates highly accurate prediction of the job runtime class was developed. The method leverages the fact that the runtimes may be represented as a mixture of two or more distributions, in order to train a classifier that will be used to predict the runtime classes of newly incoming jobs. In order to achieve high accuracy, the threshold that separates the short jobs from the long is determined during the evaluation of the classifier. In a real system the threshold can be periodically communicated to the scheduler to help deciding on the right allocation of resources for the different job classes.

The work presented is based on a single data set that was obtained, and included validation on data independent of the training data. Yet, in spite of its promising results, additional testing is required on more data sets in order to establish complete confidence in the robustness of the proposed method. However, due to the size of the data (over one million jobs), and the fact that the mixture distribution is known to be evident in the many real-world workloads, there is a strong reason to believe that with small adjustments e.g., to the number of classes, the proposed method can be tuned to sustain the workloads as well.

A uniqueness of the proposed approach is in conducting an initial step that is aimed to define runtime classes based on the empirical runtime distribution. While other existing approaches for predicting job runtime do not used categorization into classes but rather estimate the runtime numerically, the end result of all approaches is in an overall system efficiency. If efficiency can be defined and measured reliably, it can serve to compare the ultimate performance of all methods. Such assessment is under design and is planned to be conducted as a next step of the presented study.

Acknowledgment

We thank Prof. Dror G. Feitelson of the Hebrew University, Israel, for the useful comments, Evgeni Korchatov from Intel for helping to obtain and prepare the data and Eran Smadar from Intel for supporting this work.

References

[1] Feitelson, D.G., Mu'alem Weil, A.: Utilization and predictability in scheduling the IBM SP2 with backfilling. In 12th IEEE Int'l Parallel Processing Symp. (IPPS), 542–546 (1998)

[2] Tsafrir, D., Etsion, Y., Feitelson, D.G.: Modeling User Runtime Estimates, Job Scheduling Strategies for Parallel Processing (JSSPP), 1-35 (2005)

[3] Lawson, B.G., Smirni, E., Puiu, D.: Self-Adapting Backfilling Scheduling for Parallel Systems. Proceedings of the 2002 International Conference on Parallel Processing (ICPP), 583–592 (2002)

[4] Gibbons, R.: A Historical Application Profiler for Use by Parallel Schedulers. Job Scheduling Strategies for Parallel Processing (JSSPP), 58-77 (1997)

[5] Downey, A.B.: Predicting Queue Times on Space-Sharing Parallel Computers. Proc. 11th IEEE Int'l Parallel Processing Symp. (IPPS), 209-218 (1997)

[6] Smith, W., Foster, I., Taylor, V.: Predicting application run times using historical information. In the 4th Workshop on Job Scheduling Strategies for Parallel Processing (JSSPP), 122–142, Lect. Notes Comput. Sci. vol. 1459 (1998)

[7] Lee, C.B., Schwartzman, Y., Hardy, J., Snavely, A.: Are user runtime estimates inherently inaccurate? Job Scheduling Strategies for Parallel Processing (JSSPP), 253–263 (2004)

[8] Kapadia, N.H., Fortes, J.A.B., Brodley, C.E.: Predictive Application-Performance Modeling in a Computational Grid Environment. Proc. IEEE Int'l Symp. High Performance Distributed Computing (HPDC), 6 (1999)

[9] Krishnaswamy, S., Loke, S.W., Zaslavsky, A.: Estimating Computation Times of Data-Intensive Applications. IEEE Distributed Systems Online, vol. 5, no. 4 (2004)

[10] Tsafrir, D., Etsion, Y., Feitelson, D.G.: Backfilling using system-generated predictions rather than user runtime estimates. IEEE Trans. Parallel & Distributed Syst. 18(6), 789-803 (2007)

[11] Feitelson, D.G.: Metrics for mass-count disparity. In 14th Conf. Modeling, Analysis, and Simulation of Comput. and Telecomm. Syst., 61-68 (2006)

[12] Harchol-Balter, M., Crovella, M., Murta, C.: On Choosing a Task Assignment Policy for a Distributed Server System. IEEE Journal of Parallel and Distributed Computing (JPDC), vol. 59, no. 2, 204-228 (1999)

[13] Hastie, T., Tibshirani R., Friedman, J.: The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer (2001)

[14] Maimon, O., Rokach, L.: The Data Mining and Knowledge Discovery Handbook, Springer, XXXVI (2005)

[15] Benaglia, T., Chauveau, D., Hunter, D.R., Young, D.S.: Mixtools: An R Package for Analyzing Finite Mixture Models, Journal of Statistical Software, Vol. 32, No. 6, 1-29 (2006)

[16] Benaglia, T., Chauveau, D., Hunter, D.R., Young, D.S, Elmore, R., Hettmansperger, T., Thomas, H., Xuan, F.: Package 'Mixtools' - Tools for Analyzing Finite Mixture Models, Repository CRAN (2014)

[17] Dudoit, S., Fridlyand, J., Speed, T.P.: Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data. Journal of the American Statistical Association Vol. 97, No. 457, 77-87 (2002)

[18] Lewis, R.J.: An Introduction to Classification and Regression Tree (CART) Analysis. Presented at the 2000 Annual Meeting of the Society for Academic Emergency Medicine in San Francisco, California (2000)

A Novel Approach for Discovery Quantitative Fuzzy Multi-Level Association Rules Mining Using Genetic Algorithm

Saad M. Darwish
Department of Information Technology
Institute of Graduate Studies and Research University of Alexandria
Alexandria, Egypt

Abeer A. Amer
Department of Computer and Information Systems
Sadat Academy for Management Sciences (SAMS)
Alexandria, Egypt

Sameh G. Taktak
Department of Computer and Information Systems
Sadat Academy for Management Sciences (SAMS)
Alexandria, Egypt

***Abstract*—Quantitative multilevel association rules mining is a central field to realize motivating associations among data components with multiple levels abstractions. The problem of expanding procedures to handle quantitative data has been attracting the attention of many researchers. The algorithms regularly discretize the attribute fields into sharp intervals, and then implement uncomplicated algorithms established for Boolean attributes. Fuzzy association rules mining approaches are intended to defeat such shortcomings based on the fuzzy set theory. Furthermore, most of the current algorithms in the direction of this topic are based on very tiring search methods to govern the ideal support and confidence thresholds that agonize from risky computational cost in searching association rules. To accelerate quantitative multilevel association rules searching and escape the extreme computation, in this paper, we propose a new genetic-based method with significant innovation to determine threshold values for frequent item sets. In this approach, a sophisticated coding method is settled, and the qualified confidence is employed as the fitness function. With the genetic algorithm, a comprehensive search can be achieved and system automation is applied, because our model does not need the user-specified threshold of minimum support. Experiment results indicate that the recommended algorithm can powerfully generate non-redundant fuzzy multilevel association rules.**

Keywords*—Quantitative Data Mining; Fuzzy Association Rule Mining; Multilevel Association rule; Optimization Algorithm

I. INTRODUCTION

In data-mining, discovering association rules in transaction databases is frequently examined. Association rules are widely offered and are beneficial for planning and marketing. For example, they can be managed to implicate supermarket officials of what products the customers have an inclination to purchase together. Taking market basket analysis as an example, the mining problem can be explained as given a database *D* of transactions, each transaction is a set of items; find all rules that relate the carriage of one set of items with that of another set of items [1].

The classical algorithms for mining association rules are formed on binary attributes databases, which have two weaknesses. Firstly, it cannot treat quantitative attributes; secondly, it handles each item with the same weight despite that strange item may have different importance. Also, a binary association rule bears from explicit boundary problems. Besides many real world transactions consist of quantitative attributes. That is why numerous researchers have been serving on the generation of association rules for quantitative data [2] [3].

Beginning approaches for quantitative association rule mining manages distinctive partitioning for transforming other attributes to binary ones which convey from a major problem that results in information damage because of sharp limits. In other words, modern algorithms neglect or exaggerate items beside the boundary. When distributing an attribute in the data into sets comprising individual ranges of values, the users are faced with the sharp boundary problem[2]. Quantitative attributes are discretized by joining the concept of hierarchies. This manner occurs before the event of mining. For example, a concept hierarchy for age may be adopted to reconstruct the initial numeric values of this attribute by ranges [2]. To surmount this problem researchers are laboring for mining association rules for quantitative attributes. They have contributed several algorithms that tackle quantitative algorithm and reveal how they dispense with quantitative data [3] [4].

In general, fuzzy technique overcomes the main drawback of the discretize technique. Fuzzy logic produces linguistic term instead of intervals which is more nearer to the human mind. The disadvantage is that although the loss of information is small but it exists. Furthermore, the needs for fuzzy membership function to be given by an expert, which is not always straightforward and can be biased. Despite that, fuzzy association rule mining approach appeared out of the requirement to mine quantitative data uniformly present in databases efficiently [2]. There are two essential basic criteria for association rules, support(*s*) and confidence(*c*). Since the database is large and users interest about only those frequently purchased items, usually thresholds of support and confidence are predetermined by users to separate those rules that are not so attractive or beneficial. The two thresholds are called minimal support and minimal confidence respectively [5].

Genetic Algorithm (GA) is a heuristic exploration that imitates the process of natural evolving. This heuristic is

routinely applied to produce valuable explications to optimization and search obstacles. Genetic Algorithm is based on conceptions of evolution hypothesis as a fundamental policy is that only the strongest beings remain. The genetic algorithms are significant when identifying association rules because they work with the global search to determine the set of items frequency and they are less complex than other algorithms frequently worked in data mining. The genetic algorithms for discovery of association rules have been settled into usage in real problems such as commercial databases and fraud detection [6].

Earlier investigations on data mining directed on locating association rules at a single-concept level. Mining association rules at multiple concept levels may guide to the discovery of more broad and significant knowledge from data. Related item taxonomies are normally predefined in real-world purposes and can be interpreted as hierarchy trees. Terminal nodes on the trees express actual items looking in transactions; internal nodes describe classes or concepts built from lower-level nodes [7]. A simple example is given in Fig. 1. Mining multi-level association rules are motivated by several purposes, such as: (a) the multi-level association rules are more reasonable and are more interpretable for users. (b) The multi-level association rules can supply us solutions for the undesirable and undesired rules. Encouraging applications involve spatial data analysis, emergency event analysis, and network data mining [8] [9].

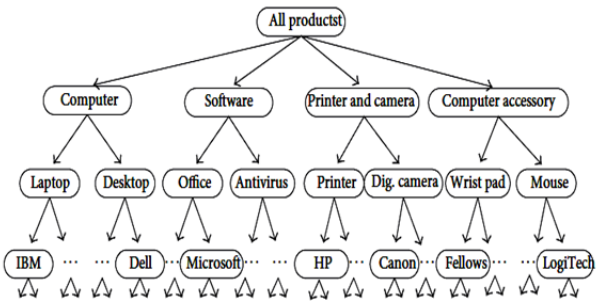


Fig. 1. The predefined taxonomy

A. Motivation & Rationale

If it arrives at quantitative association rule mining, a number of trials have been directed on performance (speed) and effectiveness (number of rules). Less effort has been converged on quality. Modern quantitative multilevel association rule mining algorithms depend on intense looks of the database to obtain regular exemplars beyond different abstraction levels [9]. However, these mining algorithms are often based on the postulate that users can blueprint the minimum support relevant to their databases [10]. Mining quantitative association rules is not a manageable enlargement of mining categorical association rules. Since the search space is unlimited, our aim is to detect a measurable set of exciting solutions (quantitative rules), near to the optimal answers. This illustrates why we have decided to solve this search problem with meta-heuristics routines, mainly genetic algorithms [11].

Regularly, when managing quantitative association rule mining several rules can be identified or inferred and confuse the user. But more importantly, some of these rules could be redundant and produce no new knowledge. Some attempt has been aimed at selling with redundant rules in flat datasets. However, datasets can have a hierarchy/taxonomy or multiple concept levels and thus redundancy in these datasets require to be adjusted. This subject is one of the phases of this research. Currently, the approach being taken is to resolve which rules are redundant and eliminate them, thus diminishing the number of rules a user has to deal with while not decreasing the information content [12]. This Paper offers an adapted version of the Apriori algorithm for mining fuzzy multi-level association rules in large databases for locating familiar item set at distinct levels of abstraction [13].

B. Research Problems

The purpose of this study is to offer to the field of data mining and in precise to multilevel quantitative fuzzy association rule mining. Hence to attain this distinct and well-defined investigation, difficulties are needed. By answering these it is likely to provide in an essential style. For our research, there are many central difficulties that we will converge on and strive to solve. These difficulties are:

- Boolean attributes can be studied as an exceptional instance of categorical attributes and it is an almost applicant to generalize the Boolean data mining algorithms. For qquantitative attributes, despite, the state is not so easy. We either have to somehow convert the quantitative association rules problem into Boolean one or to get different algorithms. Here we shall, in fact, produce an approach to discover quantitative association rules stemmed from a dataset with multiple concept levels.
- The number of rules expands exponentially with the number of items. But this complexity is undertaken with some advanced algorithms which can efficiently cut the search space [1][4]. The work picking this problem principally assists the user when scanning the rule set. Yet, the evolution of further valuable quality measures on the rules through employing genetic algorithm with fitness function (relative confidence of the association rule) to affirm the most intriguing association rules signifies the advanced aims to solve this problem [14].
- Without a priori knowledge, however, ascertaining the right intervals (discretize) for quantitative data mining can be a complex and intricate task. Moreover, these intervals may not be compact and acceptable enough for human experts to quickly gain nontrivial knowledge from those created rules. Fuzzy membership function can help to advise this problem.
- Completely and efficiently identifies no redundant association rules from datasets with a hierarchy.

C. Problem Statement

Market data in real-world usually involve quantitative values, so creating an advanced data-mining algorithm equipped to contract with quantitative data grants a challenge to workers in this study domain [4]. The multilevel association rules mining problem can be defined as follows: there are a

collection of items $I = \{i_1, i_2, ..., i_n\}$ and Γ is a classification tree that concisely clarifies the multilevel categorizing relations among items as the field awareness. i_1 is the ancestor of i_2 and i_2 is the descendant of i_1 if there is an edge in Γ from i_1 to i_2. Only leaf nodes are displayed in the database. *D* is a database of transactions where each transaction *T* in *D* is a set of items such that $T \subseteq I$. Each transaction is attached with an identifier T_{ID}. Let *P* indicates the set of positive integers, and I_v symbolizes the set $I \times p$. A couple $< x, v > \in I_v$ means the quantitative attribute *x*, with the related value *v*.I_v $\{\forall < x, l, u > \in I \times p \times p \mid l \le u\}$, l is the lower limit and u is the upper limit of p. A triple $< x, l, u > \in I_v$ stands for a quantitative x with a value in the interval [*l, u*]. Note that a transaction *T* holds an item $x \in I$ if x is in *T* or x is an ancestor of some items in *T*. In addition, a transaction *T* involves $X \subseteq I$ if *T* bears every item in X.

A multilevel association rule is an inference of the form $X \Rightarrow Y$, where $X \subseteq I, Y \subseteq I,$ and $X \cap Y = \phi$. No item in *Y* is an ancestor of any item in *X*; that is $Y \cap \text{ancestors(x)} = \phi$ .This is because a rule of the style "$x \Rightarrow \text{ancestors(x)} = \phi$" is slightly true with 100% confidence, which is redundant. Both *X* and *Y* can contain items from any level of Γ [9][15][16].

However, there are still some limitations in quantitative association rule mining, such as [3][4][10]: (1) separation of the quantitative attribute, which is adopted in the design, is not accessible for all attribute and every user. (2) users, and even experts, regularly believe difficult to provide those thresholds like the minimum support, the interest level, and the minimum confidence.(3) the search space might be very large when we contact with quantitative attributes. Finally, (4) the rules declared by the algorithm might be too many to manage with.

II. Literature Review

In this section we compare the quantitative association rule mining algorithms taking into account the form of the rules and discuss each technique advantages, disadvantages and what kind of database can be used[17][18]:

1) Discretization:

The elementary intention of this routine is to transform quantities data to Boolean by examining a separation of the numerical attributes into collections of intervals. Then, an algorithm for detecting Boolean association rules can be handled to prepare quantities rules. Two main representations of partitions are included. A fixed partition, where the assortments of intervals are disjoint and another type, where the ends of intervals are overlaid with each other.

The principal benefit of this technique, beyond being the first work done on this track, is that manipulating both categorical and numerical data correspondingly. However, situations (disjoint or overlapped) yield problems; disjoint sets damage from *Min_Sup* and *Min_Conf* thresholds and overlapped sets suffer from the cutting boundary problem. Using intervals rather than the real continuous data will inevitably result in a loss of information. The rules we make will be only an estimation of the best results. Another problem is the enlargement of the attributes dimension; the problem here is the need for more memory and time to treat these data.

2) Adjusted difference analysis:

This algorithm is based on engaging both adjusted difference analysis and discretization to discover rules between two attributes. The two attributes could be any mixture of numerical or categorical. This technique has the capability to identify positive and negative association rules and does not need any user thresholds (support and confidence). Its advantages are that it does not want any user thresholds and it has the talent to obtain a new significant objective measure of the association rules. The disadvantage of this technique is the problems of discretization as in the first technique. Also, this technique is obviously considered to be as generating a special case rule since the generated rules are always between two attributes only.

3) Fuzzy Approach based on integrating fuzzy set and fuzzy logic concepts with Apriori algorith:

It reforms numerical data into fuzzy member between [0,1] with membership function; then operate with the fuzzy member with an adjusted Apriori technique that can manipulate comfortably the extracted rules, which are stated in linguistic terms. These approaches are based on the fuzzy additions to the classical association rules mining by establishing support and confidence of the fuzzy rule. While the mining results are straightforward to interpret by human operators, two shortcomings still insist on implementing such fuzzy approaches to the original problems. One is the computational time for mining from the database, and the other is the precision of deduced rules. More formal description, as well as a survey of the existing methods of quantitative association rule mining, can be found in [20].

In the literature, several researchers have concentrated on fuzzy multilevel association rules mining [3][21-25]. Some of these methods evoked multilevel membership functions by ant colony systems and genetic algorithm without stipulating the actual minimum support. To improve the performance of computing, setting the functions for each item followed by calculating minimum supports is engaged. Other work carried benefits of the OLAP and data mining technology which conducted efficiency and adaptability [9].

Up-to-date, there exist only a few algorithms for quantitative multilevel fuzzy association rule mining (*QMLFRL*). For examples, in [4] the authors advised an *QMLFRL* based on the idea that the minimum support for an item at a higher taxonomic concept is valued as the minimum of the minimum supports of the items pertaining to it and an item minimum support for an itemset is established as the maximum of the minimum supports of the items enclosed in the itemset. Under this limitation, the characteristic of downward-closure is conserved, such that the original Apriori algorithm can be simply prolonged to find fuzzy large item sets.

With the same purpose, the authors in [26] suggested a new method of quantitative association rule extraction that can quantize the attribute by applying a clustering algorithm and learn rules simultaneously. They implemented clustering using all attributes at the same time in advance and deduced the rules from the clusters in the aspect of "association". Based on the numerical experiments, the authors have confirmed that their algorithm outperformed the conventional algorithm based on Cartesian product type quantization in terms of total precision of quantization and rule extraction.

Extra relevant work introduced in [6] to create rules based on the quantitative dataset, utilizing the notion of threshold - frequent item sets that are produced using the genetic algorithm. In this illustrations, crossover & mutation are involved to create numerous unification of the rule and can recognize co-occurrence of item sets. Here three objectives are studied: comprehensibility, interestingness, confidence, so produced rules are established as multi-objective association rules. These objectives serve to decrease search space for fitness function. Finally, optimal rules are formed that is based on distribution approach for the numeric-valued attribute (Right-hand side of a rule reveals the distribution of the values of numeric attributes such as the mean or variance).

The benefit of the preceding systems is that they carry linguistic expressions which make created rules to be much normal for human experts; but they may generate a large number of interesting association rules. Still, for many purposes, it is not periodically simple to ascertain effective association rules (matching the minimum support and confidence) among data items at low (primitive) levels of abstraction due to the sparsity of data in multidimensional space. Other associated problems cover: (1) the shortage of sufficient support for dynamically needed hierarchies; (2) algorithm efficiency cannot meet real application specifications; (3) the association between different concepts levels may be dropped; (4) Their approach enabled users to stipulate various minimum supports to different items. [4][27].

A. Research Contribution

The idea developed in this paper is partly inspired by the existing work on *QMLFRL*, buts it utilizes the genetic algorithm to compute the minimum support and minimum confidence for each level in the Taxonomy regardless of the nature of the data; thus making automatic system. Prior studies have completely investigated single-level association rules mining with GA, such as mining single objective rules and mining multi-objective rules. However, in the big data analysis setting, powerful association rules are regularly in multilevel forms and mining multilevel association rules in big data demands more efficient methods. The GA-based multilevel association rules mining method recommended in this paper is one effort to efficiently discover multilevel association rules in big data.

III. The Proposed Model

The advanced mining algorithm combines fuzzy set notions, data mining, and multiple-level taxonomy to determine fuzzy association rules in a given transaction data set deposited as quantitative values. The knowledge obtained is described by fuzzy linguistic terms, and thus simply readable by human beings. This system utilizes a top-down progressively deepening strategy to locate large itemsets [4]. In this paper, we made our primary intention toward automatically detecting minimum support and minimum confidence of each taxonomy' level by constructing a genetic algorithm based heuristic method for practical multilevel association rules mining in big datasets. By using the advantage of the genetic algorithm, which can efficiently ascertain multiple solutions concurrently in a large multidimensional problem without conducting exhaustive searches, our offered method can enhance the mining performance while preserving the wanted accuracy but bypassing the exhausting list of association rule candidates. Definitions linked to multilevel association rules are presented as follows [9][15][16]:

Definition 1: an item set, X , is a set of data items $\{X_{i,} X_j\}$, where $X_{i,} X_j \in I$. The support of an item set X in a set *S*, $\sigma(X / S)$ is the number of transactions (in *S*) which covers X against the total number of transactions in *S*. The confidence of $X \Rightarrow Y$ in *S*, $\varphi(X \Rightarrow Y/S)$, is the fraction of $\sigma(X \cup Y / S)$ in competition with $\sigma(X / S)$, i.e., the possibility that item set *Y* takes place in *S* when item set occurs X in *S*.

Definition 2: An item set X is large in set *S* at level L if the support of X is no less than its matching minimum support threshold σ'_L . The confidence of a rule $X \Rightarrow Y / S$ is high at level L if its confidence is no less than its equivalent minimum confidence threshold φ'_L .

Definition 3: a rule $X \Rightarrow Y / S$ is strong if $X \cup Y / S$ is large at the existing level and the confidence of $X \Rightarrow Y / S$ is high at the current level.

Definition 4: A fuzzy transaction denoted by *T* is given by:

$$\overline{\overline{T}} = \{(x, \mu(x)) \mid \forall x \in I\},\ 0 \le \mu(x) \le 1, \mu : I \to [0,1], \overline{\overline{T}} \subseteq T$$

where *T* is a general set of transactions, and $\mu(x)$ is degree of membership of *x*.

Definition 5: A soft quantitative transaction set that is symbolized by T'_q . Let (*F*, *E*) is a soft set over the universe *U* and $X \subseteq E$, *F* means the fuzzy power set of *U*, and *E* is a set of parameters. A set of attributes *X* is said to be supported by a transaction if:

$$T'_q = \left\{(< x, l, u >, e) \parallel \forall < x, l, u > \in I \times p \times p \mid l \le u, e \in E\right\}$$

Mining of association rules essentially focalizes at a single conceptual level. There are applications which lack to locate associations at multiple abstract planes. In a large database of transactions, where each transaction consists of a set of items and a taxonomy (is-a hierarchy) on items, it is expected to find out associations between items at any level of taxonomy. To investigate multilevel association rule mining, anyone wants to afford data at multiple-level association at multiple levels of abstraction and efficient methods for multiple level rule

mining. The first specification can be accomplished by producing concept taxonomies from the primitive level concepts to higher levels. The second condition dictates efficient methods for multilevel rule mining [15].

One modification of Apriori to multi-level datasets is the ML_T2L1 procedure [15][24]. The ML_T2L1 algorithm manages a transaction table that has the hierarchy information encoded into it. Each level in the dataset is treated separately. Firstly, level 1 (the highest level in the hierarchy) is examined for large 1-itemsets using Apriori. The list of level 1 large 1-item sets is then employed to refine and clip the transaction dataset of any item that does not have an ancestor in the level 1 large 1-itemset list and eliminate any transaction which has no common items (thus comprises only infrequent items when evaluated using the level 1 large 1-itemset list). From the level 1 large 1-itemset list, level 1 large 2-itemsets are concluded (using the cleaned dataset). Then level 1 large 3- item sets are inferred and so on until there are no more frequent item sets to find at level 1. Since ML_T2L1 specifies that only the items that are descendant from frequent items at level 1 (essentially they must descend from level 1 large 1-itemsets) can be frequent themselves, the level 2 item sets are concluded from the refined transaction table. For level 2, the large 1-itemsets are created, from which the large 2-itemsets are determined and then large 3-itemsets etc. After all the frequent itemsets are found at level 2, the level 3 large 1-itemsets are located (from the same purified dataset) and so on. ML_T2L1 reforms until either all levels are explored using Apriori or no large 1-itemsets are exposed at a level. The principal steps of the proposed system are as follows [8-10][14][24][29-31]:

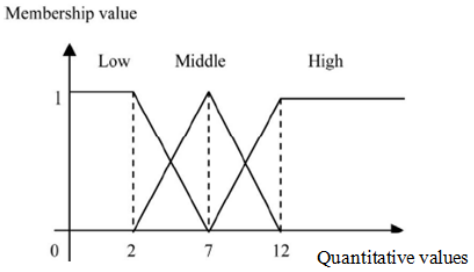


Fig. 2. The membership functions of items in Γ

Input: A group of *N* quantitative transaction data *D*, a pre-determined catalog Γ with the original items $\{i_1, i_2, ..., i_n\}$, a set of membership functions for each item in deferent levels. In our case, all the membership functions have the same style as shown in Fig. 2; but the x-axis is determined for each element in Γ based on the higher quantitative value associated with it. Finally, the parameter set minimum support α_k and minimum confidence λ_k that are acquired by genetic algorithm.

Output: A collection of fuzzy multiple-level association rules below the restrictions of optimal minimum support and confidence.

Step 1: Translate the predefined taxonomy using an arrangement of numbers and the symbol "*" by the formula, $C = \rho * 10 + i$, where i is the position number of the node at current level l, C signifies the code for the i^{th} node at current level and ρ is the code of parent of the i^{th} node at the present level.

Step 2: Interpret the item terms in the transaction data agreeing to the encoding scheme. Then set $k = 1$, $r = 1$ where k, $1 \le k \le x$ is the recent level number, x is the number of level in a given taxonomy and r denotes the number of items kept in the current frequent item sets.

Step 3: Cluster the items with the same first k digits in each transaction D_i, and add the quantities of the items in the similar sets in D_i. Symbolize the total of the j-th group I_j^k for D_i as v_{ij}^k.

Step 4: We explored several membership function for various data items for that each data item has its own features and its own membership function, then transform the value v_{ij}^k of each transaction D_i for each encoded group I_j^k into a fuzzy set f_{ijl}^k (Eq.1) by plotting v_{ij}^k on the specified membership function, where I_j^k is the j-th item on level k, v_{ij}^k is the quantitative value of I_j^k in D_i, h_j^k is the number of fuzzy areas for I_j^k, R_{jl}^k $(1 \le l \le h_j^k)$ is the l-th fuzzy region of I_j^k, f_{ijl}^k is v_{ij}^k fuzzy membership value in R_{jl}^k

$$\left(\frac{f_{ij1}^k}{R_{j1}^k} + \frac{f_{ij2}^k}{R_{j2}^k} + ... + \frac{f_{ijh}^k}{R_{jh}^k} \right). \quad (1)$$

Step 5: Assemble the fuzzy regions (linguistic terms) with membership values larger than zero to create the candidate set C_1^k; Calculate the scalar cardinality S_{jl}^k of each fuzzy region R_{jl}^k in the transaction data as $S_{jl}^k = \sum_{i=1}^{n} f_{ijl}^k$.

Step 6: Investigate if the value S_{jl}^k of each region R_{jl}^k in C_1^k is larger than or equals to the threshold α_k which is the optimal minimum support for level k obtained from implementing the genetic algorithm to the set of transactions included in this level according to Γ (see algorithm 1). If R_{jl}^k matches the threshold, place it into the large 1-itemset L_1^k for level k. That is:

$$L_1^k = \left\{ R_{jl}^k \middle| S_{jl}^k \ge \alpha_k, R_{jl}^k \in C_1^k \right\}. \quad (2)$$

Step 7: if L_1^k is null, let $k = k + 1$ and go to step 3; else, create the applicant set C_2^K from $L_1^1, L_1^2, ..., L_1^K$ to catch "level-crossing" large itemsets. The created applicant set C_2^K has to fulfill the following conditions: (1) Each 2-itemset in C_2^K must comprise at least one item in L_1^k. (2)The two regions in a 2-itemset may not have the same item name. (3) The two item names in a 2-itemset may not be with the hierarchy relation in the taxonomy. (4) Both of the support values of the two large

1–itemsets including a candidate 2-itemset must be larger than or equal to the minimum support $\alpha_{k=2}$.

Step 8: If L_1^k is null, then increase *k* by one, *r* =1 and go to step3 else set *r*=*r*+1.

(a) If *r* = 2 create the candidate set C_2^K, where C_2^K is the set of candidate itemset with 2 items on level *k* from $L_1^1, L_1^2, L_1^3, ..., L_1^k$ to learn "level-crossing" of frequent itemset. Each 2-itemset in C_2^k must contain at least one item in the L_1^k and the next item should not be its ancestor in the taxonomy. All possible 2-itemsets are composed in C_2^k.

(b) If $r > 2$, produce the candidate set C_r^k, where C_r^k is the set of candidate itemset with *r*-items on level *k* from L_{r-1}^k in a way similar to that in the preceding steps.

Step 9: For each acquired candidate *r*-itemset *S* with items $(S_1, S_2, ...S_r)$ in C_r^k:

a) Calculate the fuzzy value of *S* in each transaction datum D_i by the minimum operator as $f_{is} = \min(f_{is1}, f_{is2}, ..., f_{isr})$

b) Estimate the scalar cardinality of *S* in all the transaction data as $count_s = \sum_{i=1}^{n} f_{is}$.

c) If $count_s$ is larger than or equals to the pre-defined minimum support α_k place *S* into L_r^k.

Step 10: If L_r^k equal null then increase *K* by one and go to the next step; if not increase *r* by one and go to step 8.

Step 11: If $k > x$ then go to the next step, else set *r* =1 and go to step 3.

Step 12: create the fuzzy association rules for all frequent *r*- itemset including $S = (S_1, S_2, ..., S_r)$, *r >2* as follows:

- Catch all the rules $A \to B$ where $A \subset S$, $B \subset S$ and $A \cap B = \phi$, $A \cup B = S$.
- Calculate the confidence value of all association rules by $\frac{\sum_{i=1}^{n} min(f_{iS})}{\sum_{i=1}^{n} min(f_{iA})}$.

Step 13: Choice the rules that have confidence values not less than predefined confidence threshold λ_k, where λ_k is the predefined minimum confidence value for level *k* found by applying genetic algorithm.

Step 14: eliminate redundant rules from multi-level datasets. Herein, Rule R_1 is redundant to rule R_2 if (1) the itemset X_1 is made up of items where at least one item in X_1 is descendant from the items in X_2 and (2) the item set X_2 is entirely made up of items where at least one item in X_2 is an ancestor of the items in X_1 and (3) the other non-ancestor items in X_2 are all present in item set X_1. The additional state (4) the confidence of $R_1(C_1)$ is less than or equal to the confidence of $R_2(C_2)$.

A. Parameters extraction using genetic algorithm

A genetic algorithm is a class of investigating algorithm that is employed to automatically set the optimal minimum support and minimum confidence for each taxonomy's level. It explores a solution space for an optimal answer to a problem [28]. The algorithm generates a "population" of feasible solutions to the problem and makes them "evolve" over many generations to locate valid and better solution. The algorithm begins with a collection of solutions (represented by chromosomes) called a population. Solutions from one population are selected and managed to establish a new population. The framework of the basic genetic algorithm is as follows (see Fig. 3).

```
Procedure genetic algorithm
begin (1)
    t = 0;
    inicialize P(t);
    evaluate P(t);
    While (Not termination-condition) do
    begin (2)
        t = t + 1;
        select P(t) from P(t − 1);
        recombine P(t);
        evaluate P(t);
    end (2)
end (1)
```

Fig. 3. Structure of the genetic algorithm [29]

1) [Start] create arbitrary samples of n chromosomes (appropriate results for the problem)

2) [Fitness] assess the fitness (qualification) function f(x) of every chromosome x in the population

3) [New population] generate a new resident by iterating the subsequent steps until the new population is complete.

*Selection***:** pick two parent chromosomes from a population according to their fitness (the better fitness, the higher possibility to be chosen)

*Crossover***:** with a crossover probability, crossover the parents to produce a new generation (children). If no crossover was conducted, offspring is an accurate reflection of parents.

*Mutation***:** with a mutation probability, the GA mutates a new generation at each location (site on the chromosome).

*Accepting***:** store distinct generation in a new population.

4) [Replace] manage recently produced population for a more route of the algorithm

5) [Test] if the end condition is satisfied, stops, and returns the best solution in current population

6) [Loop] go to step 2.

Through repetition *t*, the GA keeps a population $p(t)$ of results $r_1^t, ..., r_N^t$, where r_i^t characterizes rule set that is

arbitrarily created for each level. Each solution r_i^t is gauged by the function $E(\bullet)$ and $E(r_i^t)$ is a degree of fitness of the solution. The fitness value determines the relevant power of an individual to remain and create offspring in the next production. In the next iteration (t+1) a new resident is designed on the foundation of the operations (2) and (3) [29].

B. Data Encoding

Given a randomly generated association rules for each level, the system uses the Michigan approach for encoding, in which a chromosome is a collection of all used rules; here the population consists of many rule collections. Coding in the Michigan method is binary coding, in which "1" means that a knowledge base rule will be in a knowledge base, whereas "0" means it will not be used. The key benefit of this technique is that the entire rule base is coded; therefore, it is not necessary to do the quantitative analysis of indispensable rules to see if the method functions properly, because, unlike the Pitts method, all possible rules take part in the working time of the genetic algorithm. The considerable size of the chromosome is a disadvantage. The dimension of the chromosome is dependent on the volume of the rule base and it increases exponentially depending on the number of itemsets [10][14][30].

C. Generic Operators

The frequently employed genetic operators are reproduction, crossover, and mutation. To achieve genetic operators, one must pick individuals in the population to be worked on. The collection plan is mainly based on the fitness level of the individuals exhibited in the population. For election; the system manages roulette wheel sampling fashion. In this procedure, the parents for crossover and mutations are chosen based on their fitness, i.e. if a candidate has more fitness function value more will be its opportunity to get elected. The implementation of roulette wheel sampling is performed by first normalizing the values of all applicants so that, their chances sprawl between 0 and 1, and then by applying random number function, a random number is estimated, and then matching to this value and the fitness normalized value, the candidate is elected [14].

As an individual is picked, reproduction operators only imitate it from the current population into the new population (i.e., the new generation) without transposition. The crossover operator begins with two selected individuals and then the crossover point (an integer between 1 and L-1, where L is the length of strings) is picked arbitrarily. The third genetic operator, mutation, offers random variations in the arrangements in the population, and it may irregularly have useful results: departing from a local optimum. In our GA, mutation is just to oppose every bit of the strings, i.e., changes a 1 to 0 and vice versa, with probability p_m [30].

The algorithm stops fulfilling when the decay situation is reached − i.e. when the best and worst producing chromosome in the population disagrees by less than 0.1%. It also ends execution when the total number of generations defined by the user has arrived. Besides, the algorithm bypasses forming the initial population completely randomly because it may appear in rules that will include no training data instance whereby having very low fitness. Furthermore, a population with rules that are insured to comprise at least one training instance can lead to over−fitting the data. It was shown that employing non−random initialization can reach to an elevation in the quality of the solution and can drastically decrease the runtime [24]. We, therefore, devised an initialization method which involves picking a training instance to serve as a "seed" for rule generation based on the alteration of itemsets within each level [31].

In general, genetic operator assists in controlling the heterogeneity of the population and also in blocking early concurrence to local optima [14]. Our intention is to explore fascinating association rules. Consequently, the fitness function is vital for ascertaining the interestingness of chromosome, and it does influence the convergence of the genetic algorithm. In this case, the proposed system examines two different fitness functions. The first one considers the identical confidence of the corresponding association rule as illustrated in Eq. 3, whereas the second fitness function joins the support (*sup*) and *confidence* (*conf*) attributes, which are required to define an association rule (see Eq. 4) [9][10][14]. Parameters α and β are the significant factors to equilibrium the weight of the *support* and *confidence* in the fitness function, $\alpha+\beta=1$. To mine confirmed association rules from the big database with our GA approach, the threshold of the fitness function has to be predefined; in our case, $\alpha=\beta=0.5$.

$$f_1: rconf(X \to Y) = \frac{\sup(X \cup Y) - \sup(X) \times \sup(Y)}{\sup(X)(1-\sup(Y))} \tag{3}$$

$$f_2(x \to y) = \alpha \times \sup(x \to y) + \beta \times conf(x \to y) \tag{4}$$

By adopting the recommended system, rather of producing an untold number of interesting rules in conventional mining models, only the most interesting rules are declared according to the interestingness measure determined by the fitness function. The main motivation for using GA in the learning of high-level prediction rules is that they conduct a global search and cope better with attribute cooperation than the greedy rule selection algorithms [14].

In brief, the proposed evolutionary method for quantitative association rule mining is particularly prompted by (1) partition of quantitative attribute is not accessible for every attribute and every user, (2) users, and even experts, usually feel tedious to define the minimum-support, (3) the search space might be very large when we face quantitative attributes, and (4) the rules passed might be too many to deal with [10]. However, mining association rules are not adequate of benefits; it has some defects too, first of all, the algorithmic complexity. The number of rules increases exponentially with the number of items. But this complexity is undertaken with some advanced algorithms which can efficiently clip the search space. Secondly, the obstacle of attaining rules from rules, i.e. selecting interesting rules from the set of rules.

The suggested work undertakes the second problem that essentially assist the user when scanning the rule set, and valuable quality measures on the rules are adopted based on genetic algorithm. Usually, when handling association rule mining many rules can be found or inferred and confuse the

user. But more importantly, some of these rules could be redundant and yield no new knowledge. Some attempts have been pointed at dealing with redundant rules in flat datasets, however, datasets can have a hierarchy/taxonomy or compound concept levels and thus redundancy in these datasets require to be concentrated on. This issue is one of the features of this study.

IV. Experiments and Results

In this section, we perform some experiments that have been conducted out to examine the performance of the proposed approach and confirm enhancements over the traditional method without optimization. The experiment was conducted with MATLAB software. All the experiments are handled on a laptop computer with the following specifications: Processor Intel(R) Core (TM) i5-2520MCPU@2.50GHz 2.50GHz, memory 4.00GB, and System type 64-bit operating system, x64-based processor.

A. Dataset

The dataset was hired as in [8] and can be viewed as a benchmark because it is used for comparison. This is a market basket dataset that consists of the items and amounts of items marketed in every purchasing container. This dataset consists of 1000 sales receipts of a food material repository based on the predefined taxonomy from 7 items (10000 transactions). The predefined taxonomy in the first level holds 7 nodes that describe the items worked in the test, the second level comprises 14 nodes that describe the taste or different types of a particular stock and in the third level, it also consists of 48 nodes that express the manufacturing companies and factories. The database transactions carry the name of the product and the quantity of such purchased merchandise. One item may not be employed twice in one transaction.

B. Methodology

A comparison was made between the conventional approach for mining multi-level fuzzy quantitative association rules [8] and the approach proposed in this paper that uses GA optimization technique. The objective was to find new detailed knowledge by (1) an enhancement in pronouncing multi-level optimal support and confidence that is employed to obtain interesting rules. (2) eliminating redundant rules that were encountered in the traditional approach. Both algorithms utilize a top-down progressively deepening approach to infer large itemsets and also consolidate fuzzy boundaries instead of explicit boundary intervals.

The mined rules from the proposed system are closer to the reality, and it gives the ability to mine association rules at different levels based on the optimal re-calculated mining parameters (*min_sup*, *min_conf*); unlike the traditional method that depends on the experts to determine these parameters manually. Employing GA to find these parameters makes the proposed system is context-independent and more general. In the experiments, thresholds for *min-sup* and *min-conf* were set at 0.28 and 1.7 respectively for the traditional algorithm for each taxonomy level.

In the first experiment, we examine if correct association rules can be specified with a fixed number of initial generations and in a bounded time period. Using our dataset, the initial population size ranges from 30 to 100. The results are displayed in Table 1. With a limited population, most strong association rules could be inferred in our dataset. We can decide that if the population is too small, the realization of the GA-based algorithm will be similar to the random algorithm. But if the population is too large, despite we can get full association rules immediately, but the computational complexity rises fast. However, as we can see, there is a good stability that, with a restricted population and a limited time period, most valid association rules have been mined. Therefore, we choose 50 as the default population for the dataset, which works well in our algorithm.

TABLE I. Relation Between the Number of Initial GA Population and the Number of Multilevel Association Rules (Using F1 with Michigan Encoding, Generation No. =10)

No. of Initial population	**30**	**50**	**70**	**100**
No. of association rule (Redundant)	3706	607	607	607
No. of association rule (non-redundant)	1733	349	349	349

One reason of the stability of the number of extracted rules with initial population contains 50 chromosomes is that the proposed system utilizes Michigan approach for rule coding in which a chromosome is a collection of all used rules. So, the lowest number of initial populations will contain specific rules permutations. In Michigan approach, each chromosome contains a comprehensive representation of the rules.

In the next collection of experiments, we confirmed how worthy the extracted association rules are from either the GA-based algorithm or the traditional algorithm without GA. To measure its value with the 10000 transactions, we use the formula of the fitness function f_1 using the following configuration mutation rate = 0.1, crossover rate = 0.9, generation No.=10,and initial population=50. The results in dataset are shown in Table 2.

TABLE II. Comparative Study

Methods	**No. of non-redundant rules**	**Calculated *min-sup***	**Calculated *min-conf***	**Time (Sec)**
Proposed Method with GA	349	L1=0.95 L2=0.67 L3=0.27	L1=1.12 L2=1.45 L3=1.92	500
Traditional method without GA [8]	2281	0.28 (L=1 to 3)	1.7 (L=1 to 3)	400

TABLE III. Comparative Study between the Two Fitness Functions

Fitness Function	**levels**	**Computed Fitness**	**No. of association rules (non-redundant)**
f_1 (Eq. 3)	level 1	1.90	349
	level 2	1.97	
	level 3	2.06	
f_2 (Eq.4) $\alpha = \beta = 0.5$	level 1	1.07	2249
	level 2	1.26	
	level 3	1.47	

TABLE IV. PROPOSED SYSTEM EVALUATION UNDER DIFFERENT PARAMETERS OF GA USING F_1

Parameters ratio	No. of association rules (redundant)	No. of association rules (non-redundant)
Mutation= 0.9 crossover= 0.1	751	419
Mutation=0.8 crossover=0.2	751	419
Mutation=0.7 crossover=0.3	607	349

Compatible with the outcomes above, with the advance of the time boundary, GA-based approach can catch high relevant association rules with a little more time than traditional method (25% increases in time). But in terms of quality, the proposed system extracts more interesting rules; only about 17 % of the total number of rules extracted from the comparative system. In general, a large number of extracted rules inside market basket analysis will hamper the decision-maker. The proposed system offers the most interesting rules subject to the fitness function, which is accountable for acting the assessment that imitates how optimal the solution is: the higher the number, the better the solution.

The third set of experiments was executed to compare how relevant the mined association rules are from either the fitness function that considers the relative confidence of the corresponding association rule (Eq. 3) or the fitness function that considers both the support and confidence attributes (Eq. 4). The experiment is conducted under the previous configuration for GA. The results in Table 3 reveal that the use of f_1 generates a further mined association rules rate improvement of 83% reduction in the number of extracted rules. From this experiment, we realized that the fitness function represents a critical issue in the success of genetic algorithm; this is clearly shown in the case of f_2 . Using f_2 did not bring any advantages to the GA; thus, we get the same number of extracted rules that have been obtained from the traditional method (about 2249).

The performance improvement of GA using f_1 comes from the correct extraction of interesting rules; because of calculating the support of union for items inside each rule in addition to the support of each item separately. Unlike the second function that uses the support and confidence of each rule, which represents the standard case used by many of the existing mining algorithms (e.g. Apriori).

Having compared our system to different multi-level quantitative mining algorithm, we will next explore the influence that GA factor settings have on our system, which incorporates both mutation rate and crossover rate. To retain the number of factor setting blending small, we will only fluctuate the setting for one parameter at a time while holding the setting for another parameter to its default value. In Table 4, we vary mutation rate from 0.7 to 0.9 and look at the number of association rules (non-redundant) achieved by our system on the used dataset. From the table, we can see that decreasing mutation rate will diminish the number of extracted rule (17% lower in the number of rules). This decreasing is noticeable. This decrease is due to the fact that mutation is managed to preserve genetic heterogeneity from one generation of a population to the next. In GA, mutation operators are frequently employed to give exploration and crossover operators are extensively employed to supervise the population to focalize on one the good solutions encounter so far (exploitation). Consequently, while crossover attempts to concentrate to a special point in the landscape, the mutation does its best to evade convergence and investigate more areas.

V. CONCLUSION

Really, mining quantitative association rules is an optimization obstacle rather than being an uncomplicated discretization one. In this paper, we have introduced a new genetic-based algorithm to mine multilevel association rules in big quantitative date sets that deals with quantitative attributes by accurate fuzzification the values -partitioning the values of the attribute. The proposed system uses fuzzy set concepts, multi-Level taxonomy, different pre-calculated minimum supports for each level and different membership function for each item to discover fuzzy association rules in a given transaction data set.

In our algorithm, the minimum supports and minimum confidences for each level for fuzzy quantitative association rule are defined by the genetic algorithm optimization. In this case, the employed GA combines a population initialization technique that guarantees the production of high- quality individuals; individually planned breeding operators that confirm the removal of inadequate genotypes; an adaptive mutation probability to ensure genetic heterogeneity of the population; and uniqueness testing based on both support and confidence that is employed to hold only high quality and interesting rules.

The proposed system gives the user with rules according to two interestingness metrics, which can quickly be extended if need by changing the fitness function. The results report that: In terms of mining of association rules, the proposed method keep higher precision compared with the traditional methods and the extracted rules are more close to reality. This is because of adopting various membership functions for every individual item, optimized minimum supports, and minimum confidences, and finally, the non-redundant algorithm to enhance the quality and application of the rules. Future work includes employing GA to tune the fuzzy membership function for each item.

REFERENCES

[1] J. Han, M. Kamber, and J. Pei, “Data mining concepts and techniques,” Elsevier Inc., USA, 2012.

[2] T. Hong, Y. Tung, S. Wang, Y. L. Wu, and M. T. Wu, "A multi-level ant colony mining algorithm for membership functions, "Information Sciences, vol. 182, no. 1, 2012, pp. 3-14.

[3] A. Gosain, and M. Bhugra, "A comprehensive survey of association rules on quantitative data in data mining, " the IEEE Conference on Information and Communication Technologies, India, 2013, pp.1003-1008.

[4] Y. C. Lee, T. P. Hong, and T. C. Wang, "Mining fuzzy multiple-level association rules with multiple minimum supports, " Expert Systems with Applications, vol. 34, no. 1, 2008, pp.459-468.

[5] K. Poornamala, and R. Lawrance "A general survey on frequent pattern mining using genetic algorithm, " ICTACT Journal on Soft Computing, vol. 3, issue 1, 2012, pp.440-444.

[6] D. Kanani, and S. Mishra, "An optimize association rule mining using genetic algorithm," International Journal of Computer Applications, vol. 119, issue 14, 2015, pp.11-15.

[7] E. Mahmoudi, E. Sabetnia, M. Torshiz, M. Jalal, and G. Tabrizi, "multi-level fuzzy association rules mining via determining minimum supports and membership functions, " the Second International Conference on Intelligent Systems Modeling, and Simulation, Iran, 2013pp.55-61.

[8] A. Kousari, S. Mirabedini, and E. Ghasemkhani, "Improvement of mining fuzzy multiple-level association rules from quantitative data," Journal of Software Engineering and Applications, vol. 5, no. 3, 2009,pp. 190-199.

[9] X. Yang, M. Zeng, Q. Liu, and X. Wang, "A genetic algorithm based multilevel association rules mining for big datasets" Mathematical Problems in Engineering, Vol. 2014, 2014, pp.1-10.

[10] X. Yan, C. Zhang, and S. Zhang, "genetic algorithm-based strategy for identifying association rules without specifying actual minimum support, " Expert Systems with Applications vol. 36, no. 2, 2009, pp. 3066-3076.

[11] S. Aouissi, A. Vrain, and C. Nortet, "Quantminer: a genetic algorithm for mining quantitative association rules", the 20th International Conference on Artificial Intelligence , 2007, pp. 1035–1040.

[12] G. Shaw, Y. Xu, and S. Geva, "Utilizing non-redundant association rules from multi-level datasets" the IEEE International Conference on Web Intelligence and Intelligent Agent Technology, vol. 03, Australia, 2008, pp.681-684.

[13] P. Gautam, and K. R. Pardasani, "Algorithm for efficient multilevel association rule mining", International Journal on Computer Science and Engineering, vol. 2, no.5, 2010, pp.1700-1704

[14] S. Manish, A. Agrawal, and A. Lad, "Optimization of association rule mining using improved genetic algorithms," the International Conference on Systems, Man and Cybernetics, vol. 4, USA, 2004, pp.3725-3729.

[15] V. Ramana, M. Rathnamma, and A. Reddy, "Methods for mining cross level association rule in taxonomy data structures" International Journal of Computer Applications, vol. 7, no.3, 2010, pp.28-35.

[16] S. Saraf, N. Adlakha, and S. Sharma, " Soft set approach for mining quantitative fuzzy association patterns in databases, " International Journal of Advanced Research in Computer Science and Software Engineering, vol. 3, issue 11, 2013, pp.359-369.

[17] M. Delgado, N. Manín, M. J. Martín-Bautista, D. Sánchez, and M. -A. Vila" Mining fuzzy association rules: an overview, "Studies in Fuzziness and Soft Computing, Springer Berlin Heidelberg , vol. 164, 2005, pp.351-373.

[18] G. Attila, "A fuzzy approach for mining quantitative association rules, " ACTA CYBERN, vol. 15, no.2, 2001, pp.305-320.

[19] W. Toshihiko, and H. Takahashi, "A study on quantitative association rules mining algorithm based on clustering algorithm, " International Journal of Biomedical Soft Computing and Human Sciences, vol. 16, no.2, 2010, pp.59-67.

[20] R. Sridevi, and E. Ramaraj, "A general survey on multidimensional and quantitative association rule mining algorithms, " International Journal of Engineering Research and Applications, vol. 3, issue 4, 2013, pp.1442-1448.

[21] P. Gautam, N. Khare, and K. R. Pardasani," A model for mining multilevel fuzzy association rule in database, " Journal of Computing, vol. 2, issue 1, 2010, pp.58-64.

[22] O. Oladipupo, C. Ayo, and C.Uwadia, " a fuzzy association rule mining expert-driven Approach to Knowledge Acquisition," African Journal of Computing and ICT, vol. 5, no. 5, 2012,pp.53-60.

[23] S. Prakash, M. Vijayakumar, and R.Parvathi, " A Novel Method of Mining Association Rule with Multilevel Concept Hierarchy", International Journal of Computer Applications, vol. 5, no. 5, 2011, pp.26-29.

[24] G. Shaw, Y. Xu, and S. Geva, "Eliminating redundant association rules in multi-level datasets, " the 4th International Conference on Data Mining, USA, 2008pp.1-8.

[25] E. Ayetiran, and A. Adeyemo, "A data mining-based response model for target selection in direct marketing," International Journal Information Technology and Computer Science, vol. 4, no. 1, 2012, pp. 9-18.

[26] W. Toshihiko, and H. Takahashi, "A study on quantitative association rules mining algorithm based on clustering algorithm," Journal of the Biomedical Fuzzy Systems Association, vol.16, no.2, 2010, pp.59-67.

[27] Y. Wan, Y. Liang, and L. Ding, "mining multilevel association rules from primitive frequent itemsets," Journal of Macau University of Science and Technology, vol.3, issue 1, 2009,pp.10-19.

[28] R. Haldulakar, J. Agrawal, "Optimization of Association Rule Mining through Genetic Algorithm," International Journal on Computer Science and Engineering, vol. 3, no. 3, 2011, pp.1252-1259.

[29] M. Kayaa, R. Alhajj, "Genetic algorithm based framework for mining fuzzy association rules, " Fuzzy Sets and Systems, vol. 152, 2005, pp. 587–601.

[30] S. Tiwari, M. K. Rao" Optimization in association rule mining using distance weight vector and genetic algorithm , "International Journal of Advanced Technology & Engineering Research, vol. 4, Issue 1, 2014, pp.79-84.

[31] P. Wakabi-Waiswa, and V. Baryamureeba, "Mining high quality association rules using genetic algorithms", In Proceedings of the twenty second Midwest Artificial Intelligence and Cognitive Science Conference, USA, 2009, pp. 73-78.

A Model for Facial Emotion Inference Based on Planar Dynamic Emotional Surfaces

Ruivo, J. P. P.
Escola Politécnica
Universidade de S˜ao Paulo
S˜ao Paulo, Brazil

Negreiros, T.
Escola Politécnica
Universidade de S˜ao Paulo
S˜ao Paulo, Brazil

Barretto, M. R. P.
Escola Politécnica
Universidade de S˜ao Paulo
S˜ao Paulo, Brazil

Tinen, B.
Escola Politécnica
Universidade de S˜ao Paulo
S˜ao Paulo, Brazil

***Abstract*—Emotions have direct influence on the human life and are of great importance in relationships and in the way interactions between individuals develop. Because of this, they are also important for the development of human-machine interfaces that aim to maintain a natural and friendly interaction with its users. In the development of social robots, which this work aims for, a suitable interpretation of the emotional state of the person interacting with the social robot is indispensable. The focus of this paper is the development of a mathematical model for recognizing emotional facial expressions in a sequence of frames. Firstly, a face tracker algorithm is used to find and keep track of faces in images; then the found faces are fed into the model developed in this work, which consists of an instantaneous emotional expression classifier, a Kalman filter and a dynamic classifier that gives the final output of the model.**

***Keywords*—emotion recognition, facial emotion, Kalman filter, machine learning**

I. Introduction

Emotions influence the human behavior and the way individuals interact and relate to other members of society. They permeate one's daily life and determine how people react to the various situations they encounter in their routines.

Studies indicate that people with impairments to express or recognize feelings end up having great difficulty keeping even casual relationships [1]. Emotions also help the body prepare for specific external events. For example, the fear people may experience when they see a large object coming fastly towards them stimulates blood circulation in their legs, allowing them to act promptly and respond trying to avoid the object.

Computer interfaces that can understand the emotional state of its users can communicate more naturally compared to interfaces without this capability. Affective computing comes to deal with the integration of the concept of emotion in the computational area [2].

Emotions are characterized by signs in voice, speech and body movements, which are recognized regardless of culture, possibly being a legacy of human evolution and not a result of personal experiences of the individual [3]. Particularly in the face, the most obvious signs are presented in the regions of the mouth, eyes and eyebrows. Ekman and Friesen showed evidence for the hypothesis of universality of emotional facial expressions in intercultural studies with illiterate populations of Papua New Guinea and investigated the influence of the cultural phenomena [3].

Works from Ekman [4] [5] propose the existence of six major universal emotions: joy, sadness, surprise, fear, anger and disgust. An emotional display can either be classified as belonging to one category, such as joy, or more than one category, forming composite emotions, such as the mixture of fear and angry, or joy and surprise.

This study aims to identify five basic emotional states: Happiness, Sadness, Anger and Fear, plus the Neutral state, which could be understood as the absence of emotions. The model proposed in this work does not try to describe short-lived or rapidly changing emotions (micro expressions, in the works of Ekman), but focuses on trying to detect lasting emotional states people may be subject to. The dynamic model for emotion recognition presented in this work is a novel model based on the work of [6].

The rest of this paper is organized as follows. In Section II are reviewed previous works regarding automatic emotion inference. The adopted methodology, including face detection, feature extraction, instantaneous emotion recognition and dynamic emotion recognition is then presented at Section III. Section IV describes the results obtained. Finally, the conclusions and future work directions are presented on Section V.

II. Bibliographic Review

There are three main approaches to emotions classification: discrete model, dimensional model and the approach based on evaluation mechanisms [7].

The discrete model arranges emotions in categories, like the basic emotions of Ekman. Categorization of emotions is an intuitive and practical way to identify them, even if a large number of classes is necessary in order to classify all of the known affective states. Many of the works developed in the area utilizes this approach [8] [7] [9] [10].

The dimensional model seeks to describe the emotions by means of some criteria or dimensions. Two key dimensions are valence and arousal [10] [11]. Valence transmits how the person feels under the influence of a certain emotion, and can assume continuous values ranging from extreme sadness, for negative valence, to extreme happiness, for positive valence. Arousal is associated to the possibility of an individual to take or to perform an action under influence of an emotion, and can assume continuous values ranging from an extremely passive attitude, for negative arousal, to an extremely active

attitude, for positive arousal. Some authors [12] suggest other dimensions for the model, such as dominance. Dominance is related to the control somenone has over a situation while under the influence of an emotion, and can assume continuous values ranging from total lack of control to total control of the situation. The dimensional model avoids the need for an extensive list of categories. Emotions are identified depending on its position on the model's axes. However, because of the limited number of dimensions this approach deals with, the projection of an emotion to the model's axes could cause loss of information [7].

The evaluation approach classifies emotional displays based on a set of assessments of the event that caused such display. For a given emotion, it is evaluated how relevant it is the event that elicited the emotion, what are its implications, the individual's ability to deal with these implications, and what is the significance of that event for the society the individual inhabits [13]. This approach is less simple and intuitive when compared to the others, as it requires a detailed analysis of the situations that elicited the emotions.

Pantic [14] suggests automatic recognition of facial emotion expressions to be done in three main steps: face detection, extraction of relevant features of the face and emotion classification

Face detection is a crucial step in the recognition of expressed emotions, and comprises of locating faces in still images or image sequences. In several works, such images are obtained under conditions that helps face detection algorithms, like the capture of the face in frontal orientation, without occlusions, and under uniform lighting conditions. However, in real situations, these conditions rarely can be reproduced, which makes the problem more challenging. Consequently, an ideal method of facial detection should deal with problems such as the different scales and orientations the human face may take, besides having to consider possible partial occlusions of the face and changes in the lighting conditions.

Extraction of face relevant characteristics has the purpose of generating a feature vector to be used for the emotion identification. It seeks to describe the face through certain categorical or numerical information that should contribute to the recognition of the emotional state of the analyzed person. These characteristics may be based on features of the human face such as eyebrows, nose and mouth, or may be based on mathematical models. These models, in turn, may follow an analytical approach, in which the face is represented by a set of points or patterns of interest that contain specific regions of the face; or they may follow a holistic approach, in which the face is seen as a unit, with its particular shape and texture. Hybrid approaches also exist, in which features of the two above-mentioned approaches are combined. Different scales and orientations of the face, as well as partial occlusion and noise, hamper the execution of this step.

The extracted features vector should then be used to estimate the expressed emotion via a classification algorithm. In this step, any of the approaches presented for emotion classification may be used; however, much of the work done in the area uses the discrete approach [15]. The classification of the facial emotion expressions is done by machine learning algorithms trained with the feature vectors extracted from the samples of one or more training databases. Examples of these algorithms are Support Vector Machines (SVMs), Decision Trees and Neural Networks (NNs).

The present work introduces a fourth step to the process proposed by [14] and includes the usage of a continuous emotional classifier model, following the line of work of [9] and [16]. This step was introduced so that the model would be able to detect long-lasting emotional states rather than instantaneous emotional displays; also, it should help with the minimization of the influences of natural noises, like laugh and speech, that deform the face and difficult the determination of someone's facial emotion expression.

Figure 1 presents a flow diagram of the steps aforementioned.

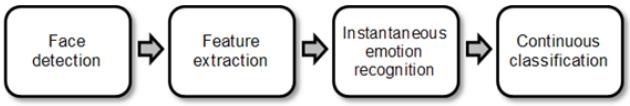


Fig. 1: Flow diagram for the proposed model.

One way to describe one's facial emotion expression is to use the Facial Actions Coding System (FACS) [17]. This system defines 44 Action Units (AU), each one representing the facial movements caused by muscle activity in a specific region of the face. Studies show that a particular subset of 15 of these AUs have greater relevance in the communication between humans [18].

FACS can be understood as an abstraction layer of the underlying facial muscle activity. Through the identification of the level of activity of the relevant AUs, one can infer the related muscles' activities and the corresponding facial expression. FACS defines, for example, involuntary and sincere expression of happiness as the activation of AUs numbers 6 and 12, that is, the lifting of the cheeks and the lateral and vertical extent of the lips, respectively. A forced (faked) expression shows only activation of AU 12 instead. This differentiation is possible because AU 12, which is the contraction of the zygomatic major muscle, is voluntary, while AU 6, contraction of the orbicularis oculi muscle, occurs involuntarily.

Furthermore, FACS brings into consideration the duration and intensity of AUs. Spontaneous muscle activations are in the range 250ms to 5 seconds, depending on the AU [19]. Rules for determining the intensity of each AU are also determined on FACS, for example, as the degree of elevation of the corner of the lips to the AU 12 or the wrinkle density over the nose for AU 44.

As noted in [9] [20], two different categories of properties could be extracted from faces: geometric properties and appearance properties.

Methods based on geometric properties look for characteristic regions of the face, such as eye contour, representing the shape and geometry of the features to be studied. For the extraction of data in video, one approach is the optical flow, as in [21], with tracking of characteristic points. Another approach are three-dimensional methods [22], which were developed along with the development of three-dimensional videos. In the solution presented in [23], the Active Shape

Model [24] and a Kalman filter were used to locate specific areas such as mouth and eyes in each frame of a video.

Appearance-based methods, however, search for changes in texture, such as wrinkles on the face. These methods can be used to describe the whole face or specific regions of interest [25] [18].

Following Figure 1, the next step, emotion classification, can be based on neural networks (NN), support vector machines (SVM) or hidden Markov chains (HMM) [9] [22] [22] [26], among other algorithms [18] [23].

It should also be noted that the humans' emotions detection system is not perfect, and emotions are not always interpreted correctly [14]. Donato [21] shows that people who had no training were able to correctly identify emotions in about 80% of a set of photos, but trained people, such as those passing throught FACS training, have a hit rate of about 90%. For Russell [27], however, a number of studies show that the rate of recognition by individuals varies according to the experimental conditions, ranging from about 55% to about 95%; also, negative emotions, such as anger and sadness, have a significantly lower accuracy recognition rate than positive ones.

The instantaneous emotion recognition model presented in this work is based on the work of Loconsole et al. [28]. In the referred work, an emotion classifier (namely, a random forest) based on geometric facial features is trained and used to differentiate images of faces expressing five emotional states: Joy, Sorrow, Surprise, Fear, Disgust and Anger. The authors analyze the accuracy their model achieved with and without calibration with neutral faces and considering different quantities of learned facial expressions. Also, they compare the accuracy of their model with that of other authors' models, and conclude their model achieved higher accuracy for the experiments made.

III. Methodology

This section briefly introduces the methods and techniques used to implement each of the steps shown in the diagram of Figure 1.

A. Face Detection

In this step, the Chehra Face Tracker is used [29]. This tracker detects and keeps track of faces in input images. It can be classified as a discriminative tracker, as it uses facial landmarks and discriminative functions to describe the current state of the face of a person, rather than a generative tracker, which would seek parameters that would maximize the probability of the deformable model to reconstruct a given face [29].

The Chehra Face Tracker uses an incremental parallel cascade of linear regressions to train the model, which has a better performance on face tracking in videos when compared to both the parallel cascade of linear regressions and the sequential cascade of linear regressions, showing better adaptation over time and robustness to environment changes on the face [29].

The tracker is capable of handling new training samples without having to retrain the model from scratch. It can also automatically tailor the model to the subject being tracked and to the imaging conditions, hence becoming person-specific over time [29].

B. Feature Extraction

Once the face tracker is able to fit the face model on one of the found faces in the image, one can proceed to extract features of interest from it.

The process of choosing what features to extract is not trivial, as the chosen feature set should be one that describes the studied concepts (in this case, the five facial emotion expressions: Happiness, Sadness, Anger and Fear, plus the Neutral state), so the trained classifier may have a better chance of learning how to properly differentiate amongst samples of these concepts. Loconsole [28] presents a feature set which is intended to differentiating among facial displays of Ekman's six basic emotions. This set comprises of two kinds of features: linear features and eccentricity features. While the linear features are determined by calculating the normalized linear distances between two given landmarks outputted by the face tracking model, the eccentricity features are given by the eccentricity measures of ellipses fitted over groups of three facial landmarks.

In the present work, Loconsole feature set is adopted with some new features added to it. The added features were chosen based on facial cues Ekman found to be of relevance in the process of facial emotion recognition [4]. The complete set of features adopted is described in Table I (refer to Figure 2 for the landmark's labels referenced in the table).

Table I: Extracted feature set

Name	Measure	By
F1	$\overline{UEBl_{m7y}Ul_{m3y}}/DEN$	[28]
F2	$\overline{U_{m1y}SN_y}/DEN$	[28]
F3	$\overline{D_{m2y}SN_y}/DEN$	[28]
F4	$\overline{EBlr_{Mx}EBrl_{Mx}}/DEN$	Us
F5	$\overline{A_{My}D_{m2y}}/DEN$	Us
F6	$\overline{B_{My}D_{m2y}}/DEN$	Us
F7	$\overline{A_{My}U_{m1y}}/DEN$	Us
F8	$\overline{B_{My}U_{m1y}}/DEN$	Us
F9	$\overline{EBlr_{My}Elr_{My}}/DEN$	Us
F10	$\overline{EBrl_{My}Erç_{My}}/DEN$	Us
F11	$\angle(A_m, D_{m2}, B_m)$	Us
F12	$\angle(A_M, U_{m1}, B_M)$	Us
F14	$\angle(EBll_M, EBl_{aux}, EBlr_M)$	Us
F13	$\angle(EBrr_M, EBr_{aux}, EBrl_M)$	Us
F15	$\angle(EBllm_m, EBlr_M, EBl_{aux})$	Us
F16	$\angle(EBrr_M, EBrl_M, EBr_{aux})$	Us
F17	$Ecc(A_M, B_M, D_{m2})$	[28]
F18	$Ecc(A_M, B_M, D_{m2})$	[28]
F19	$Ecc(Ell_M, Elr_M, UEl_{m3})$	[28]
F20	$Ecc(Ell_M, Err_M, DEl_{m4})$	[28]
F21	$Ecc(Erl_M, Err_M, UEr_{mr})$	[28]
F22	$Ecc(Erl_M, Err_M, UEr_{m6})$	[28]
F23	$Ecc(EBll_M, EBlr_M, UEBl_{m7})$	[28]
F24	$Ecc(EBrl_M, EBrr_M, UEBr_{m8})$	[28]

In Table I, $\overline{(P_1P_2)}$ represents the linear distance between points P_1 and P_2, and the indices x and y are used to represent the horizontal and vertical points' coordinates, respectively. The notation $\angle(P_1, P_2, P_3)$ represents the internal angle between points P_1, P_2 and P_3, in radians. Finally, $Ecc(P_1, P_2, P_3)$ represents the eccentricity of an ellipse fitted over the points P_1, P_2 and P_3. The measure of eccentricity of an ellipse is given by the formula below (refer to Figure 3).

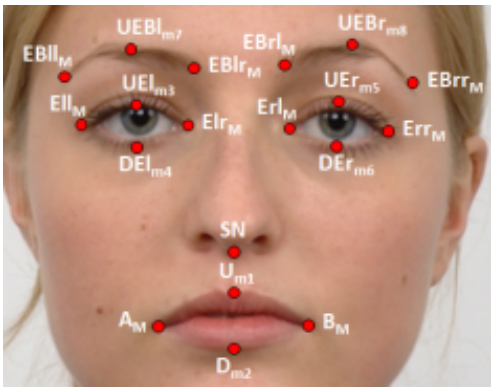


Fig. 2: The facial landmarks considered for the feature extraction process (taken from [28]).

$$Ecc(P_1, P_2, P_3) = \sqrt{\frac{\left(\frac{P_{1x} - P_{3x}}{2}\right)^2 + \left(\frac{P_{1y} - P_{2y}}{1}\right)^2}{\left(\frac{P_{1x} - P_{3x}}{2}\right)^2}} \quad (1)$$

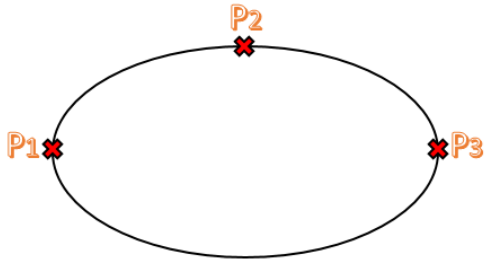


Fig. 3: An ellipse and the necessary points to the calculation of its eccentricity.

Feature F4 is a measure of the horizontal distance between the inner points of the eyebrows. This distance should be smaller in angry faces (which usually present the inner points of the eyebrows closer together) and bigger in surprised faces (which usually present the inner points of the eyebrows farther apart), for example.

Features F5 and F6 are measures of the vertical distances between the leftmost and the rightmost points of the mouth and the bottommost point of the mouth, respectively. These features should be helpful in differentiating facial expressions that present open mouths (like an angry expression, with exposed teeth) and closed mouths (like in a neutral expression). Also, they should be helpful in detecting if the analysed face is currently speaking or not.

Features F7 and F8 are similar to F5 and F6, but measure the vertical distances between the leftmost and the rightmost points of the mouth and the topmost point of the mouth. They have the same purpose features F5 and F6 have.

Features F9 and F10 are measures of the vertical distance between the inner points of the eyebrows and the inner points of the eyes. These features should help to differentiate facial expressions that present the inner corners of the eyebrows lifted (like in a surprised expression) from facial expressions that present the inner corners of the eyebrows lowered (like in an angered expression).

Feature F11 is the measure of the inner angle formed by the leftmost and rightmost points of the mouth with the bottommost point of the mouth. Feature F12 is the measure of the inner angle formed by the leftmost and the rightmost points of the mouth with the topmost point of the mouth. Together, they should be helpful in describing if the mouth is closed of opened, similarly to the features F5 to F8.

Features F13 and F14 are the measures of the inner angles formed by the corner of the eyebrows with the central point of each eyebrow. They should be helpful in describing if the eyebrows are arched (like in a surprised facial expression) or flat (like in an angered expression).

Features F15 and F16 are the measures of the inner angles formed by the outter corner and center points of the eyebrows with the inner corners of the eyebrows. They have the same purpose of the features F13 and F14.

Some of the points used to calculate the features aren't directly output by the face tracker algorithm adopted in this work, and must be calculated before the features can be computed. These points are: UEl_{m3}, UEr_{m5}, EBl_{aux} and EBr_{aux}. The Equations 2, 3, 4 and 5 describe how each of these points are obtained. EBl_{aux} and EBr_{aux} are not facial landmarks, but auxiliar points used in conjunction with the landmarks to calculate some of the chosen features.

$$UEl_{m3} = \left(\frac{Ell_{Mx} + Elr_{Mx}}{2}, \frac{Ell_{My} + Elr_{My}}{2}\right) \quad (2)$$

$$UEr_{m5} = \left(\frac{Erl_{Mx} + Err_{Mx}}{2}, \frac{Erl_{My} + Err_{My}}{2}\right) \quad (3)$$

$$UEr_{m5} = (Elr_{Mx} - Erl_{Mx} + EBlr_{Mx}, \; Elr_{My} - Erl_{My} + EBlr_{My}) \quad (4)$$

$$UEr_{m5} = (Erl_{Mx} - Elr_{Mx} + EBrl_{Mx}, \; Erl_{My} - Elr_{My} + EBrl_{My}) \quad (5)$$

C. Databases

Once the feature set is chosen, the next step is to choose one or more databases to extract these features from. These databases should contain samples of all of the concepts the machine learning algorithm is expected to learn.

In the present work, both Cohn-Kanade Plus[30] and MMI Facial Expression [31] Databases are used to train the instantaneous facial emotion classifier model.

The Cohn-Kanade Plus (or CK+) Database comprises of 486 sets of pictures from 97 posers. Each set contains a sequence of pictures depicting a person acting the onset of a particular target emotion and each sequence is labeled as a sample of that particular represented target emotion. All of the sets start with a neutral expression and evolve into a particular target emotion expression.

The CK+ Database contains, but is not limited to, sequences of all of the studied basic emotions, that is: Happiness, Sadness, Anger and Fear; but doesn't contain sets labeled as Neutral. For the purpose of this work, for each selected set, the first picture of the sequence is taken as a Neutral sample and the last picture of the sequence is taken as a sample of the sequence's target emotion. To avoid one emotion being predominant over the others in the training set, which could degrade the quality of the training process, the limit of samples for each target emotion is set to be the number of samples available for the scarcer target emotion. After the features are extracted from the chosen sets, 129 samples are generated by this process.

The MMI Facial Expression Database comprises of over 2900 videos and images of 75 posers. Only part of these videos are labeled as samples of basic emotion, so just a subset of the database is effectively utilized in this work. The selected videos show humans acting a full emotional cycle of a particular target emotion, that is, all of the three phases of the emotional display are represented: onset, apex and offset. All of the selected videos start with a Neutral face expression, which progresses to a target emotion expression and then regresses back to the Neutral display.

Similarly to the CK+ Database, the MMI Facial Expression Database contains, but is not limited to, videos of all of the studied basic emotions, but doesn't contain samples of Neutral displays. Since the videos aren't labeled at a frame-level and considering there is no preliminary indication of which of the frames represent the emotion's apex, one must first manually annotate the frames' target emotions before they can extract the features from them.

That said, all of the 74 videos chosen from this database were annotated in the following manner: the authors would watch the videos and pinpoint four instants of interest. The first instant (referred to as t_1 from here forth) represents the start of the emotional onset in the video; the second instant (t_2) represents the emotional onset's ending and the beginning of the apex; the third instant (t_3) represents the apex's ending and the beginning of the emotional offset; finally, the fourth instant (t_4) represents the emotional offset's ending.

With these instants annotated, a frame-level categorization of the videos is created: the frames before t_1 and after t_4 (inclusive) are classified as Neutral samples; the frames between t_2 and t_3 (inclusive) are classified as that video's target emotion samples; finally, the frames between t_1 and t_2 and t_3 and t_4 are classified partially as Neutral samples and partially as that video's target emotion samples.

However, not all of the generated samples were used to train the classifier. The first and the last frames of each video were chosen to compose the Neutral set of the database; also, windows of size $n = 10frames$ were built around the center of the apex region (that is, around the middle frame between t_2 and t_3) in each video, and all of the frames within these windows were taken as samples of that video's target emotion. The value of n was chosen empirically, and aimed to stablish a balance between the quantity of Neutral samples and the quantity of the other four emotions' samples. Also, care was taken so the created windows would never exceed their boundaries, that is, a window would never start at an instant before t_2 nor would it end after t_3.

After the features are extracted from the chosen pictures, 809 samples are generated by the described process.

It's worth saying both of the adopted databases contain videos and images of faces in profile and in other non-frontal orientations. However, different head orientations may cause the selected features to vary considerably for samples of the same target emotion. This could hamper the classifier's learning process and, for that reason, only videos and images containing emotional displays in frontal-oriented faces are used to train the classifier.

Finally, one should take note that all the sample images contained in these databases were acted, and not naturally elicited.

D. Instantaneous Facial Emotion Recognizer

The instantaneous facial emotion recognizer is a machine learning algorithm trained over the training set extracted through the previously described procedure.

The Random Forest learning algorithm is adopted in this work, as it was shown to have good accuracy on the work of Loconsole [28] when compared to other algorithms. The learner's accuracy and other statistics of interest are presented further in Section IV

The information fed into the dynamic classifier, however, is not simply the category output by the instantaneous classifier for a given sample, but rather, a measure of confidence that the classifier has for that sample to belong to each of the considered classes. The confidence measure used was the normalization of the number of votes each class received by the weak learners. Suppose, as an example, that a particular sample is classified by a random forest containing 100 random trees, and that 70 trees vote for the sample to belong to the Happiness class and the rest of the trees vote for it to belong to the Neutral class; in that case, the confidence measure for the sample to belong to the Happiness class would be 70%, the confidence measure for the sample to belong to the Neutral class would be 30% and the confidence measure for the sample to belong to the other classes would be 0. So, given a sample S_1, the output of the instantaneous classifier that is fed into the dynamic model is a vector

of the form $V_1 = (Pr_{1n}, Pr_{1h}, Pr_{1s}, Pr_{1a}, Pr_{1f})$, where $Pr_{1n}, Pr_{1h}, Pr_{1s}, Pr_{1a}$ and Pr_{1f} are the confidence levels for S_1 to belong to the Neutral, Happiness, Sadness, Anger and Fear categories, respectively.

E. Kalman Filter

After the instaneous facial emotion classifier is properly trained, its outputs can be fed into the dynamic classifier, which will output the model's final prediction for the samples. However, aiming to eliminate high frequency noises, these outputs are firstly processed by a Kalman filter before they are inserted into the dynamic model. This section describes this filter and highlights the advantages of its usage.

As a natural consequence of the use of video frames to analyze the facial features of a person, different sources of noise can affect the classification algorithm.

It is assumed that the emotions are represented by the data initially fed in the training phase, which are gathered under controlled conditions; thus, effects such as face deformation resulting from speech, light source variations and unexpected face motions should be minimized. Furthermore, the objective of the model is to enhance the presentation of the slow and continuous emotions in spite of the instantaneous ones, sp a low pass filter should be used.

Kalman filtering is the solution proposed to this model, being a filter that has a good performance on linear systems with zero mean Gaussian noise on both the model and in the process of data acquisition. The empirical evidence presented in [9] supports this choice.

Being x_s the state variable of a linear system and y, the output of the filter for a single emotion, the filtered signal related to one of the emotions being analyzed, 6 and 7 describe the Kalman filter.

$$\dot{x_s} = x_s \tag{6}$$

$$\dot{F_a} = y = \frac{Kx_s}{\tau} \tag{7}$$

In the above equations, K is the filter's gain and τ is the time constant. There are two steps for the filtering, the first being the prediction step and the second the update step. The update is only run when new information from the sensors – in this case the output of the instantaneous emotion analyzer – is available. If the delay between data acquisitions is higher than the delay between filter steps calculations there will be some steps in which only the prediction steps will be run.

The prediction step is described by 8 and 9, where $x_{s,t}$ is the current state $x_{s,t-1}$ is the previous state, w is the noise covariance, p the covariance of the state variable on the t state. Note that the update step always assumes that the state variable has not changed, only the covariance of the system.

$$x_{s,t} = x_{s,t-1} \tag{8}$$

$$p = p + w/\tau^2 \tag{9}$$

The update step is described by equations 10, 11 and 12, where m is the residual covariance, v is the covariance of the observation noise and r_t and y_t are the filter input and output at instant t. This input corresponds to the output of the instantaneous emotion classifier. The state variable now has its value updated and, consequently, the output of the Kalman filter has its value proportionally changed.

$$m = \frac{\frac{pK}{\tau}}{p\left(\frac{K}{\tau}\right)^2 + v} \tag{10}$$

$$x_{s,t} = x_{s,t} + m(r_t - y_t) \tag{11}$$

$$p = (1 - \frac{mk}{\tau})p \tag{12}$$

Note that these equations describe the filtering process for a single class (that is, the filtering of outputs of a particular emotion). The full model is represented by applying these equations for each emotion separately.

However, neither w nor v are known, and have to be estimated by an optimization algorithm, which is described in Section III-G.

F. Dynamic Model

After the instantaneous output is filtered, it is ready to be fed into the dynamic model.

The dynamic model proposed here does not aim to describe rapid emotional variations a person may be subject to, but rather, it tries to describe more lasting emotional states. Suppose, for illustration purposes, that a man is talking to a dear friend of him that he has not seen for a while. One may expect the overall conversation to elicit a pleasant emotion. However, during this conversarion, he happens to see a person throwing trash in the street; it infuriates him for a while, but he rapidly get back to talking to his friend and forgets the sight that angered him. If pictures of his face were fed into the proposed model during this entire event, one should expect the model to detect the overall pleasant emotion of the conversation (that is, if it was pleasant enough so that his facial expression indicated so); however, his temporary enragement should not modify the output of the model.

The dynamic model is based on the work of [9], and utilizes the concept of Dynamic Emotional Surfaces (DESs).

As name indicates, DESs are surfaces that aim to describe the dynamics of transitions between different emotional states. In this work, a planar surface is adopted, and it is partitioned in four quadrants, one for each of the considered basic emotions: Happiness, Sadness, Anger and Surprise. Centered in the intersection of the four areas, there is the Neutral area, which represents the absence of emotions. Figure 4 illustrates the model's DES.

Located on the $+45°$ and $-45°$ diagonals of this plane, there are four Emotional Attractors (EAs), one for each of

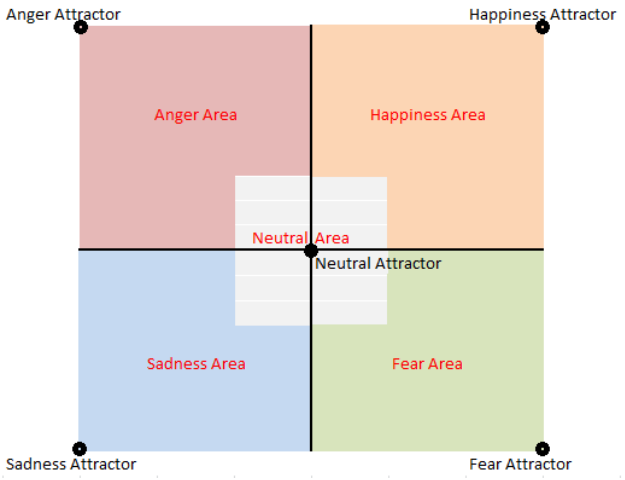


Fig. 4: A representation of the planar DES used in this work.

the considered non-Neutral emotions, and each located in its corresponding quadrant. Refer to Figure 4. The Happiness, Fear, Sadness and Anger attractors are located on the points $PA_{Happiness}$, PA_{Fear}, $PA_{Sadness}$ and PA_{Anger}, respectively, and the Neutral attractor is located on the point $PA_{Neutral}$.

Let to slide upon the plane, there are Emotional Particles (EPs), one for each analyzed subject. The location of a particle in a given instant indicates the model's output emotion for that instant, according to the equation below, where $P = (P_x, P_y)$ is an EP's position and $f(P)$ is the model's output in the considered instant.

$$f(P) = \begin{cases} \text{Happiness}, & \text{if } P_x > K_{nr} \text{ and } P_y > K_{nr} \\ \text{Sadness}, & \text{if } P_x < -K_{nr} \text{ and } P_y < -K_{nr} \\ \text{Anger}, & \text{if } P_x < -K_{nr} \text{ and } P_y > K_{nr} \\ \text{Fear}, & \text{if } P_x > K_{nr} \text{ and } P_y < -K_{nr} \end{cases} \tag{13}$$

In the equation 13, K_{nr} is a constant that determines the width and height of the Neutral area.

The EAs are responsible for pulling EPs towards them. The stronger the confidence level the Kalman filter outputs for a given emotion, the stronger the pull velocity for that emotion's attractor will be. If at the instant $\bar{t}$ Kalman filter outputs a confidence level of $Pr_E(\bar{t})$ for emotion E, then E's attractor velocity, $VA_E(\bar{t})$, is given by Equation 14.

$$VA_E(\bar{t}) = K_{avm} Pr_E(\bar{t}) \tag{14}$$

The parameter K_{avm} is the attractors velocity modulator parameter, which value, like the Kalman filter parameters w and v, is also found via an optimization algorithm.

The dynamics for EPs are described by equations 15 and 16, where $P(t)$ and $V(t)$ are particles' position and velocity, $Pr_E(t)$ is the confidence measure for emotion E and $VA_E(t)$ is the attractor's E pull velocity, all at instant t.

$$P(t) = P(t-1) + V(t) \tag{15}$$

$$V(t) = \begin{cases} VA_{Neutral}, & \text{if } max(Pr_E(t)) = Pr_{Neutral}(t) \\ \sum\limits_{E=\bar{e}} VA_E, & \text{if not} \end{cases} \tag{16}$$

where $\bar{e}$ is the subset $\{Happiness, Sadness, Anger, Fear\}$. Also, the position of the particle is never let to exceed the rectangle delimited by the four non-Neutral EAs.

The noise smoothing introduced by the Kalman filter and the intrinsic inertia presented by the proposed model make so that natural facial noises - like the mouth movements caused by laughter or speech - should have its influence on the predictions diminished, when in comparison to the instantaneous classifier.

G. Parameters Optimization

As the previous sections explained, some of the model parameters can not be known *a priori*, and are better defined via an optimization process. These parameters are: Kalman filter's noise covariance (w), Kalman filter's covariance of the observation noise(v) and DES's attractors velocity modulator (K_{avm}).

The optimization process here adopted is based on the simulated annealing algorithm, and can be described by the pseudo-code presented below.

In the above pseudo-code, T_0, T_{room} and T_{curr} are the initial, room and current temperatures of the optimizer, in that order; p_curr, p_{la} and p_{sol} are the current iteration's parameters, the last accepted solution's parameters and the final solution parameters, respectivelty; e_0, e_{curr}, e_{la} and e_{sol} are the initial energy, the current iteration's energy, the last accepted solution's energy and the final solution's energy, in that order; dr is the temperature decay rate and Pr_{acc} is the probability that a solution will be accepted by the algorithm.

Note that an iteration's energy, e_{curr} is obtained by the function $calculateEnergy(p_{curr}, dataset)$, which considers both the current value of the parameters being optimized and a dataset chosen for the optimization. The MMI Facial Emotion Database's previously selected 74 videos were used to extract the energy measure; however, this time they were considered in their full-length. The adopted energy measure is the number of frames the model misclassified in the iteration.

A proposed solution is always accepted if it causes the system's energy to decrease in comparison to the last accepted solution's energy. However, even if a solution causes the energy to increase, it has a chance of being accepted that is proportional to the iteration's current temperature and inversely proportional to the energy increase it causes. This measure helps the optimizer to avoid getting stuck in local minima.

If a solution is accepted, its parameters and energy are stored to serve as comparison data for the next iteration. However, a solution is only stored as a final solution if its energy is smaller than the last accepted final solution.

// Initializations:
$T_0 = 200°C$;
$T_{room} = 20°C$;
$T_{curr} = T_0$;
$p_{curr} = randomizeParameters()$;
$p_{la} = p_{curr}$;
$p_{sol} = p_c$;
$e_0 = +\infty$;
$e_{curr} = e_0$;
$e_{la} = e_0$;
$e_{sol} = e_0$;
$dr = 0.99995$;
// Iterations:
while $(T_{curr} > T_{room})$ **do**
  $e_c = calculateEnergy(p_{curr}, dataset)$;
  if $(e_{curr} < e_{la})$ **then**
    $Pr_{acc} = 1$;
  else
    $Pr_{acc} = e^{(e_{la} - e_{curr})/T_{curr}}$;
  end
  if $(Rnd(0,1) > Pr_{acc})$ **then**
    $p_{la} = p_{curr}$;
    $e_{la} = e_{curr}$;
    if $(e_{la} > e_{sol})$ **then**
      $p_{sol} = p_{la}$;
      $e_{sol} = e_{la}$;
    else
      $p_{curr} = p_{la}$;
    end
  end
  $p_{curr} = moveAround(T_{curr})$;
  $T_{curr} = T_{curr} \times dr$;
end

At the end of every iteration, the parameters are varied through the function $moveAround(T_{curr})$, which takes into consideration the iteration's temperature - the higher the temperature, the more the parameters are allowed to variate -, and the optimizer temperature is made to decay by a constant rate dr.

IV. Tests and Results

This section presents the results of the tests realized on the model. These tests are presented separately for the instantaneous facial emotion classifier, for the parameters optimization algorithm and for the dynamic facial emotion classifier.

A. Tests on the Instantaneous Facial Emotion Classifier

Tests were made to measure the quality of the instantaneous classifier. Since a poorly trained classifier may compromise the overall performance of the model, the quality of its outputs should be analyzed with caution.

The random forest learning algorithm discards the need for procedures like cross-validation, bootstrap or separate test sets for estimating the classifier's accuracy. During the training of each of the weak learners (that is, of each tree of the forest), an out-of-bag set (or "oob set", containing roughly 1/3 of the complete training set) is created for that learner. The oob set is used to validate the accuracy of that particular tree. After all the trees have finished training, the following procedure is used to calculate an estimation of the accuracy of the learner for samples stranger to the training set:

1) Each sample contained in the complete training set is considered separately;
2) All trees that contain a particular sample in their oob sets are used to classify that sample, and a vote counting is used to decide to what class it belongs to. The procedure is repeated for all samples of the complete training set;
3) The random forest's accuracy is given by the number of samples of the set classified correctly divided by the number of samples classified incorrectly by that process...

Through the described procedure, an accuracy estimate of approximately 90% was obtained.

The analysis of the learner's confusion matrix allows one to observe how its predictions are distributed amongst the different classes. Table II presents the confusion matrix for the trained random forest.

Table II: The confusion matrix for the trained random forest

	Neutral	Anger	Fear	Happiness	Sadness
Neu.	201	5	4	8	20
Ang.	0	168	1	0	5
Fea.	0	0	168	0	1
Hap.	0	0	2	189	0
Sad.	0	1	0	0	113

One can notice that the Neutral class is the one with more misclassified samples, even if considering relative numbers. Also, more than half of the misclassified Neutral samples are categorized as Sadness samples, which suggests that the boundaries between these two classes is the less obvious for the classifier, at least on the considered dataset.

B. Tests on the Parameters Optimization Algorithm

Tests were made with values of K_{nr} (which determines the height and width of the neutral area of the plane) varying from 1 to 5, with unitary increments. Also, for all the tests, the attractors were positioned on the points $PA_{Happiness} = (10, 10)$, $PA_{Fear} = (10, -10)$, $PA_{Sadness} = (-10, -10)$, $PA_{Anger} = (-10, 10)$ and $PA_{Neutral} = (0, 0)$. Figure 5 shows the model's accuracy history for the best optimization achieved - that is, for the optimization that reached the lowest energy on the used dataset.

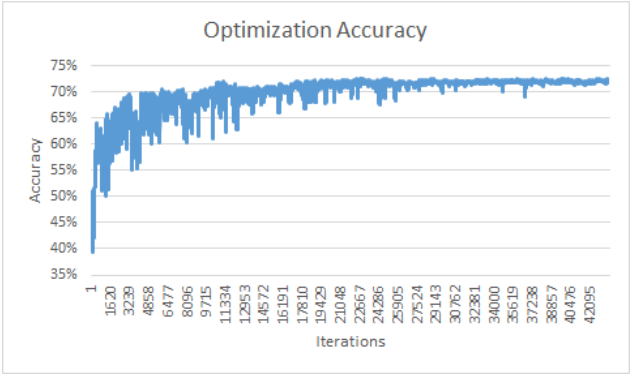


Fig. 5: The accuracy curve for the best optimization case.

It is possible to see an overall increase of the optimization accuracy as the iterations progress; also, the accuracy curve seems to converge to a value of about 72% by the end of the process. The accuracy achieved with the instantaneous model for the same dataset is of 64%. This fact suggests that the use of the dynamic model was beneficial even for a dataset with videos that do not contain too many facial noises caused by factors like laughter or speech.
The best accuracy was reached for a value 1 of K_{nr}.

C. Tests on the Dynamic Facial Emotion Classifier

To test the developed dynamic classifier, a test was run on the video "S43_an_2" of the eNTERFACE'05 Database [32], the same analyzed in the work of [9].
This video depicts the face of an angered person as she irritatedly proclaims a certain sentence. The presence of facial noises in the video is relevant for the experiment, as it allows for the analysis of how well the dynamic model is able to deal with such noises. Also, this is the first experiment that utilizes a video entirely stranger to the datasets used for training the instantaneous classifier and for the optimization process.
Because the video "S43_an_2" is simply classified as an Anger video, and since there is no information about whether any other emotional displays are considered to be present in it, all of its frames were considered as Anger samples and fed into the model.
Figures 6 and 7 present the dynamic model output and the instantaneous classifier output for each frame of the video, respectively. The accuracy achieved with the dynamic model was of 89%, while the accuracy achieved with the instantaneous classifier was of 64%. This result suggests that the dynamic model successfuly dealt with a considerable portion of the facial noises presented in the video. Note that not only the dynamic model achieved a higher accuracy on the video, but its outputs seem to be more reliable. With exception of the last frame, all frames in the video were classified as Anger or Neutral frames by the dynamic model, and there are less variations between different emotional states; the classifications attributed by the instantaneous model, however, flicker more rapidly and between a larger number of emotional states. One could argue that the result achieved by the dynamic model is more useful than the one achieved by the instantaneous classifier if it was to be used to control an automated system like a social robot - maybe the social robot would not be able to react as well to a flickering input as it would react to a more stable one.

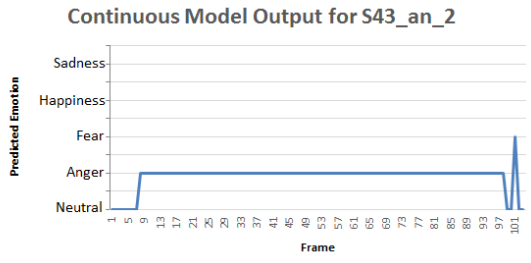


Fig. 6: The output of the continuous model for the video "$S43\_an\_2$".

Finally, Figure 8 presents the trajectory on which the EP traveled throughout the video. Note that the particle rapidly

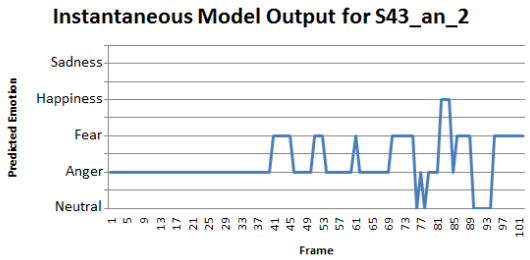


Fig. 7: The output of the instantaneous classifier for the video "$S43\_an\_2$".

progresses to the Anger area, where it remains until the latter parts of the video, regressing back to the neutral area and then to the Fear area by the end of the video. The transition to the Fear area is probably due to the considerably large number of Fear predictions outputted by the instantaneous classifier in the latter parts of the video.

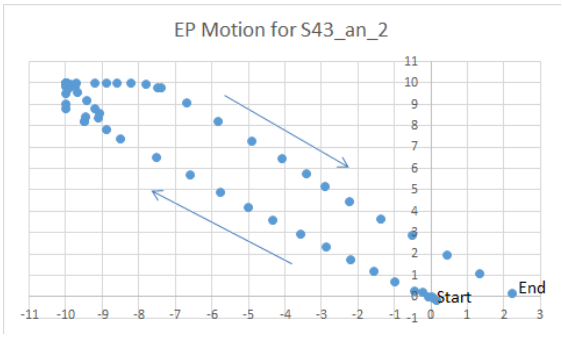


Fig. 8: Motion of the EP throughout the video "$S43\_an\_2$".

V. Conclusions

In the present work, an innovative dynamic emotion recognition model was presented. This model comprises of the conjugation of a machine learning algorithm, a Kalman filter and an original dynamic model that aims to describe durable emotional states and to minimize facial noises like deformations caused by laughter and speech. A simulated annealing algorithm was utilized to optimize the model's parameters.
The model has shown good performance when compared to the instantaneous emotion classifier trained in the present work: while the former achieved an accuracy rate of 72% over the chosen dataset, the latter presented an accuracy rate of just 64%, on the same dataset.
When tests on a sample stranger to the datasets utilized to train the instantaneous classifier and to optimize the model's parameters, the dynamic model once again outmatched the instantaneous model: not only it achieved a higher accuracy rate (89% against 64%), but it also provided a much more stable output.
As target objectives for future works, the following tasks are proposed:

1) Execute more tests on the dynamic model, in order to better analyze its accuracy and the way it describes the progression of emotional expressions in faces;

2) Utilize larger datasets to train the instantaneous model and to optimize the dynamic model;
3) Utilize datasets that contain faces deformed by natural facial noises, like laugther or speech, for the training and optimization of the model;
4) Study possible changes the proposed planar DES may need to better describe the way emotions manifest themselves in human faces;
5) Increase the number of considered emotions and study how the DES should be changed to accommodate this change.

REFERENCES

[1] P. Ekman, "Basic emotions," in *Handbook of Cognition and Emotion*, T. Dalgleish and M. J. Power, Eds. Chichester, UK: John Wiley Sons, Ltd, 2005, ch. 3, p. 45–60.

[2] R. W. Picard, "Affective computing," MIT Media Lab, Perceptual Computing, Cambridge, MA, Tech. Rep. 295, 1995.

[3] P. Ekman, E. R. Sorenson, and W. V. Friesen, "Pan-cultural elements in facial displays of emotion," *Science*, vol. 164, no. 3875, pp. 86–88, 1969.

[4] P. Ekman, "A linguagem das emoções," *São Paulo: Lua de Papel*, 2011.

[5] P. Ekman and W. V. Friesen, *Unmasking the face: A guide to recognizing emotions from facial clues*. Ishk, 2003.

[6] R. A. M. Gonçalves, "Um modelo matemático para inferência computacional de estado emocional a partir de detectores de expressões faciais," M. Eng. thesis, Universidade de São Paulo, 2012.

[7] Z. Zeng, M. Pantic, G. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 31, no. 1, pp. 39–58, 2009.

[8] M. Zarkowski, "Identification-driven emotion recognition system for a social robot," in *Methods and Models in Automation and Robotics (MMAR), 2013 18th International Conference on*. IEEE, 2013, pp. 138–143.

[9] R. A. M. Gonçalves, D. R. Cueva, M. R. Pereira-Barretto, and F. G. Cozman, "A model for inference of emotional state based on facial expressions," *Journal of the Brazilian Computer Society*, vol. 19, no. 1, pp. 3–13, 2013.

[10] A. Rabie, B. Wrede, T. Vogt, and M. Hanheide, "Evaluation and discussion of multi-modal emotion recognition," in *Computer and Electrical Engineering, 2009. ICCEE'09. Second International Conference on*, vol. 1. IEEE, 2009, pp. 598–602.

[11] M. K. Greenwald, E. W. Cook, and P. J. Lang, "Affective judgment and psychophysiological response: Dimensional covariation in the evaluation of pictorial stimuli," *Journal of psychophysiology*, 1989.

[12] J. A. Russell and A. Mehrabian, "Evidence for a three-factor theory of emotions," *Journal of research in Personality*, vol. 11, no. 3, pp. 273–294, 1977.

[13] D. Sander, D. Grandjean, and K. R. Scherer, "A systems approach to appraisal mechanisms in emotion," *Neural networks*, vol. 18, no. 4, pp. 317–352, 2005.

[14] M. Pantic and L. J. M. Rothkrantz, "Automatic analysis of facial expressions: The state of the art," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 22, no. 12, pp. 1424–1445, 2000.

[15] P. Ekman and W. V. Friesen, "Facial action coding system," 1977.

[16] B. Tinen, "Sistema de identificação de emoções por expressões faciais com operação ao vivo," Eng. thesis, Universidade de São Paulo, 2014.

[17] P. Ekman and W. Friesen, "Facial action coding system: A technique for the measurement of facial movements," *Consulting Psychologist*, vol. 2, 1978.

[18] M. Valstar and M. Pantic, "Fully automatic facial action unit detection and temporal analysis," in *Computer Vision and Pattern Recognition Workshop, 2006. CVPRW'06. Conference on*. IEEE, 2006, pp. 149–149.

[19] B. Fasel and J. Luettin, "Automatic facial expression analysis: a survey," *Pattern recognition*, vol. 36, no. 1, pp. 259–275, 2003.

[20] A. K. Jain and S. Z. Li, *Handbook of Face Recognition*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2005.

[21] G. Donato, M. S. B. J. C. Hager, P. Ekman, and T. J. Sejnowski, "Classifying facial actions," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 21, no. 10, pp. 974–989, 1999.

[22] S.-S. Liu, Y.-T. Tian, and D. Li, "New research advances of facial expression recognition," in *Machine Learning and Cybernetics, 2009 International Conference on*, vol. 2. IEEE, 2009, pp. 1150–1155.

[23] J. Hamm, C. G. K. R. C. Gur, and R. Verma, "Automated facial action coding system for dynamic analysis of facial expressions in neuropsychiatric disorders," *Journal of neuroscience methods*, vol. 200, no. 2, pp. 237–256, 2011.

[24] T. F. Cootes, J. C. Taylor, D. H. Cooper, and J. Graham, "Active shape models-their training and application," *Computer vision and image understanding*, vol. 61, no. 1, pp. 38–59, 1995.

[25] S. Du, Y. Tao, and A. M. Martinez, "Compound facial expressions of emotion," *Proceedings of the National Academy of Sciences*, vol. 111, no. 15, pp. E1454–E1462, 2014.

[26] M.-H. Yang, D. J. Kriegman, and N. Ahuja, "Detecting faces in images: A survey," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 24, no. 1, pp. 34–58, 2002.

[27] J. A. Russell, "Is there universal recognition of emotion from facial expressions? a review of the cross-cultural studies." *Psychological bulletin*, vol. 115, no. 1, p. 102, 1994.

[28] C. Loconsole, C. R. Miranda, G. Augusto, A. Frisoli, and V. C. Orvalho, "Real-time emotion recognition-novel method for geometrical facial features extraction." in *VISAPP (1)*, 2014, pp. 378–385.

[29] A. Asthana, S. Zafeiriou, S. Cheng, and M. .Pantic, "Incremental face alignment in the wild," *Science*, 2014.

[30] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*. IEEE, 2010, pp. 94–101.

[31] M. Pantic, M. Valstar, R. Rademaker, and L. Maat, "Web-based database for facial expression analysis," *Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on*, 2005.

[32] O. Martin, I. Kotsia, B. Macq, and I. Pitas, "The enterface'05 audio-visual emotion database," in *Data Engineering Workshops, 2006. Proceedings. 22nd International Conference on*. IEEE, 2006, pp. 8–8.