Research Journal of Applied Sciences, Engineering and Technology 14(8): 291-298, 2017
DOI:10.19026/rjaset.14.4953
ISSN: 2040-7459; e-ISSN: 2040-7467

Submitted: December 22, 2016 Accepted: April 11, 2017 Published: August 15, 2017

Research Article

Fuzzy Discretization based Classification of Medical Data

[1]M. Shanmugapriya, [1]H. Khanna Nehemiah, [1]R.S. Bhuvaneswaran, [2] Kannan Arputharaj and [1]J. Dhalia Sweetlin
[1]Ramanujan Computing Centre,
[2]Department of Information Science and Technology, Anna University, Chennai-600025, India

Abstract: Discretization is one of the commonly used data preprocessing technique to improve the efficiency of the knowledge extraction process on clinical data. Generally, clinical data contains numeric attributes with continuous values. Data discretization simplifies the original data by transforming continuous data attribute values into a finite set of intervals. Although discretization is capable of handling continuous attributes on clinical data, there are cases where discretization is not an appropriate technique for handling continuous attributes. There are instances where attribute values are vague, imprecise and have multiple distributions with different classes, which challenges the process of mining in clinical data. Hence, there is a need for fuzzy discretization to pre-process the clinical data before mining. The aim of this study is to derive fuzzy discretization from crisp-interval discretization using geometric approach for constructing fuzzy sets, where overlapping region between the fuzzy sets is represented as geometric area. This study comprises of three steps: First, non-overlapping fuzzy sets are constructed using intervals generated from crisp-interval discretization. Second, area of overlapping between the fuzzy sets is computed based on the geometric approach and an average area of overlapping is estimated. Third, fuzzy sets are redesigned based on the estimated average area of overlapping. Fuzzy discretizations for three, five and seven intervals have been examined using Pima Indian Diabetes dataset (PID) and Bupa Liver Disorder dataset (BLD) taken from the University of California Irvine machine learning repository. The variation in performance of crisp and fuzzy discretization methods is measured using six classification approaches namely, tree based approach, probabilistic induction based approach, rule-based approach, network learning approach, kernel-based approach and distance-based approach and a rule-based fuzzy inference system. The results show that the classification accuracy remains stable with less deviation across different classifiers with varying intervals.

Keywords: Classification, fuzzy discretization, fuzzy set, interval discretization, membership function, overlapping area

INTRODUCTION

Electronic Medical Records (EMR) stores enormous clinical data that describes the health care examinations undergone by the patients. These data contain hidden knowledge that can be used in developing Clinical Decision Support Systems (CDSSs). The CDSS assists the clinician in the decision making activities such as diagnosis, prognosis and treatment. The process of Knowledge Discovery in Databases (KDD) plays a vital role in extracting the knowledge from clinical data. Data pre-processing and data mining is an important step in KDD. Data pre-processing is the task of improving the quality of data for mining process. The task includes data cleaning, data integration, data transformation and data reduction. Data reduction methods comprises of numerosity reduction, dimensionality reduction and data discretization. Data discretization plays an important role in obtaining cognitively relevant human interpretation and in speeding up the computation process with a reduced level of data (Mittal and Cheong, 2002; Russell and Norvig, 1995).

Data discretization converts the continuous valued attributes into a range of discrete intervals. This conversion can also affect the performance of predictors and classifiers (Kianmehr *et al*., 2008; Ishibuchi *et al*., 2001). There has been several works in the literatures that emphasize the importance of performing data discretization before mining process (Maslove *et al*., 2013; Mittal and Cheong, 2002). Based on the learning approach, data discretization methods have been classified into two types namely, Supervised and Unsupervised (Dougherty *et al*., 1995). Supervised discretization is possible only when the class information is present in the database. If no class

Corresponding Author: M. Shanmugapriya, Ramanujan Computing Centre, Anna University, Chennai-600025, India

information is available, unsupervised discretization is preferred. Shanmugapriya *et al*. (2016b) have used unsupervised interval discretization methods in their study and have applied to four medical data sets namely Cleveland Heart Disease (CHD) data set, Chronic Kidney Disease (CKD) data set, Pima Indians Diabetes (PID) data set and BUPA Liver Disorder (BLD) data set. The performance of the various classifiers such as C4.5 decision tree, Support Vector Machine (SVM), K-Nearest Neighbor (K-NN), Classification Based on Association rules (CBA), Bayes and Multilayer Perceptron (MLP)are analyzed by varying intervals. Normally, interval discretization has been used for handling continuous attributes in many machine learning techniques such as decision trees, Bayesian networks and association rule mining (Quinlan, 1996).

Even though interval discretization is enough to handle the continuous attributes, it may not be appropriate in situations, where there is no clear boundary to set the interval limits. Since fuzzy logic is a convenient and well known tool for handling continuous attributes with unclear boundary (Zeinalkhani and Eftekhari, 2014; Zimmermann, 1996), in this study, fuzzy based discretization has been used to discretize the data with continuous attributes. Moreover, in Clinical Decision Support System (CDSS), there is a need for human reasoning in order to handle continuous attributes (Pal *et al*., 2012). Fuzzy logic can better represent the continuous attributes in human understandable manner. To handle vague and imprecise continuous attributes in the data set, fuzzy discretization is performed on the dataset (Mehta *et al*., 2009). There are several works on deriving fuzzy discretization from interval discretization (Roy and Pal, 2003; Zeinalkhani and Eftekhari, 2014). Ishibuchi and Yamamoto (2003) have examined two methods of generating fuzzy discretization from interval based discretization. The first method was based on trapezoidal membership function, which is a linear function. The second method extended the trapezoidal membership function to piecewise linear function. In both methods, the overlap grades were assigned based on the parameters of adjacent membership functions. For experimentation, they used three interval discretization methods namely equal width intervals, equal-frequency intervals and minimum entropy intervals. Their proposed approach was tested using wine data set with 13 continuous valued attributes and sonar dataset with 60 continuous valued attributes by varying overlapping grades. The datasets were collected from the University of California Irvine (UCI) Machine Learning repository. The discretized data sets were classified using fuzzy rule-based classifier. From the results, it was observed that classification ability was increased for some cases and degraded for other cases by the increase in the overlap grades.

Zeinalkhani and Eftekhari (2014) proposed a two-step algorithm to generate the membership functions. In the first step, discretization algorithm divides the domain of attributes to several partitions and in the second step, a membership function is defined on each partition. The generated partitions were transformed into fuzzy membership functions. Transformations were performed based on four approaches: First approach was based on partition width, second one was based on standard deviation of examples inside the partition, third one was based on Neighbor Partition Coverage Rate (NPCR) and the last one was based on Partition Coverage Rate (PCR). Furthermore, they proposed a membership function generation algorithm, called Fuzzy Entropy Based Fuzzy Partitioning (FEBFP). For experimentation, they considered several discretization methods including equal width and equal frequency. Datasets were taken from UCI machine learning repository and also from Knowledge Extraction based on Evolutionary Learning (KEEL) dataset repository (Alcalá-Fdez *et al*., 2011). Their experimental result shows that membership functions defined by partition coverage rate and partitions generated by Zeta discretization algorithm performed well.

Ishibuchi *et al*. (2001) compared fuzzy discretization with standard non-fuzzy discretization using fuzzy rule-based system. For the experimentation they used wine data set taken from UCI machine learning repository. Wine data set is a three-class pattern classification problem with 178 patterns and 13 continuous attributes. In fuzzy discretization, they discretized each attribute of wine data set into fuzzy sets with linguistic terms, where each fuzzy set is characterized using triangular membership function. They have designed the fuzzy sets based on the following two conditions: 1) The sum of neighboring membership functions is always 1 and 2) Crossing points of neighboring membership functions coincide with the threshold values in the interval discretization. They generated the linguistic rules using linguistic terms. For the generation of linguistic rules, they have extended the definition of basic rule selection criteria such as support and confidence. For non-fuzzy discretization, they used entropy based discretization method. They compared fuzzy and non-fuzzy discretization approaches using fuzzy rule-based system on wine dataset. From the result, it was observed that higher classification accuracy (95%) was obtained using fuzzy discretization. For non-fuzzy discretization, they observed only 93% accuracy.

Fazzolari *et al*. (2014) presented a multi-objective evolutionary method to improve accuracy-interpretability trade-off of fuzzy rule-based classification systems. This method works in three stages namely fuzzy discretization, rule base extraction and concurrent tuning of both membership functions in database and the selection rules in the rule base. In the first stage, fuzzy discretization algorithm has been designed to generate suitable granularities for defining the initial fuzzy partitions of the database. In the second

stage, rule base associated to the fuzzy partitions (obtained in the first step) was created by extracting candidate fuzzy association rules. In the final stage, multi-objective evolutionary algorithm was designed to perform the tuning of membership functions concurrently in the database and the selection of rules in the rule base. The proposed method was tested with 35 datasets taken from KEEL dataset repository, including small size datasets and high dimensional and large scale datasets. The obtained results show that the multi-objective evolutionary approach improves the precision, with respect to the results obtained using a single-objective approach. The knowledge of the domain expert is utilized to design fuzzy sets in most of the existing works. To overcome this dependency on an expert, in this study a geometric approach is used for designing the fuzzy set. Geometric representation is preferred as human reasoning can be better represented.

This study presents a geometric approach using SimE, for deriving fuzzy discretization from Equal Width (EW) interval discretization method (Shanmugapriya *et al*., 2016a). The proposed approach was tested with two medical datasets namely Pima Indians Diabetes dataset and Bupa Liver Disorder Dataset with three intervals. Fuzzy discretization is derived from interval discretization (EW) in three steps: In the first step, data sets are discretized into intervals using equal width discretization method. In second step, fuzzy sets are created using the boundary values of each interval derived from the equal width discretization method. The adjacent fuzzy sets will have no similarity (overlapping area) between them because it is derived from crisp intervals. Setnes *et al*. (1998) has suggested that fuzzy sets with optimal overlapping area can improve the semantic representation and performance of any fuzzy based system. Since, an area of overlapping between the adjacent fuzzy sets is preferred, it is estimated by investigating many studies (Allahverdi, 2009; Muthukaruppan and Er, 2012; Samuel *et al*., 2013) on fuzzy classification of medical data.

In third step, fuzzy sets created in step two are redesigned with the estimated area of overlapping using SimE. The proposed approach is evaluated through fuzzy rule-based classification for the considered intervals on Pima Indian Diabetes dataset and Bupa Liver Disorder dataset.

MATERIALS AND METHODS

Two Clinical datasets taken from the University of California Irvine machine learning repository have been used for this experimentation: Pima Indian Diabetes (PID) dataset with 9 attributes and 768 instances, Bupa Liver disorder (BLD) dataset with 7 attributes and 345 instances. Each dataset contains discrete, categorical and continuous attributes. Table 1 shows the description, type and range of the attributes in PID dataset. It has six continuous attributes, one discrete attribute and one categorical attribute for representing class label (presence and absence of the diabetes). Table 2 depicts the description, type and range of the attributes in BLD dataset. It includes five continuous attributes and one categorical attribute for representing the class label. (presence and absence of the liver disorder).

Fuzzy set: Fuzzy set is a set whose element has a degree of membership. A fuzzy set A in X is a set of ordered pairs (Zadeh, 1965):

$$A = \left\{ \left(x, \mu_A(x) \right) \middle| \, x \epsilon X \right\} \quad (1)$$

where, X is called universe of discourse and $\mu_A(x)$: $X \rightarrow [0,1]$ is the membership function which maps each element x of X to a value between 0 to 1. Generally, membership functions are identified and designed by the domain expert (Bera *et al*., 2014).

Table 1: Description of Pima Indian diabetes dataset

Attribute name	Description	Type	Range
Preg	Number of times pregnant	Discrete	0-17
Glucose	Plasma glucose concentration a 2 hours in an oral glucose tolerance test	Continuous	0-199
Bp	Diastolic blood pressure	Continuous	0-122
Skin	Triceps skin fold thickness	Continuous	0-99
Insulin	2-Hour serum insulin	Continuous	0-846
BMI	Body mass index	Continuous	0-67.1
Pedi	Diabetes pedigree function	Continuous	0-2.42
Age	Age of the person	Discrete	21-81
Class	Diabetes/Non-Diabetes	Categorical	0-1

Table 2: Description of Bupa liver disorder dataset

Attribute name	Description	Type	Range
Mcv	Mean corpuscular volume	Continuous	65-103
Alkphos	Alkaline phosphotase	Continuous	23-138
Sgpt	Alamine aminotransferase	Continuous	4-155
Sgot	Aspartate aminotransferase	Continuous	5-82
Gammagt	Gamma-glutamyltranspeptidase	Continuous	5-297
Drinks	Number of half-pint equivalents of alcoholic beverages drunk per day	Continuous	0-20
Class	Liver disorder presence/absence	Categorical	0-1

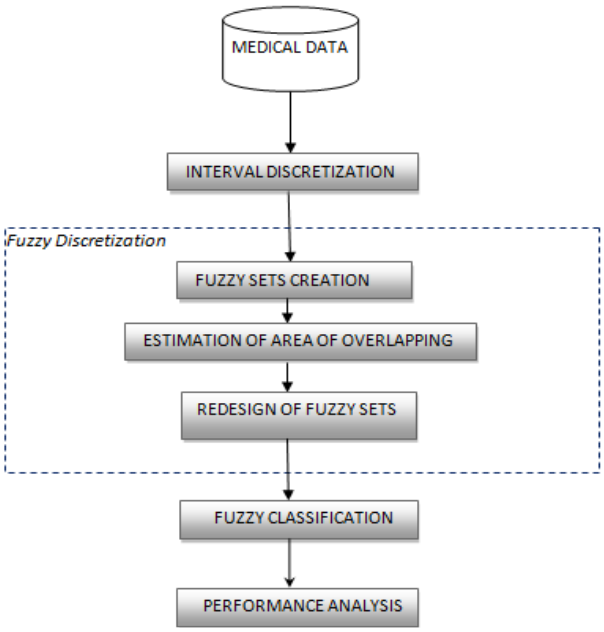

Fig. 1: System framework

Membership Function (MF): Membership function is used to characterize the fuzziness in the fuzzy set. It defines the degree of an element's membership in a fuzzy set. The value of the membership is in the range [0, 1]. There are different types of MFs namely, triangular, Gaussian, trapezoidal, bell shaped and sigmoidal.

In this study, triangular membership function is used for characterizing the fuzzy sets. The triangular membership function for a fuzzy set X is defined using Eq. (2) (Kaufmann, 1975; Klir and Yuan, 1991):

$$Triangular(X; a, b, c) = \begin{cases} 0, x \le a. \\ \frac{(x-a)}{(b-a)}, a \le x \le b. \\ \frac{(c-x)}{(c-b)}, b \le x \le c. \\ 0, c \le x. \end{cases} \quad (2)$$

where, x is an element of a fuzzy set X, a and c represents the lower and the upper boundary of fuzzy set X and b represents the center of the fuzzy set X.

Fuzzy set similarity: Similarity is a measure of approximate equality between the fuzzy sets. The similarity measure of fuzzy sets has been applied in many fields such as classification, clustering, image processing, fuzzy reasoning and decision making (Setnes *et al*., 1998; Zwick *et al*., 1987). Different kinds of similarity measures are proposed in literatures (Pappis and Karacapilidis, 1993). In most of the existing works researchers have estimated the similarity based on elements of the sets. Shanmugapriya *et al*. (2016a) in their previous work they have proposed an algorithm called Similarity Estimator (SimE), for estimating the similarity between fuzzy sets using a geometric approach.

SYSTEM FRAMEWORK

The proposed method has four phases namely, the Interval Discretization, Fuzzy Discretization, Fuzzy rule-based Classification and Performance Analysis. The framework of the proposed method is given in Fig. 1. The details of each phase are explained in the following sub sections:

Interval discretization: In the first phase, continuous data of PID and BLD dataset are discretized into k equal sized intervals (I_1, I_2..., I_k) using EW discretization method. Each interval I_k is denoted by its lower limit l_k and upper limit u_k as $I_k = [l_k, u_k]$. The width of an interval(w) can be computed using the Eq. (3), (4) and (5) respectively (Liu *et al*., 2002):

$$w = \frac{V_{max} - V_{min}}{k} \quad (3)$$

$$V_{max} = \max\{v_1, v_2, v_3, \ldots v_n\} \quad (4)$$

$$V_{min} = \min\{v_1, v_2, v_3, \ldots v_n\} \quad (5)$$

where, V_{max} and V_{min} are the maximum and minimum values of an attribute $V, v_i \in V, i = \{1,2,3 \ldots n\}$; n is the number values in each attribute; k is the number of cut points specified by the user. In this study, three values of k have been examined: k = {3, 5, 7}. The $k + 1$ cut points are $V_{min} + w, V_{min} + 2w, \ldots, V_{min} + (k-1)w$. Non overlapping intervals are obtained in this phase. There is no overlapping between the intervals boundaries.

Fuzzy discretization: In this phase, fuzzy discretization is obtained from the interval discretization in three steps:

Step 1: In this step, fuzzy sets (A_1 A_2..., A_k) are constructed for each attribute of the PID and BLD datasets. Triangular membership function (Kaufmann, 1975) is used for constructing the fuzzy set for each of the interval generated from interval discretization method. Parameters (a, c, b) needed for constructing fuzzy sets using triangular membership function are obtained from the lower and upper bound of the intervals. The parameters a and c represent the lower and upper boundary of the fuzzy set respectively. The parameter b represents the center of the fuzzy set. This step results the non-overlapping fuzzy sets of all the attributes derived from the crisp intervals. Figure 2 shows the three non-overlapping fuzzy sets (Small, Normal, Big) of Mean Corpuscular Volume (Mcv) attribute of BLD data set derived using EW discretization method.

The Mcv attribute value ranges between 0-200. The Fuzzy set 'Small' is defined with the values [0, 78, 39], the fuzzy set 'Normal' is defined with the values [78, 93, 86] and the fuzzy set 'Big' is defined with the values [93,103, 98].

Step 2: Fuzzy sets generated in step 1 have no overlapping area because it is derived from crisp-intervals. In this step, to design the fuzzy sets with overlapping area, an average area of overlapping between the fuzzy sets is estimated. This estimation is arrived after investigating many studies (Allahverdi, 2009; Muthukaruppan and Er, 2012; Samuel *et al*., 2013) on fuzzy classification of medical data.

Step 3: In this step, the fuzzy sets obtained in step1 are redesigned with the estimated area of overlapping using SimE algorithm. SimE computes the area of overlapping by partitioning the region of overlapping into geometric structures and summing the area of resulting polygons. To obtain the area of overlapping with an estimated value, fuzzy sets are redesigned by adjusting the parameters of triangular membership function (a, b and c). This step results overlapping fuzzy sets.

Figure 3 shows the overlapping fuzzy sets of mean corpuscular volume (mcv) attribute of BLD data set obtained after redesigning. The fuzzy set 'Small' is defined with the values [0, 82, 39], the fuzzy set

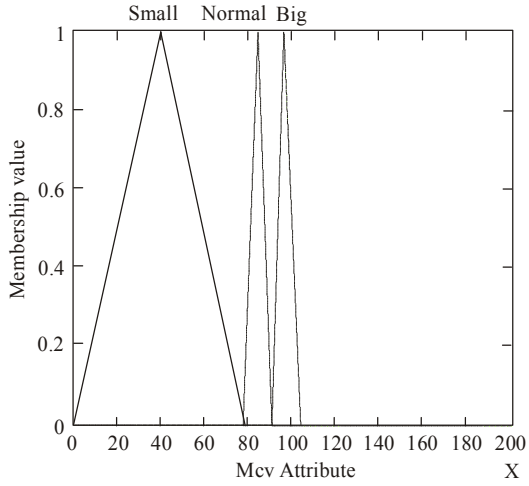

Fig. 2: Non-overlapping fuzzy sets of mean corpuscular volume (Mcv)

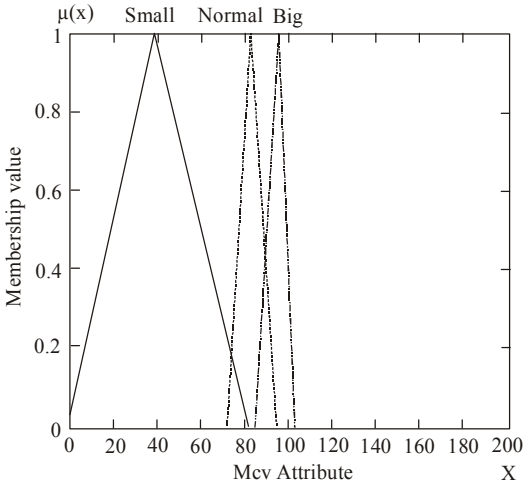

Fig. 3: Overlapping fuzzy sets of mean corpuscular volume (Mcv)

'Normal' is defined with the values [72, 95, 84] and the fuzzy set 'Big' is defined with the values [85, 103, 97].

Fuzzy rule-based classification: In this phase, Mamdani-type Fuzzy Inference System (FIS) is used as fuzzy rule-based classification system for classifying the presence or absence of a disease. Fuzzy inference system is modeled in the following four steps:

Step1: Fuzzification: This process involves the transformation of all the input attributes into the corresponding fuzzy sets with linguistic terms using the function defined in Eq. (2). Inputs of the fuzzy inference system are the generated fuzzy sets, values of the attributes and the rule set. In this study, the proposed geometric approach for fuzzy discretization is used to generate the fuzzy sets (Fuzzified output).

Step 2: Fuzzy rule set generation: Rule set is created by including all possible combinations of attributes and

classes. The rule set is characterized by a set of IF-THEN rules in which the antecedents and the consequents involve linguistic terms. In this study, fuzzy rules are generated by defining the crisp rules with linguistic terms. Discretized data obtained from the interval discretization are given to the Partial Decision Tree (PART) algorithm for generating crisp rules (Exarchos *et al*., 2012).

Step 3: Fuzzy inference: This process maps a given fuzzy input to a fuzzy output using the rules contained in the rule set. This mapping provides a basis from which decisions can be made. The inference process receives its inputs from fuzzification process and the rule set. This is obtained by performing the following steps as discussed by Rajasekaran and Vijayalakshmi Pai (2007):

Step 3.1: Apply fuzzy AND operator on the antecedent part of the rule.
Step 3.2: Analyze the implication from antecedent to consequent, using the rules in the rule set.
Step 3.3: Aggregate the consequents across the rules into single output.

Step 4: Defuzzification: Fuzzy inference process returns the inference value of an instance. In this step, the inference value is mapped into crisp output using Mean of Maximum (MoM) defuzzification method (Naaz *et al*., 2011). Accuracy of fuzzy-rule based classification is computed based on the defuzzified value. The above steps are performed for each interval discretization (3, 5 and 7).

Step 5: The steps one through four are repeated for each interval discretization (3, 5 and 7).

Performance analysis: The performance of the EW interval discretization method with fuzzy discretization method is analyzed and compared using six traditional classifiers namely Associative classifier (CBA), Decision tree classifier (C4.5), Support Vector Machine (SVM), Multi-layer Perceptron classifier (MLP), Naïve Bayes classifier (NB) and k-Nearest Neighbour classifier (kNN) and a rule-based Fuzzy Inference System (FIS) by varying the discretization intervals namely three, five and seven. Performance evaluation parameters namely, Classification Accuracy (CA), Sensitivity (SN) and Specificity (SP) are computed using Eq.(6), Eq. (7) and Eq. (8) respectively:

$$CA = \frac{TP+TN}{TP+TN+FP+FN} \quad (6)$$

$$SN = \frac{TP}{TP+FN} \quad (7)$$

$$SP = \frac{TN}{TN+FP} \quad (8)$$

where, TP, TN, FP and FN represent the true positives, true negatives, false positives and false negatives respectively. Sensitivity measures the proportion of positives that are correctly identified as positive. Specificity measures the proportion of negatives that are correctly identified as negative. Accuracy measures the proportion of the total number of predictions that are correct.

RESULTS AND DISCUSSION

This study is implemented using MATLAB R2013a. The proposed approach is tested with two datasets namely Pima Indian Diabetes dataset and Bupa Liver Disorder dataset. All the continuous attributes in the data set are fuzzified using the proposed fuzzy discretization approach. The performance of the fuzzy discretization approach is evaluated using fuzzy rule-based classification system. Fuzzy toolbox available in MATLAB R2013a is used for building fuzzy inference system. The data is split into training (75% of the data) and testing data (25% of the data).

Fuzzy inference system is modeled using training data and it is tested using test data. For each dataset (PID and BLD) and for each discretization interval, performance of fuzzy rule-based classification system is analyzed and compared with six crisp-interval discretization based classifiers. Table 3 depicts the results obtained in the experimentation. For Pima Indian diabetes dataset, the six traditional classifiers achieved an average accuracy of 71.63%, 70.618% and 70.40% for three, five and seven intervals respectively and it is depicted in Fig. 4. For the same dataset, Fuzzy discretization based classifier obtained an accuracy of 55.46%, 64.58% and 52.99% for three, five and seven intervals respectively.

Fuzzy discretization based fuzzy classification obtained the highest accuracy of 64.58% at seven interval. In Bupa Liver Disorder dataset, traditional classifiers achieved an average accuracy of 57.66%, 55.94% and 55.99% for three, five and seven intervals respectively. For the same dataset, Fuzzy discretization based classifier obtained an accuracy of 53.04%, 51.01% and 48.11% for three, five and seven intervals respectively and it is depicted in Fig. 5. There is a drop in the performance values as an expert is not involved in designing the fuzzy sets. Fuzzy discretization based classifier yielded the highest accuracy of 53.04% at interval three.

CONCLUSION

Generally medical data has huge number of continuous attributes with uncertainty, vague and impreciseness. Proper handling of such attributes improves the performance of the decision making system in medical domain. Representation of medical data in human understandable form requires discretization. Although interval discretization is capable of handling continuous attributes, it is not able

Table 3: Classification performance evaluation for fuzzy and crisp-interval discretization

		SVM			KNN			C4.5		
Dataset	No. of Intervals	*Acc	Sen	Spec	Acc	Sen	Spec	Acc	Sen	Spec
PID	3	73.823	0.589	0.818	72.650	0.604	0.792	73.043	0.563	0.820
	5	68.628	0.246	0.922	71.476	0.537	0.810	73.963	0.529	0.852
	7	69.137	0.320	0.890	68.749	0.485	0.796	74.475	0.607	0.818
BLD	3	61.739	0.615	0.619	57.391	0.603	0.545	57.971	0.585	0.573
	5	59.710	0.775	0.426	52.173	0.686	0.363	54.782	0.698	0.403
	7	60.289	0.757	0.454	54.492	0.680	0.414	58.260	0.863	0.312
		CBA			Bayes			MLP		
Dataset	No. of Intervals	Acc	Sen	Spec	Acc	Sen	Spec	Acc	Sen	Spec
PID	3	65.106	0.000	1.000	73.561	0.562	0.828	71.625	0.534	0.814
	5	63.930	0.422	0.756	75.267	0.604	0.832	70.444	0.515	0.806
	7	65.106	0.000	1.000	76.307	0.653	0.822	68.628	0.503	0.784
BLD	3	52.662	0.964	0.099	58.840	0.544	0.630	57.391	0.558	0.585
	5	53.930	0.877	0.209	57.971	0.733	0.431	57.101	0.591	0.551
	7	55.064	0.835	0.286	58.550	0.739	0.437	49.275	0.532	0.454
		Fuzzy								
Dataset	No. of Intervals	Acc	Sen	Spec						
PID	3	55.4688	0.1493	0.7720						
	5	64.58	0.0187	0.9820						
	7	52.9948	0.0896	0.7660						
BLD	3	53.0435	0.4943	0.5286						
	5	51.0145	0.4545	0.5680						
	7	48.1159	0.5739	0.3846						

*Acc- Accuracy; Sen- Sensitivity; Spec- Specificity

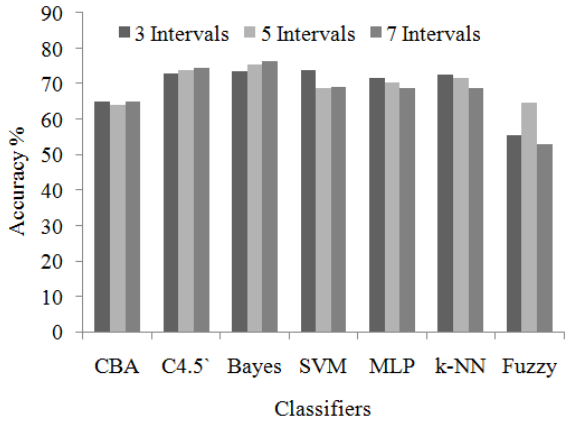

Fig. 4: Performance of classifiers for Pima Indian diabetes dataset

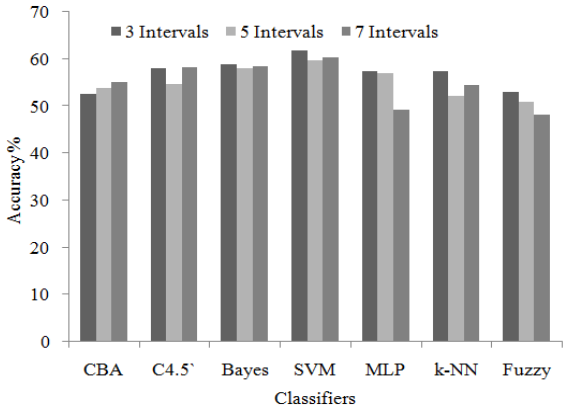

Fig. 5: Performance of classifiers for Bupa liver disorder dataset

to handle data with vagueness and uncertainty where there is multiple overlapping data distribution. In order to handle such data, this study proposes a method for fuzzy discretization where each attribute is discretized into set of overlapping fuzzy sets. The proposed fuzzy discretization method is examined using fuzzy rule based classification system. Then it is compared with six traditional classification approaches. The results obtained from this study show that the classification accuracy remains stable with less deviation across different classification approaches. However, the proposed fuzzy classifier provides better accuracy than the existing traditional classifiers at least in one interval. Further work in this direction can be the use of fuzzy logic in other classifiers to provide a hybrid classifier that can improve the accuracy further.

REFERENCES

Alcalá-Fdez, J., A. Fernández, J. Luengo, J. Derrac, S. García, L. Sánchez and F. Herrera, 2011. KEEL data-mining software tool: Data set repository, integration of algorithms and experimental analysis framework. J. Mult-Valued Log. S., 17: 255-287.

Allahverdi, N., 2009. Some applications of fuzzy logic in medical area. Proceeding of the IEEE International Conference on Application of Information and Communication Technologies (AICT), pp: 1-5.

Bera, S., A.J. Gaikwad and D. Datta, 2014. Selection of fuzzy membership function based on probabilistic confidence. Proceeding of the International Conference on Control, Instrumentation, Energy and Communication (CIEC), pp: 612-615.

Dougherty, J., R. Kohavi and M. Sahami, 1995. Supervised and unsupervised discretization of continuous features. Proceeding of the 12th International Conference on Machine Learning, 12: 194-202.

Exarchos, T.P., A.T. Tzallas, D. Baga, D. Chaloglou, D.I. Fotiadis, S. Tsouli, M. Diakou and S. Konitsiotis, 2012. Using Partial decision trees to predict Parkinson's symptoms: A new approach for diagnosis and therapy in patients suffering from Parkinson's disease. Comput. Biol. Med., 42(2): 195-204.

Fazzolari, M., R. Alcalá and F. Herrera, 2014. A multi-objective evolutionary method for learning granularities based on fuzzy discretization to improve the accuracy-complexity trade-off of fuzzy rule-based classification systems: D-MOFARC algorithm. Appl. Soft Comput., 24: 470-481.

Ishibuchi, H. and T. Yamamoto, 2003. Deriving fuzzy discretization from interval discretization. Proceeding of the 12th IEEE International Conference on Fuzzy Systems, 1: 749-754.

Ishibuchi, H., T. Yamamoto and T. Nakashima, 2001. Fuzzy data mining: Effect of fuzzy discretization. Proceeding of the IEEE International Conference on Data Mining (ICDM), pp: 241-248.

Kaufmann, A., 1975. Introduction to the Theory of Fuzzy Subsets, V.1: Fundamental Theoretical Elements. Academic Press, San Diego.

Kianmehr, K., M. Alshalalfa and R. Alhajj, 2008. Effectiveness of fuzzy discretization for class association rule-based classification. In: An, A., S. Matwin, Z.W. Raś and D. Ślęzak (Eds.), Foundations of Intelligent Systems. ISMIS, 2008. Lecture Notes in Computer Science, Springer, Berlin, Heidelberg, 4994: 298-308.

Klir, G.J. and B. Yuan, 1991. Fuzzy Sets and Fuzzy Logic. Prentice-Hall, Englewood Cliffs, NJ.

Liu, H., F. Hussain, C.L. Tan and M. Dash, 2002. Discretization: An enabling technique. Data Min. Knowl. Disc., 6(4): 393-423.

Maslove, D.M., T. Podchiyska and H.J. Lowe, 2013. Discretization of continuous features in clinical datasets. J. Am. Med. Inform. Assn., 20(3): 544-553.

Mehta, R.G., D.P. Rana and M.A. Zaveri, 2009. A novel fuzzy based classification for data mining using fuzzy discretization. Proceeding of the WRI World Congress on Computer Science and Information Engineering, 3: 713-717.

Mittal, A. and L.F. Cheong, 2002. Employing discrete bayes error rate for discretization and feature selection tasks. Proceeding of the IEEE International Conference on Data Mining (ICDM-2002), pp: 298-305.

Muthukaruppan, S. and M.J. Er, 2012. A hybrid particle swarm optimization based fuzzy expert system for the diagnosis of coronary artery disease. Expert Syst. Appl., 39(14): 11657-11665.

Naaz, S., A. Alam and R. Biswas, 2011. Effect of different defuzzification methods in a fuzzy based load balancing application. Int. J. Comput. Sci., 8(5).

Pal, D., K.M. Mandana, S. Pal, D. Sarkar and C. Chakraborty, 2012. Fuzzy expert system approach for coronary artery disease screening using clinical parameters. Knowl-Based Syst., 36: 162-174.

Pappis, C.P. and N.I. Karacapilidis, 1993. A comparative assessment of measures of similarity of fuzzy values. Fuzzy Set. Syst., 56(2): 171-174.

Quinlan, J.R., 1996. Improved use of continuous attributes in C4.5. J. Artif. Intell. Res., 4: 77-90.

Rajasekaran, S. and G.A. Vijayalakshmi Pai, 2007. Neural Networks, Fuzzy Logic and Genetic Algorithms: Synthesis and Applications. Prentice Hall, New Delhi, India.

Roy, A. and S.K. Pal, 2003. Fuzzy discretization of feature space for a rough set classifier. Pattern Recogn. Lett., 24(6): 895-902.

Russell, S.J. and P. Norvig, 1995. Artificial Intelligence: A Modern Approach. Prentice-Hall, Englewood Cliffs, NJ.

Samuel, O.W., M.O. Omisore and B.A. Ojokoh, 2013. A web based decision support system driven by fuzzy logic for the diagnosis of typhoid fever. Expert Syst. Appl., 40(10): 4164-4171.

Setnes, M., R. Babuška, U. Kaymak and H.R. van Nauta Lemke, 1998. Similarity measures in fuzzy rule base simplification. IEEE T. Syst. Man Cy. B, 28(3): 376-386.

Shanmugapriya, M., H. Khanna Nehemiah, R.S. Bhuvaneswaran, K. Arputharaj and J. Jabez Christopher, 2016a. SimE: A geometric approach for similarity estimation of fuzzy sets. Res. J. Appl. Sci. Eng. Technol., 13(5): 345-353.

Shanmugapriya, M., H. Khanna Nehemiah, R.S. Bhuvaneswaran, K. Arputharaj and J. Dhalia Sweetlin, 2016b. Unsupervised discretization: An analysis of unsupervised discretization approaches for clinical datasets. Res. J. Appl. Sci. Eng. Technol., (Accepted for Publication).

Zadeh, L.A., 1965. Fuzzy sets. Inform. Control, 8(3): 338-353.

Zeinalkhani, M. and M. Eftekhari, 2014. Fuzzy partitioning of continuous attributes through discretization methods to construct fuzzy decision tree classifiers. Inform. Sciences, 278: 715-735.

Zimmermann, H.J., 1996. Fuzzy Set Theory-and Its Applications. 3rd Edn., Kluwer Academic Publishers, Norwell, MA, USA.

Zwick, R., E. Carlstein and D.V. Budescu, 1987. Measures of similarity among fuzzy concepts: A comparative analysis. Int. J. Approx. Reason., 1(2): 221-242.