Applied Soft Computing 49 (2016) 137–145

Contents lists available at ScienceDirect

Applied Soft Computing

journal homepage: www.elsevier.com/locate/asoc

A computational intelligence technique for the effective diagnosis of diabetic patients using principal component analysis (PCA) and modified fuzzy SLIQ decision tree approach

Kamadi V.S.R.P. Varma [a,*], Allam Appa Rao [b], Thummala Sita Mahalakshmi [a], P. V. Nageswara Rao [a]

[a] Department of Computer Science and Engineering, GITAM University, Visakhapatnam, India
[b] CRRao AIMSCS, UoH Campus, Hyderabad, India

ARTICLE INFO

Article history:
Received 29 September 2015
Received in revised form 17 June 2016
Accepted 27 June 2016

Keywords:
Computational intelligence technique
Fuzzy decision tree
Fuzzification
Knowledge inference systems
Gini index
SLIQ
Data reduction

ABSTRACT

Knowledge inference systems are built to identify hidden and logical patterns in huge data. Decision trees play a vital role in knowledge discovery but crisp decision tree algorithms have a problem with sharp decision boundaries which may not be implicated to all knowledge inference systems. A fuzzy decision tree algorithm overcomes this drawback. Fuzzy decision trees are implemented through fuzzification of the decision boundaries without disturbing the attribute values. Data reduction also plays a crucial role in many classification problems. In this research article, it presents an approach using principal component analysis and modified Gini index based fuzzy SLIQ decision tree algorithm. The PCA is used for dimensionality reduction, and modified Gini index fuzzy SLIQ decision tree algorithm to construct decision rules. Finally, through PID data set, the method is validated in the simulation experiment in MATLAB.

1. Introduction

Diabetes is this century's stunning health hazard which has most significance throughout the world. Diabetes has severe health complications like diabetic cardiomyopathy, myopathy, neuropathy, retinopathy, nephropathy and many other complications which affect the financial and social life of people [1]. Type1 and Type2 are the well known and the most occurring types of diabetes. Type1/Juvenile diabetes occurs due to the body immune system. Diabetes mellitus is the most common type of diabetes which occurs due to ineffective use of insulin by the body cells or inefficient insulin secretion from the pancreas. The people suffering from this disease may lead a healthy and happy life if they can manage the disease properly through proper medication and healthy diet under the supervision of a diabetician. So, early detection plays a vital role in the diagnosis of diabetes disease and it can further prevail over the diabetes health complications.

The early diagnosis of the disease may reduce other health complications. Doctors, medical practitioners or dialectologists are now depending on computers for effective and efficient disease diagnosis. The computer diagnosis system uses knowledge discovery methods, which are known as Computational Intelligence techniques (CI). The CI techniques may be strengthened through effective and sophisticated knowledge discovery mechanisms.

In this research article the authors proposed a hybrid model through combining the PCA with modified fuzzy SLIQ gini index based decision tree algorithm.

The rest of the paper is organized as follows. Section 2 describes the literature review. Data description is presented in Section 3. Proposed model description and flow chart is presented in Section 4. Section 5 deals with the performance metrics and k-fold cross validation approaches used in the present method. The overall result discussion is presented in Section 6. Finally the paper is concluded in the Section 7.

2. Littttterature review

Classification is one of the data mining techniques which discover valid hidden knowledge from data. The mostly used decision tree models are ID3 and C4.5 [2,3] which use entropy as split mea-

* Corresponding author.
E-mail addresses: gitamvarma@gmail.com, varma.6680@gmail.com (V.S.R.P.V. Kamadi).

http://dx.doi.org/10.1016/j.asoc.2016.05.010
1568-4946/

Table 1
Description of PID Data Set.

Attribute	Variable	Abbreviation	Specification
A1	V1	Pregnant	Number of times pregnant
A2	V2	Glucose	Plasma glucose concentration at 2 h. in an oral glucose tolerance test-(mg/dl)
A3	V3	DBP	Diastolic blood pressure-(mm Hg)
A4	V4	TSFT	Triceps skin fold thickness-(mm)
A5	V5	INS	2-Hour serum insulin-(mu U/ml)
A6	V6	BMI	Body mass index-(kg/m2)
A7	V7	DPF	Diabetes pedigree function
A8	V8	Age	Age
CLASS		DM	1. tested_Positive:(diabetic) 2. tested_Negative:(non diabetic)

sure. Gini index based decision tree algorithm SLIQ was proposed by Mehta et al. in 1996 [4]. Traditional decision tree algorithms segregates crisp set recursively till the total set belongs to either of the classes which has sharp boundary problem [5].

Different decision tree algorithms are proposed depending upon different types of attribute split measures like information gain, gain ratio, gini index and entropy measure. To overcome the sharp boundary problem, vagueness and ambiguity of data sets, different Fuzzy Decision Tree (FDT) induction algorithms are developed. FuzzyID3 algorithm was proposed by combining the fuzzy sets with induction decision tree [6].

A membership function describes the measure of degree of equality of an element to fuzzy set. Identification of relevant fuzzy membership function plays a vital role in the fuzzy decision trees. Selection of fuzzy membership function depends on the expertise and it is done manually. The fuzzy membership function is designed in such a fashion where the closer values to the split makes fuzzy membership value nearer to numeric value one. There are many fuzzy membership functions available in the literature like triangular, trapezoidal, bell shaped, Gaussian and many more. Chandra B. and Paul Varghese [7] used triangular fuzzy membership function while implementing gini index fuzzy SLIQ decision tree algorithm. They also used new fuzzy membership function to implement fuzzyfying gini index based decision trees and they achieved promising improvement in the efficiency. Kamadi V.S.R.P. Varma et al. [8] used Gaussian fuzzy membership function for a better diagnosis of diabetic patients using computational intelligence approaches through fuzzy SLIQ decision trees.

Shankaracharya et al. [9] presented a review article on Computational Intelligence techniques on early diabetes detection. They discussed different approaches used by different authors on Pima Indian Diabetes (PID) data sets like data analysis through logistic regression, clustering techniques, support vector machines, neural networks, Neuro Fuzzy Inference Systems (NFIS), expert systems and Modified Mixture of Experts (MIME) techniques.

From the literature survey, it is evident that much work is not adopted by combining dimensionality reduction approach with fuzzy SLIQ decision trees. This motivated us to take up this approach to investigate classification accuracy on PID data set. With this, the authors also propose a new fuzzy membership function in the methodology.

3. Data description

Knowler et al. made studies on Pima Indian women of at least 21 years old and living at Phoneix, Arizona, USA. According to Knowler's reports, the incidence and prevalence of diabetes among Pima Indians is higher [10]. Sankaracharya et al. presented a review on computational intelligence in early diabetes diagnosis where the article is articulated with different authors' approaches tested with PID dataset [9] (Tables 1 and 2).

Table 2
Brief Statistical Analyze of Diabetes Disease Dataset.

Attribute	Mean	Standard deviation	Min/max
1	3.8	3.4	1/17
2	120.9	32	56/197
3	69.1	19.4	24/110
4	20.5	16	7/52
5	79.8	115.2	15/846
6	32	7.9	18.2/57.3
7	0.5	0.3	0.0850/2.3290
8	33.2	11.8	21/81

4. Proposed model

4.1. PCA and FSDT method layout

The detailed work flow the proposed method is shown in Fig. 1.

4.2. Principal Component Analysis (PCA)

4.2.1. Data standardization

From the literature it is evident that data with different parameters, different units and scales requires data standardization process. Normalization or Standardization of data plays a vital role in data mining. The transformation equation is as follows.

$$X' = \frac{\left(X - \bar{X}\right)}{S_X} \tag{4.1}$$

Where,

$\bar{X}$ = Mean
X = Attribute Value
S_X = Standard Deviation

4.2.2. Dimensionality reduction

Pattern Recognition Techniques use dimensionality reduction approaches to obtain better accuracy of classification algorithms. Dimensionality reduction is a pre-processing stage which reduces the high dimensionality data into a manageable size, keeping the original information intact. The dimensionality reduction has three important stages as follows:

- Extraction of factors using Factor Analysis
- Verify the validity and internal consistency
- Calculating the new scales

4.2.2.1. Factor analysis. Factor analysis is a statistical technique to express the inconsistency among the observed variables in terms of factors. It searches for combined variations in response to unseen veiled variables. The set of variables in a data set is reduced using

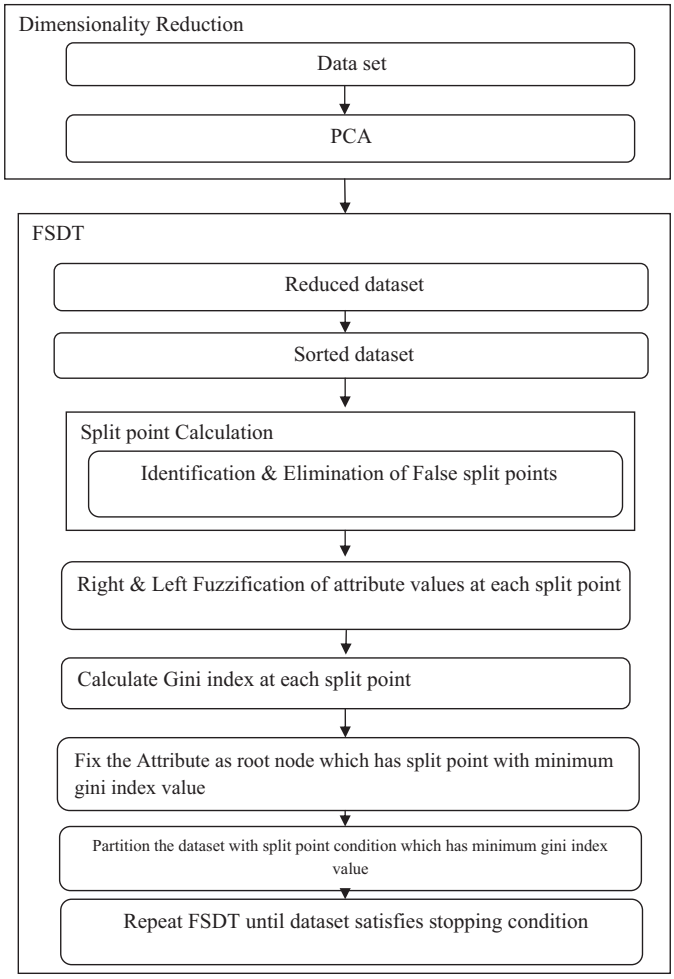

Fig. 1. Method layout.

the information observed about the inter dependencies between observed variables. The common factor model can be described as,

$$X_i = a_{i1}f1 + a_{i2}f2 + \ldots\ldots\ldots\ldots\ldots + a_{in}fn + e_i \quad (4.2)$$

Where,

X_i is the ith variable

a_{ij} is the jth factor loading for the ith variable: j = 1,2.....n

$f1, f2.....fn$ are uncorrelated factors and ei is the error term

The factor analysis has been performed with statistical package for social sciences SPSS 16.0. There are several methods available in SPSS 16.0 to extract the factors. The results used in this paper are taken from principal factor analysis.

4.2.2.2. Major points in factor analysis.

- Identification of the low communality variables and
- Identification of variable without loading on any factors or a variable with cross loadings.

Step 1: Identify the low communality variables that are less than 0.5 which has less required levels of justification. The communalities of the Pima Indian data set variables are shown in Table 3. All variable values are greater than 0.5 after extraction hence all variables have propound importance in the justification.

Step 2: Check the factor loadings in unrotated factor matrix. In this Principal Component Analysis technique we adopted varimax rotation to minimize the cross loadings. The factor loading matrix which is also known as Rotated Component Matrix is shown in Table 4.

Table 3
Measures of Communalities.

Communalities					
Variable	Initial	Extraction	Variable	Initial	Extraction
V1	1.000	0.800	V5	1.000	0.801
V2	1.000	0.783	V6	1.000	0.826
V3	1.000	0.528	V7	1.000	0.964
V4	1.000	0.759	V8	1.000	0.832

Extraction method: principal component analysis.

Table 4
Measures of Rotated Component Matrix.

Rotated Component Matrix				
	Component			
	1	2	3	4
V8	0.900			
V1	0.893			
V3	0.511			
V6		0.888		
V4		0.854		
V5			0.880	
V2			0.847	
V7				0.971

Table 5
KMO and Bartlett's Test.

KMO and Bartlett's Test		
Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		0.633
Bartlett's Test of Sphericity	Approx. Chi-Square	467.454
	df	28
	Sig.	0.000

Bryman et al. [11] proposed that Kaiser Meya Olkin (KMO) is one of the check measures to confirm the sufficiency and soundness of the statistics. The KMO is used to exhibit proportion of variance in variable. The variable having value 0.5 or less is considered to be improper otherwise the variable is proper. In this work we obtained KMO value as 0.633 which is above the acceptable lower value 0.5 and significance value is 0.000 which is less than the lower range value 0.005. These measures illustrate the suitability of the factor analysis used for the model. The KMO and Bartlett's measures are shown in Table 5. Total variance measures are shown in Table 6.

4.2.2.3. Verification of internal consistency and validity. The internal consistency and validity of the factors are verified using three methods.

- Factor loadings of the attributes
- Average variance extracted
- Calculation of Cronbach's Alpha

4.3. Factor loadings of the attribute

The value of the factor loadings is one of the important features for the examination of the validity of the factors with Confirmatory Factor Analysis (CFA) [12]. An acceptable level of factor loadings is 0.5, but 0.7 or above is suggested. Out of 8 factor loadings one is 0.5 and the remaining all are above 0.7 which is an excellent indication of the soundness of the factors.

Table 6
Total Variance Measures.

Total Variance Explained									
Component	Initial Eigen values			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2.651	33.143	33.143	2.651	33.143	33.143	1.946	24.319	24.319
2	1.532	19.146	52.289	1.532	19.146	52.289	1.771	22.134	46.453
3	1.191	14.883	67.172	1.191	14.883	67.172	1.565	19.568	66.020
4	0.919	11.486	78.658	0.919	11.486	78.658	1.011	12.638	78.658
5	0.656	8.198	86.857						
6	0.437	5.460	92.317						
7	0.340	4.249	96.565						
8	0.275	3.435	100.000						

Extraction Method: Principal Component Analysis.

4.4. Average variance extracted

The average percentage of Variance Extracted (VE) among set of crated objects is a summary indicator of convergence. This is calculated using the below mentioned formula.

$$VE = \frac{\sum_{i=1}^{n} \lambda_i^2}{n} \quad (4.3)$$

where, λ is standardized factor loading, n is number of items

The measured value of VE is 0.5, which indicates satisfactory levels.

4.5. ***Calculation of*** *Cronbach's Alpha va****lue***

Internal consistency among the variables is identified using Cronbach's Alpha value. This indicates the close relation among the set of items that are newly computed using PCA. This is measured as a function of number of items and the average inter-correlation among items. According to Nunnally et al. [13] the Cronbach's Alpha values of 0.6 and above are considered as the sufficient for testing the reliability of dimensions. The Cronbach's alpha is measured using the below equation.

$$\alpha = \frac{N\overline{C}}{\gamma + \left((N-1)*\overline{C}\right)} \quad (4.4)$$

where,

N is the number of items and $\overline{C}$ is the average inter correlation and $\overline{\gamma}$ is the average variance.

4.5.1. Calculation of new scales

From the literature there are three different approaches available to calculate the new scales.

4.5.1.1. First method. The surrogate variables which has highest loading on each factor are selected as representative variables for each factor.

4.5.1.2. Second method. In this method the items in the scale are summed and averaged.

4.5.1.3. Third method. In this approach the factor scores are computed using composite measure of each factor computed for each variable.

In our work, the new scales are calculated using second method.

4.6. Split point calculation

The records in the data are split into two separate child nodes using impurity measure. The best split is chosen based on the impurity of the child nodes. The smaller the degree of impurity, the more skewed the class measure distribution. Different types of impurity measures are available in the literature as follows [14]. A best attribute value is chosen as split point as a decision node for binary partitioning of the records with respect to the class information. Chandra B and Paul Varghese used the average of the records of a sorted attribute if the class change occurs among simultaneous records in the list of attributers in fuzzy SLIQ decision tree algorithm and they eliminated the split points with same attribute values. The author Kamadi V.S.R.P. Varma et al. proposed a Gaussian Gini index Fuzzy SLIQ Decision Tree algorithm (GG-FSDT) [8]. The author proposed false split points identification and elimination which decreases the computational effort. In this article, the author used the false split point identification and elimination approach to identify the split measures. The best split is identified using one of the node impurity measures known as gini index approach.

False Split Point Identification and Elimination Criteria

1. Simultaneous class change occurs between the similar records of a sorted attribute list
2. The similar records in the attribute have same records above or below the split point with different class information of a sorted attribute list.

4.7. Fuzzy membership function

A membership function provides the degree of similarity measure of an element to a given fuzzy set. Membership functions are chosen randomly by the authors with evident knowledge and experience of membership functions or intended by using machine learning techniques. There are different types of fuzzy membership functions that are used in the earlier methods like triangular, trapezoidal, Gaussian, bell-shaped and etc. Chandra B and Paul Varghese used triangular fuzzy membership functions. Hameed used Gaussian membership function for improving the reliability and robustness of students' evaluation systems and achieved improving results because it has less degree of freedom, smooth transition between levels and accurate representation of input-output association [15].

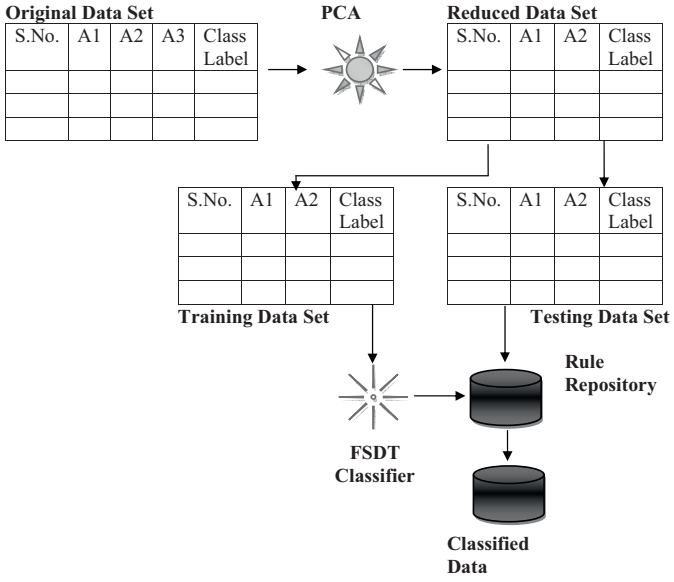

Fig. 2. Working model of PCA and FSDT classifier.

4.7.1. Triangular fuzzy membership function

$$\text{fuzzy value} = \begin{cases} \dfrac{lw}{lp + lw - val}; \text{val}\langle lp \\ 1.0; \text{lp} \le \text{val} \le \text{rp} \\ \dfrac{rw}{val - rp + rw}; \text{val}\rangle \text{rp} \end{cases} \quad (4.5)$$

where,

If attribute value ≤ splitpoint

$lw = \alpha * \sigma$
$rw = 0.0$
$lp = \text{splitpoint-}\beta$
$rp = \text{splitpoint} + \beta$

If attribute value ≥ splitpoint

$lw = 0.0$
$rw = \alpha * \sigma$
$lp = \text{splitpoint-}\beta$
$rp = \text{splitpoint} + \beta$

β takes values between zero and one
σ denotes standard deviation
lw and *rw* are the parameters for the left and right sides of the attribute split point
α average value of the distinct classes in the dataset

4.7.2. Gaussian fuzzy membership function

$$fuzzy\ \ value = exp^{\frac{-(val-splitpoint)}{2(\sigma)^2}2} \quad (4.6)$$

where, '*val*' represents the attribute value in the data set, σ denotes standard deviation.

4.7.3. Proposed fuzzy membership function

$$fuzzyvalue = \begin{cases} 1 - \dfrac{1}{2}\exp(\sigma(val\text{- }split\ \ point)) \ ; if\ val \le split\ point \\ 1 - 1/[1 - 2\ \exp((\sigma(val\ -\ split\ \ point)))]\ ; if\ val \ge split\ \ point \end{cases} \quad (4.7)$$

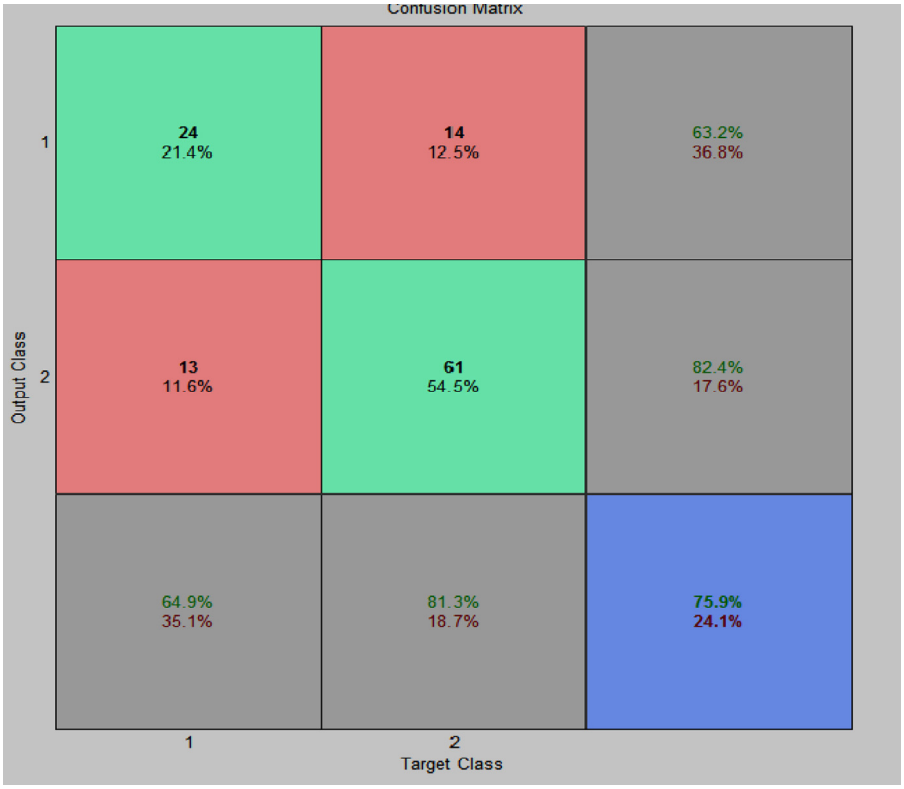

Fig. 3. Test set 1 Confusion Matrix.

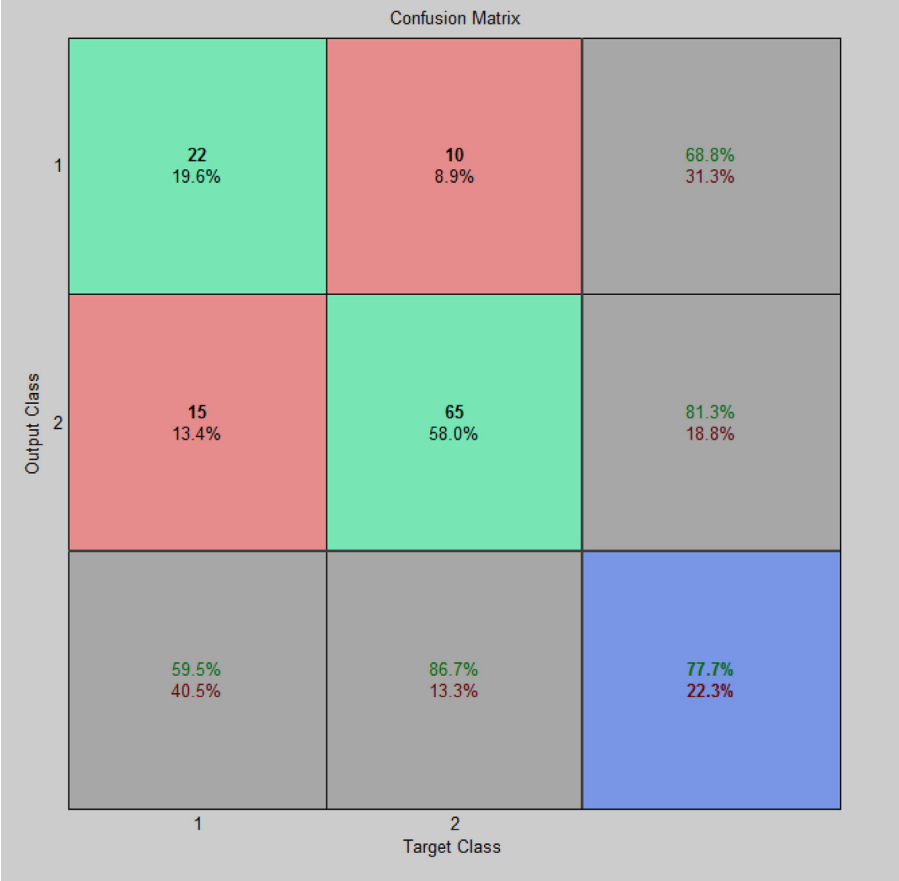

Fig. 4. Test set 2 Confusion Matrix.

where, '*val*' represents the attribute value in the data set, σ denotes standard deviation.

In our proposed method the fuzzy membership function is implemented to achieve less degree of freedom and smooth transition between levels of the fuzzy values. The proposed membership function keeps maximum fuzzy values closer to the split point.

4.8. Gini index calculation

One of the impurity measures which were used earlier in many models is gini index measure. The gini index is calculated at every split point of the ordered attribute list and the best split point is identified as the split point with minimum gini index. The corresponding attribute is taken as root node and the remaining records in the data set are separated using the best split point. Chandra B and Paul Varghese used the below mentioned gini index formula [7,16].

$$D\left(X_j\right) = \sum_{v=1}^{v} \frac{N^{(v)}}{N^{(u)}} \left[1 - \sum_{k=1}^{C} \left(\frac{N_{wk}^{(v)}}{N^{(v)}}\right)^2\right] \tag{4.8}$$

where, X_j = Selected split point

$N^{(v)}$ = Sum of v^{th} partition's records fuzzy membership values

$N^{(u)}$ = Sum of u^{th} partition's records fuzzy membership values

$N_{wk}^{(v)}$ = Sum of the product of fuzzy-membership values of the attribute and the fuzzy-membership values of the corresponding records with class w_k in partition (v)

C = Total number of distinct classes in the dataset

V = Total number of partitions

4.9. Experimental design

The working model of PCA and FSDT classifier is shown in Fig. 2. The algorithm of this model is explained as follows.

Algorithm:

Step 1: Read the training data set.

Step 2: Arrange the each attribute in sorted order and calculate the split points.

Step 3: Identify and eliminate false split points to obtain the actual split points in the corresponding attribute list.

Step 4: Apply the fuzzy membership function at each split point to convert the attribute information as fuzzy membership values.

Step 5: Measure the Gini index at each split point.

Step 6: The split point with minimum gini index is identified as the best split condition and the corresponding attribute as the root node.

Step 7: Construct a binary decision tree with obtained root node and split information.

Step 8: Repeat the process from step 2 for each node if the stopping criterion is not satisfied.

4.10. Stopping criterion

- If all the records in the data set belong to the same class
- If all have similar records in the attribute list
- If the dataset is empty

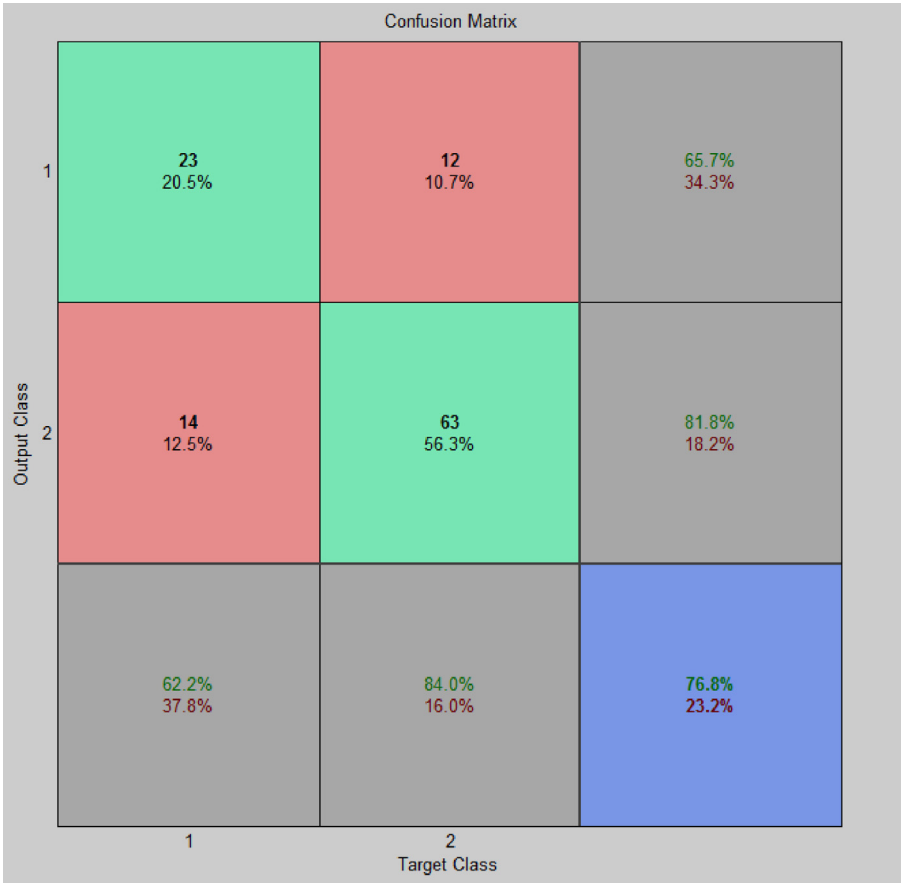

Fig. 5. Test set 3 Confusion Matrix.

Table 7
Performance Metrics.

Measure	Formula
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$
Sensitivity	$\frac{TP}{TP+FN}$
Specificity	$\frac{TN}{TN+FP}$

Table 8
New Dimensions Identified using PCA.

New Attribute	Feature Combinations
X1	Pregnant, DBP & Age
X2	TSFT & BMI
X3	Glucose, INS
X4	DPF

Table 9
Average Variance Extracted using PCA.

Factor	Average Variance
Factor 1	0.62228
Factor 2	0.74893
Factor 3	0.74590
Factor 4	0.94281

Table 10
Measures of Cronbach's Alpha Value.

Factor	Cronbach's Alpha	Number of items
1	0.731	3
2	0.774	2
3	0.708	2

Table 11
Number of Split Points Obtained in Each Fold.

S.No.	Fold	Split Points		
		FSDT	GGFSDT	PCA + FSDT (Proposed Method)
1	Training set-1	2762	1652	768
2	Training set-2	5541	3388	1478
3	Training set-3	6099	3889	1656

Table 12
Summary of the Confusion Matrix for both Training and Testing.

Fold	Testing		
	Accuracy	Sensitivity	Specificity
1	75.9	64.9	81.3
2	77.7	59.5	86.7
3	76.8	62.2	84.0

5. Performance metrics and k-fold cross validation

The most common metrics of performance measure in medical diagnosis classification techniques are accuracy, sensitivity and specificity. Accuracy is defined as the ability of the model to correctly predict the class label of previously unseen or new data. Sensitivity measures the ability of the method to identify the occurrence of target class accurately. Specificity measures the ability of the method to separate the target class [16]. The Accuracy, Sensitivity and Specificity are measured as follows (Table 7).

Table 13
Classification Accuracy for PID Data Set.

S. No	Method	Author	Average testing accuracy (%)	References	Year
1	C 4.5	J.R. Quinlan	65.06	[3]	1993
2	SLIQ	Mehta. M., R. Agarwal, and J. Rissanen	67.92	[4]	1996
3	k-NN	Ster. Dbobnikar	71.9	[20]	1996
4	CART	Ster. Dbobnikar	72.8	[20]	1996
5	MLP	Ster. Dbobnikar	75.2	[20]	1996
6	CART-DB	Shang. Breiman	74.4	[22]	1996
7	Naïve Bayes	Friedman	74.5	[23]	1997
8	GNG	Deng. Kasabov	74.6	[21]	2001
9	GCS	Deng. Kasabov	73.08	[21]	2001
10	RBF	Kayaer K.	68.23	[24]	2003
11	GFDT	B. Chandra, P. Paul Varghese	74.94	[18]	2009
12	SSVM	Purnami S.W. et al.	76.73	[25]	2009
13	SGFDT	Hongze Qie, Haitang Zhang	74.09	[19]	2011
14	GGFSDT	K.V.S.R.P Varma et al.	75.8	[8]	2013
15	PCA-FSDT	**Our Method**	**76.8**		

where, TP: True positives, TN: True negatives, FP: False positives, FN: False negatives.

k-fold cross validation is the best quantity for classifier performance. The k-fold cross-validation method segments the data into k equal-sized partitions [17]. The average of the k different test results gives the accuracy of the algorithm. In this article we used 3-fold cross validation.

6. Results and Discussion

In this paper the authors developed a Computational Intelligence technique using Principal Component Analysis and modified Gini index- Fuzzy SLIQ Decision Tree algorithm for the diagnosis of diabetes. We have proposed an expert system which has two stages. In the first stage the PID data set with eight attributes is reduced to four attributes using Principal Component Analysis (PCA) and in the second stage decision rules are constructed using modified Fuzzy SLIQ Decision Tress algorithm (FSDT). The method performance was verified by 3-fold cross validation approach in each fold. Out of 336 whole data set 112 items are used for testing and the reaming 224 items are used for training the algorithm. PCA is used to extract the local, global, structural and statistical features and dimensionality reduction. Statistical Package for Social Sciences SPSS 16.0 is used to Principal Component Analysis. The validity of the factors extracted with this method is confirmed using Confirmatory Factor Analysis (CFA) technique. The significance of the separation between the identified dimensions is verified using Cronbach's Alpha value. According to PCA results out of eight attributes of PID data set finally four factors are identified. The new attributes that are identified through PCA is presented in Table 8. Factors and the corresponding average variance extracted by them are presented in Table 9. The Cronbach's Alpha values are as shown in the Table 10.

The factor scores are computed using composite measure of each factor computed for each variable and this new factor scores are used as input to the modified fuzzy SLIQ decision tree algorithm. In each fold out of 336 whole data set 112 items are used for testing and the remaining 224 items are used for training the algorithm. Table 11 presents the number of split points obtained with Fuzzy SLIQ Decision Tree algorithm (FSDT), Gaussian Gini index Fuzzy SLIQ Decision Tree algorithm and the proposed method (PCA + FSDT). The results show that the proposed method reduces computational effort by reducing the number of split points for each training fold. The reduced split points in the training set directly reduces the gini index calculation in the corresponding training set. However the average accuracy achieved with the proposed model is considerably high when compared with the previous models FSDT and GGFSDT.

Summary of the confusion matrix for testing performance metrics are shown in Figs. 3–5 .

Table 12 presents the test accuracy, sensitivity and specificity obtained with 3 folds of the PID test data set. The method achieves an average test accuracy of 76.8% which is a prominent result when compared with the earlier model test accuracy with the PID data set which is presented in the Table 13.

7. Conclusion

In this paper the computational intelligence technique for diabetes diagnosis was simulated using Principal Component Analysis and Modified Fuzzy SLIQ Decision Tree algorithm with the help of SPSS and MATLAB. The work was carried out with 336 data records collected from UCI machine learning repository. The accuracy obtained with this model was 76.8%, which outperformed when compared with the earlier models. The accuracy of the model can be improved by using relevant and better fuzzy membership functions which are applicable to diabetes clinical data. In future work we are planning to extend the model for the diagnosis of other diseases with better fuzzy membership functions.

Acknowledgement

Allam Appa Rao acknowledges the financial support of DST-IRHPA Scheme vide Lr No. IR/SO/LU/03/2008/1 dated 24-12-2010.

References

[1] Y. Zhang, The economics costs of undiagnosed diabetes, Popul. Health Mange. 12 (2) (2009) 95–101.
[2] J.R. Quinlan, Introduction of decision tree, Mach. Learn. 1 (1986) 86–106.
[3] Quinlan J.R., Morgan Kaufmann, 1993. Programs for Machine Learning. San Mateo, California.
[4] M. Mehta, R. Agrawal, J. Riassnen Avignon, SLIQ: A fast scalable classifier for data mining in Extending Database Technology, Springer, France, 1996, pp. 18–32.
[5] Myung Won Kim, Ara Khil, Joung Woo, Ryu Efficient Fuzzy Rules for Classification, Proceeding of the International Workshop on Integrating AI and Data Mining (AIDM'06) (2006).
[6] M. Umano, et al., Fuzzy Decision Trees by Fuzzy Id3 Algorithm and Its Application to Diagnosis Systems, vol. 3, Fuzzy Systems, IEEE, 1994. Fuzzy Systems, IEEE World Congress on Computational Intelligence, Oriando, FL, 2016, pp. 2113–2118.
[7] B. Chandra, P. Varghese Paul, Fuzzy SLIQ decision tree algorithm, IEEE Trans. Syst. Man Cydernetics-Part B 38 (October (5)) (2008) 1294–1301.
[8] V.S.R.P. Kamadi Varma, et al., A computational intelligence approach for a better diagnosis of diabetic patients, Comput. Electr. Eng. (August) (2013).
[9] Shankaracharya, Computational intelligence in early diabetes diagnosis: a Review, Rev. Diabetic Stud. 7 (January (4)) (2011) 252–262.
[10] C. William Knowler, et al., Diabetes incidence and prevalence in pima indians: a 19-fold greater incidence than in Rochester, Minnesota, Am. J. Epidemiol. 108 (6) (1978) 497–505.

[11] A. Bryman, D. Cramer, Quantitative data analysis with SPSS release 8 for Windows, in: A Guide for Social scientists, London, Routledge, 1999.
[12] T. Sitamahalakshmi, et al., Extraction of dimensions for telugu character recognition: a case study, research India publication, Adv. Comput. Sci. Technol. 4 (2011) 113–126, ISSN 0973-6107.
[13] J.C. Nunnally, I.H. Bernstein, Psychometric Theory, 3rd ed., McGrawHill, New York, 1994.
[14] Pang-Ning Tan, Vipin Kumar, Introduction to Data Mining, Pearson, 2007.
[15] I.A. Hameed, Gaussian membership functions fro improving the reliability and robustness of students, evaluation system, Expert Syst. Appl. (April) (2011).
[16] Jiawei Han, Micheline Kamber, Jian Pei, Data Mining Concepts and Techniques, Morgan Kaufmann, Waltham, 2012.
[17] G. Walker, A. Kadam, Predicting breast cancree dater survivability: a comparison of three data mining methods. Delen, D, Artif. Intell. Med. 34 (2) (2005) 113–127.
[18] B. Chandra, P. Paul Varghese, Fuzzifying Gini Index based decision trees, Expert Syst. Appl. 36 (2009) 8549–8559.
[19] Qiu Hongze, Haitang Zhang, Fuzzy SLIQ decision tree based on classification sensitivity I, J. Mod. Educ. Comput. Sci. 5 (August) (2011) 18–25, China.
[20] B. Ster, A. Dobnikar, Neural networks in medical diagnosis: comparison with other methods, in: International Conference on Engineering Applications with Neural Networks, London, 1996, pp. 427–430.
[21] D. Deng, N. Dunedin Kasabov, On-line pattern analysis by evolving self-organizing maps, 5th Biannual Conference on Artificial Neural Networks and Expert Systems (ANNES) (2001) 46–51.
[22] N. Shang, L. Breiman, Distribution based trees are more accurate, Springer ICONIP96, Hong Kong, 1996, pp. 133–138.
[23] N. Friedman, D. Geiger, M. Goldszmit, Bayesian networks classifiers, Mach. Learn. 29 (1997) 131–163.
[24] K. Kayaer, T. Yildirim, Medical diagnosis on Pima Indian diabetes using general regression neural networks. Proceedings of the international conference on artificial neural networks and neural information processing, Istanbul (2003) 181–184.
[25] S.W. Purnami, A. Embong, J.M. Zain, A New smooth support vector machine and its applications in diabetes disease diagnosis, J. Comput. Sci. 5 (2009) 1006–1011.

Kamadi V.S.R.P. Varma received the M.Tech degree in Computer Science and Technology from Andhra University, India in 2008. Working towards the Ph.D. degree at Dept.CSE, JNTUH-Hyderabad. Presently working as Assistant Professor, Department of CSE, GITAM University, Visakhapatnam. Research interests include Data Mining, Fuzzy Logic, Genetic Algorithms, Big Data and Bioinformatics.

Allam Appa Rao was the first to receive Ph.D from Andhra University in Computer Engineering. Research contributions in Computer Engineering are vital, varied and vast. Received "Srinivas Ramanujan Birth Centenary Award" from Indian Science Congress Association (ISCA). Research interest includes Bioinformatics, Soft computing, Big Data and Computational Biology.

T. Sita Mahalakshmi received M.Tech (CST) from Andhra University and Ph.D from Nagarjuna University and had 25 years of teaching and research experience. Presently working as Professor, Department of CSE, GITAM University, Visakhapatnam. Worked on hand written telugu script recognition using soft computing techniques. Research interests include Genetic algorithms, Neural Networks and Data Mining.

P.V. Nageswara Rao received M.Tech (CST) from Andhra University and Ph.D from Nagarjuna University and had 20 years of teaching and research experience. Presently working as Professor, Department of CSE, GITAM University, Visakhapatnam. Research interests include Bioinformatics, Soft Computing and Data Mining.