Medical Data Classification Using Similarity Measure of Fuzzy Soft Set Based Distance Measure

Saima Anwar Lashari, Rosziati Ibrahim, and Norhalina Senan
Faculty of Computer Science and Information Technology,
Universiti Tun Hussein Onn Malaysia,
86400 Parit Raja, Batu Pahat, Johor, Malaysia.
saima@uthm.edu.my

***Abstract*—Medical data classification plays a crucial role in many medical imaging applications by automating or facilitating the delineation of medical images. A considerable amount of literature has been published on medical images classification based on data mining techniques to develop intelligent medical decision support systems to help the physicians. This paper assesses the performance of a new classification algorithm using similarity measure fuzzy soft set based distance based for numerical medical datasets. The proposed modelling comprises of five phases explicitly: data acquisition, data pre-processing, data partitioning, classification using FussCyier and performance evaluation. The proposed classifier FussCyier is evaluated on five performance matrices': accuracy, precision, recall, F-Micro and computational time. Experimental results indicate that the proposed classifier performed comparatively better with existing fuzzy soft classifiers.**

***Index Terms*—Medical Data Classification; Similarity Measure; Fuzzy Soft Set; Distance Measure.**

I. INTRODUCTION

Medical data classification is a type of multifaceted optimization problem which involves accurate and precise diagnosis of diseases. Several data mining techniques have been proposed and exist for medical data classification; however, the classification accuracy of these models is limited and insufficient in cases where the relationships of input/output datasets are complex and/or non-linear [1]. In recent times, more significant emphasis is given to the utilization of classifiers in medical diagnosis. By using quantitative measures and machine learning techniques, medical diagnostic tools provide automated procedures for objective decisions. Accordingly, machine learning techniques provide second perspective in medical data analysis in addition to knowledge-based approaches. Hence, varied methods and approaches are utilized to achieve better accuracy of medical data classification [2].

In data mining, categorization is formulated to make a forecast of the memberships in a group for data instances. This process utilizes complex analysis of data to determine data connections in huge datasets. Due to its complex features, medical databases provide complications for pattern extortion [8]. There are two approaches to data mining: statistical and machine learning algorithms. The processes in data mining are classified into descriptive and predictive (Figure 1). Descriptive mining tasks provide the general data properties in the database. For Predictive mining tasks, inference is made on the data for predictions [9] whereby forecast is made on explicit values based on patterns identified by known results. Descriptive data mining, without having any predefined target, provides characteristics and descriptions for the data set.

Data mining is a recent approach and it encounters challenges. Extracting useful information from data is a complex process as it has to control different data types, progressive deterioration of data mining algorithms, significant data mining results, data mining representation request and results. Most of recent efforts reviewed in this paper are more related to this second direction of development which is predictive, data mining where ideas are motivated from concepts of pattern recognition, image processing, and computer vision). However, with this in mind, it is important to realize that must medical data analysis efforts are heavily influenced, if not fundamentally driven by, the particular image datasets being utilized and the clinical or biological tasks that underlie the need for medical data analysis.

Solutions resulted from classification algorithm are commendable but as of now, none is diverse and flexible to be accepted generally in the medical data classification community. Categorical variables in medical data are occasionally useful to arrive at decisions and to generalize information. Categorical data (e.g. classification of disease and non-disease groups) is handy for data mining technique and also easy to extract medical information. Meanwhile, in conducting comparative studies, classification researches heavily rely on stored repositories of data (such as UCI repository [12] it allows new algorithm ideas to test its plausibly on known problems.

Data mining techniques, which are a recent application in the medical domain, are applied in mining medical data, which comprises of association rule mining for finding frequent patterns, prediction, classification and clustering. To date, there have been many research on this and intelligent and decision support systems have been developed to make more accurate diagnosis and prediction of diseases especially in predicting heart diseases, lung and breast cancer and remote health monitoring.

Table 1 provides the summary of medical data classification regarding the resolved difficulties that are solved, convenience in medical data mining or implementation of the tools. Therefore, selected researches on classification performance of different classifiers are summarised. Here is shown the effort made for data classification. Nevertheless, it is obvious that benchmarking to determine the best classification algorithm for medical data classification is still lacking.

Table 1
Summary of medical data classification

Author(s), Year	Medical Dataset	Technique (s)	Comments
Zuo et al, 2013	Parkinson Disease	Fuzzy K-NN approach	Familiarized an adaptive Fuzzy K-NN approach for diagnosing the disease [31]
Long , 2015	heart disease	rough sets based attribute reduction and interval type-2 fuzzy logic system (IT2FLS)	heart disease diagnosis system using rough sets based attribute reduction an IT2FLS
Ghofrani, 2014	*X-Ray dataset*	*SVM, Euclidean Distance& PNN*	
Polat *et al*. 2007	Breast Cancer and Liver Disorders dataset	Fuzzy-AIRS	Modeling and analysis of medical data

Uncertainties affect the image analysis and the most problematic issue in image analysis and pattern recognition research is classification [3]. Fuzzy set theory is important in formalizing uncertainties for medical diagnosis and prognosis [4-5]. In order to manage uncertainty in the decision making, the use of fuzzy set theory has given rise to a lot of new approaches [6][7][11]. A new method namely Soft Set Classifier (SSC) was put forward by Mushrif et al., [8] to classify natural textures using the notions of soft set theory. Later, Handaga et al., [9] demonstrated Fuzzy Soft Set Classifier (FSSC), a new application of soft set for numerical data classification, emphasizing a more general concept based on similarity measure between two fuzzy soft sets. The method is capable of handling parameters in the form of real numbers but FSSC has high complexity. Therefore, the present research is carried out due to limitations of the earlier studies and lack of work on the similarity fuzzy soft set based distance measure [10]. The purpose of the present study is to enhance the accuracy of the medical data classification by offering a new classifier named as FussCyier. In appraising the performance of FussCyier, the existing fuzzy soft set classifiers SSC and FSSC were set as a benchmark for the proposed FussCyier. Five performance measures were used to evaluate the performances of these three classifiers, which are classification accuracy, precision, recall, F-Macro and computational time.

The rest of the paper is organized as follows. Section 2 presents a brief review of similarity measure fuzzy soft set. Section 3 presents the proposed methodology. Section 4 presents results and discussion. Finally, Section 5 concludes this work with future directions for medical data classification.

II. Similarity Measure Fuzzy Soft Set

Measuring similarity between two entities is a basic task in classification and clustering fields. Similarity measures provide concrete numbers for how many dissimilar patterns, signals, images or sets are alike [11]. Majumdar & Samantra [12] had carried out a research on the similarity of fuzzy soft set and investigate its application in medical diagnosis to ascertain whether a person contracts a certain disease or otherwise.

Baccour et al., [11] showed properties of fuzzy similarities from the literature and discuss their validation to the common existing properties. Kalaiselvi & Inbarani, [13] used fuzzy soft set similarity measure for gene expression data to obtain more accurate predictions on the stages among cancer genes while the informative genes are identified using Entropy filtering.

A measure of similarity or dissimilarity defines the resemblance between two samples or objects. For a similarity measure the resemblance increases with each increment of its value. The opposite applies for a dissimilarity measure (i-e. a distance measure). The less the values are, the more significant the resemblance is. Thus, similarity and distance measures are dual concepts, if they are normalized similarity=1-distance measure and vice versa [11].

III. Proposed Modelling

As shown in Figure 1, the proposed modelling comprises of five phases that are data acquisition, data pre-processing, data partitioning, classification using FussCyier and performance evaluation.

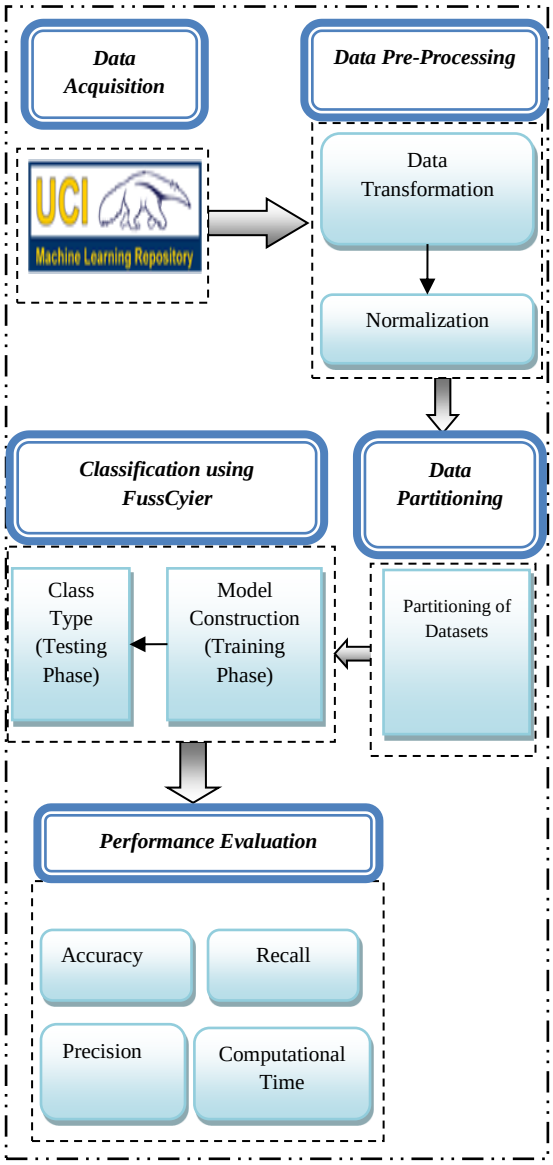

Figure 1: Proposed modelling

Data acquisition is one of the crucial elements to design and develop a successful classifier. Data collections have been done from University of California at Irvine (UCI) machine learning repository. The repository is which makes data collection much easier while also maintaining transparency of any published materials.

Table 2 provides description of all dataset, in which most of the datasets have real numerical features while some have multiclass labels. Dataset includes Pima Indians diabetes, Indian Liver Patient Dataset, liver disorder, Statlog (heart) and dermatology [14].

Table 2
Dataset Description

No.	Dataset	Description
1	Pima Indians Diabetes	i:583 f:8, c:2
2	ILPD	i:583 f:10, c:2
3	Liver Disorder	i:345, f:7, c:2
4	Statlog (Heart)	i:270, f:13, c:2
5	Dermatology	i:366, f:33,c:6

Abbreviations:
instance: i
features: f

Data pre-processing is a step taken on raw data to obtain the best recital of which has significant impact on the performance of classification algorithm. Normalization is a paramount step for data pre-processing because some machine learning methods do not handle continuous attributes, in addition to other important reasons. Firstly, the transformed data in the set of interval $[0,1]$ are more cognitively relevant for human understanding. Secondly, computation process takes place quicker.

Feature fuzzification (normalization) is done by dividing each attribute value with the largest value at each attribute [10].

$$e_{fi} = \frac{e_i}{\max(e_i)} \tag{1}$$

where $e_i, i = 1,2,.....,n$ is the old attribute

e_{fi} is attribute with new value between $[0,1]$

For data partition, a general course in data mining is to segregate data into training and testing sets. Using 70:30 data split between training and testing ensures that maximum posterior classifier will be constructed based on soft set.

Classification using FussCyier involves two steps: model construction (training phase) and class type (testing phase) as shown in Figure 2. For the training phase, the average value of each parameter from all objects with the same class label is calculated to construct fuzzy soft set model as shown in Equation 2.

For testing phase, the FussCyier method applies the distance between two fuzzy soft set as stated in the work of Baccour [9], as illustrated in Equation 3. Since, FussCyier measures the distance between features vectors, intuitively, small distances correspond to higher similarity. Lastly, the maximum score from the distance measure is computed to determine class label for the test data as shown in Equation 4.

Training phase

1. Given N samples obtained from the data class w.
2. Calculate the cluster center vector $E_w \quad i = 1,2,..,N$ using equation below

$$E_w = \frac{1}{N}\sum_{i=1}^{N} E_{wi} \tag{2}$$

3. Obtain Fuzzy soft set model (F_W, E), is a cluster centre vector for every class w having D features
4. Repeat the steps 2 and 3 for all W classes

Testing phase

1. Obtain the unknown class data
2. Obtain a Fuzzy soft set model for unknown class data $(\tilde{G}, E)$, compute similarity measure based on distance between $(\tilde{G}, E)$ and (F_w, E) for each w using Equation (3)

$$S(F,G) = \frac{1}{n}\sum_{i=1}^{n}\left(1 - \left|\mu_F^i - \mu_G^i\right|\right) \tag{3}$$

3. Assign the unknown data to class w if distance measure is maximum

$$w = \arg\left[\max_{w=1}^{W} S(G, F_w)\right] \tag{4}$$

Figure 2: Classification Using FussCyier

Five performance metrics were utilized to evaluate the performance of the proposed FussCyier method: classification accuracy, precision, recall, F-Macro and computational time.

A. Accuracy

The overall accuracy of a classifier is anticipated by dividing the total correctly classified positives tuples and negatives tuples t by the total number of samples. The formula is as follows in Equation 5 [8].

$$(\text{OCA})_i = \frac{\sum_{i=1}^{n}(\text{True Classification})_i}{(\text{Total number of cases})_i} \tag{5}$$

where i the class number
n is the total number of classes

B. Micro Averaged F Measure

In micro averaging, *F* measure is computed globally over all category decisions and obtained by summing over all individual decisions as follow [9].

Micro averaged *F* measure is then computed as:

$$F(\textit{micro}\ \text{average}) = \frac{2\pi\rho}{\pi + \rho} \tag{6}$$

where ρ *is* recall

π *is* precision

C. Recall

Recall is a measure of the ability of predication model to select instances of a certain class from the dataset and correspond to the true positive rate [9].

$$\text{Re}\,call = \frac{tp}{tp+fn} \quad (7)$$

D. Precision

Precision is a measure of the accuracy if a class has been predicted [9].

$$\text{Pr}\,ecision = \frac{tp}{tp+fp} \quad (8)$$

IV. Results and Discussion

Table 3 illustrates the highest classification rate occurs with dermatology dataset with accuracy 99.34%, precision 97.33%, recall 98.69%, Specificity 48.77. FMI 98.88 with CPU time 0.3061 seconds whereas classification accuracy rate with Pima Indian Diabetes classification accuracy rate occurs with 81.42%, precision 77.07%, recall 67%, Specificity 65.22. FMI 80.94 with 0.0194 seconds. Liver dataset shows accuracy 48.96%, precision 68.14%, recall48.50 %, Specificity 41.75 FMI 54.90 with CPU time 0.0045 seconds.

Table 3
Performance Analysis of FussCyier

UCI Datasets	Performance Measure					
	Accuracy	Precision	Recall	Speci-ficity	FMI	CPU Time
PID	81.42	77.07	67.00	65.22	80.9	0.01
ILPD	90.62	45.76	88.20	71.43	59.22	0.021
Liver dataset	48.96	68.14	48.50	41.75	54.9	0.009
Statlog (Heart)	84.52	85.33	80	55.56	77.78	0.0054
Dermato -logy	99.34	97.33	98.69	48.77	98.88	0.3061

PID: Pima Indians Diabetes
ILPD: Indian Liver Patient Dataset

Table 4 shows performance analysis of fuzzy soft based classifiers. It is experimentally demonstrated that this method yields high accuracy for five dataset namely Pima Indian diabetes, ILPD, liver, Statlog (Heart) and dermatology, when compared to existing fuzzy soft set classifiers such as FSSCT and FSSSM [3]. FussCyier does not perform well with Liver dataset where accuracy was observed as 48.96 which is lower than FSSCT 51.17 and FSSSM 53.01 respectively.

Table 4
Performance Analysis of different classifiers

UCI Datasets	FSSCT	FSSSM	FussCyier
Pima Indian Diabetes	70.35	70.22	81.42
ILPD	64.19	78.52	90.62
Liver dataset	51.17	53.01	48.96
Statlog(Heart)	82.72	77.04	84.52
Dermatology	82.97	97.03	99.34

Table 5 illustrates performance of different classification algorithms on real-world dataset Pima Indian diabetes. Naïve Bayes gives accuracy 77.86, recall .83, precision .83, C4.5 offered accuracy 78.22, recall .86, precision .81, whereas SVM accuracy 77.47, recall .77, precision .77, while FussCyier shows better classification accuracy 81.42, recall .67, precision .77 compared with base classifiers.

Table 5
Accuracy Comparison of algorithms for Pima Indian Diabetes dataset

Algorithms/Techniques	Accuracy	Recall	Precision
Naïve Bayes [15]	77.8646	.83	0.83
C4.5[15]	78.2252	0.864	0.814
SVM [15]	77.474	0.778	0.77
k-NN [15]	77.7344	0.892	0.792
FSSCT [3]	70.35	-	-
FSSSM [3]	70.22	-	-
FussCyier	81.42	.67	0.77

To sum up, data mining techniques is one of the accomplished and standard tools in detecting diseases at early stages and has become commonplace in the recent years due to the heightened performance in classification. The main purpose of these data mining techniques is to give the most efficient algorithms that designates given data in several characteristics. Moreover, these algorithms are paramount and required for automatic classification tools. FussCyier shows enhanced performance in terms of accuracy, precision and recall. Therefore, this classifier has the potential to be further enhanced and expanded by being incorporated with preprocessing techniques and other classification algorithms.

V. Conclusion

In this paper, the practicality of similarity fuzzy soft set based on distance measure has been investigated for classifying medical datasets. Five datasets from UCI machine learning and six performance parameters were used to determine the effectiveness of the proposed classification algorithm. Based on the results, it is experimentally proven that the proposed classifier demonstrates better performance compared to the existing techniques.

Acknowledgment

This work is supported by Research, Innovation, Commercialization, and Consultancy (ORICC), Vote No D001, Universiti Tun Hussein Onn Malaysia.

References

[1] W. Raghupathi, “Data mining in healthcare” Healthcare Informatics: Improving Efficiency through Technology, Analytics, and Management, 353-372, 2016.

[2] I. Yoo, P. Alafaireet, M. Marinov, K. Pena-Hernandez, R. Gopidi, J. F Chang, and L. Hua, “Data mining in healthcare and biomedicine: a survey of the literature,” Journal of medical systems, 36(4), 2431-2448, 2012.

[3] S. A. Lashari and R. Ibrahim, “Performance Comparison of Selected Classification Algorithms Based on Fuzzy Soft Set for Medical Data,” In Advanced Computer and Communication Engineering Technology(pp. 813-820), 2015. Springer International Publishing.

[4] S. A. Begum and O. M. Devi, “Fuzzy algorithms for pattern recognition in medical diagnosis,” Assam University Journal of Science and Technology, 7(2), 1-12, 2011.

[5] L. A. Zadeh, “Fuzzy sets,” Information and control, 8(3), 338-353, 1965.

[6] S. A. Lashari, R. Ibrahim, & N. Senan, Soft set theory for automatic classification of traditional Pakistani musical instruments sounds.

In *Computer & Information Science (ICCIS)*, 2012 *International Conference on* (Vol. 1, pp. 94-99). IEEE.

[7] N. Senan, R. Ibrahim, N. M. Nawi, & M. M. Mokji, (2009). Feature extraction for traditional malay musical instruments classification system. In *Soft Computing and Pattern Recognition, 2009. SOCPAR'09. International Conference of* (pp. 454-459). IEEE.

[8] M. M. Mushrif, S. Sengupta and A. K. Ray, "Texture classification using a novel, soft-set theory based classification algorithm," In Computer Vision–ACCV 2006 (pp. 246-254), 2006, Springer Berlin Heidelberg.

[9] B. Handaga, T. Herawan and M. M. Deris, "FSSC: An Algorithm for Classifying Numerical Data Using Fuzzy Soft Set Theory," International Journal of Fuzzy System Applications (IJFSA), 2(4), 29-46, 2012.

[10] S. A. Lashari, R. Ibrahim, N. Senan, I. T. R. Yanto and T Herawan, "Application of Wavelet De-noising Filters in Mammogram Images Classification Using Fuzzy Soft Set," In International Conference on Soft Computing and Data Mining (pp. 529-537), 2016, Springer, Cham.

[11] L. Baccour, A. M., Alim and R. I. John, "Some Notes on Fuzzy Similarity Measures and Application to Classification of Shapes, Recognition of Arabic Sentences and Mosaic." IAENG International Journal of Computer Science, 41(2), 81-90, 2014.

[12] P. Majumdar and S. K. Samanta, "Generalised fuzzy soft sets," Computers & Mathematics with Applications, 59(4), 1425-1432, 2010.

[13] N. Kalaiselvi and H. H. Inbarani, "Fuzzy Soft Set Based Classification for Gene Expression Data," arXiv preprint arXiv:1301.1502, 2013.

[14] A. Asuncion, & D. J. Newman, (2007). UCI Machine Learning Repository [http://www. ics. uci. edu/~ mlearn/MLRepository. html]. Irvine, CA: University of California. School of Information and Computer Science, 12

[15] P. C. Thirumal and N. Nagarajan, "Utilization of data mining techniques for diagnosis of diabetes mellitus-a case study," ARPN Journal of Engineering and Applied Science, 10(1), 2015.

[16] K .Polat,, S. Şahan , H. Kodaz & S. Güneş, "Breast cancer and liver disorders classification using artificial immune recognition system (AIRS) with performance evaluation by fuzzy resource allocation mechanism,". *Expert Systems with Applications*, *32*(1), 172-183, 2007.

[17] N. C. Long, P.Meesad & H. Unger, "A highly accurate firefly based algorithm for heart disease prediction," *Expert Systems with Applications*, *42*(21), 8221-8231, 2015.

[18] W. L. Zuo, Z. Y. Wang, T. Liu & H. L. Chen, " Effective detection of Parkinson's disease using an adaptive fuzzy k-nearest neighbor approach" *Biomedical Signal Processing and Control*, *8*(4), 364-373,2013.

[19] Ghofrani, F., Helfroush, M. S., Danyali, H., & Kazemi, K. (2014). Improving the performance of machine learning algorithms using fuzzy-based features for medical x-ray image classification. *Journal of Intelligent & Fuzzy Systems*, *27*(6), 3169-3180.

[20] P. C. Thirumal and N. Nagarajan, "Utilization of data mining techniques for diagnosis of diabetes mellitus-a case study," ARPN Journal of Engineering and Applied Science, 10(1), 2015.