Performance Research on Medical Data Classification using Traditional and Soft Computing Techniques

Zahid Ansari, Quazi Mateenuddin H., Ansari Abdullah

ABSTRACT--- The world today has made giant leaps in the field of Medicine. There is tremendous amount of researches being carried out in this field leading to new discoveries that is making a heavy impact on the mankind. Data being generated in this field is increasing enormously. A need has arisen to analyze these data in order to find out the meaningful and relevant hidden patterns. These patterns can be used for clinical diagnosis. Data mining is an efficient approach in discovering these patterns. Among the many data mining techniques that exists, this paper aims at analyzing the medical data using various Classification techniques. The classification techniques used in this study include k-Nearest neighbor (kNN), Decision Tree, Naive Bayes which are hard computing algorithms, whereas the soft computing algorithms used in this study include Support Vector Machine (SVM), Artificial Neural Networks (ANN) and Fuzzy k-Means clustering. We have applied these algorithms to three kinds of datasets that are Breast Cancer Wisconsin, Haberman Data and Contraceptive Method Choice dataset. Our results show that soft computing based classification algorithms better classifications than the traditional classification algorithms in terms of various classification performance measures.

Keywords— Medical data Mining, Classification, Soft computing techniques

I. INTRODUCTION

In the field of medical, there is a tremendous amount of data being generated [1]. This data if properly processed can provide useful information. For instance, a medical study could find a link between the successes of a medicine with the patient's height. The reality might be that when the data includes information on parameters including height, body weight, eye color, shoe size and more some connections may appear to be important just by chance. We can implement big data systems that can analyze DNA in a matter of minutes. When these systems go through vast amounts of medical data, they find patterns. These patterns could someday lead us to find cures to some of the deadliest diseases. Governments can now analyze data regarding the drugs that are prescribed to patients by doctors in the public sector. This helps them gain a clear picture of the types of drugs that are being prescribed, thereby helping to understand whether the patients are receiving the most up to date medicines.

Since there is tremendous amount of medical data managing them is a difficult task. Different data mining techniques are available. Data can be handled well if it is properly classified, for example, we can classify different medical data like BMI of a person as lean, normal, fat and obese. Some of the important applications of data mining techniques in the field of medicine include health informatics, medical data management, patient monitoring systems, analysis of medical images for unknown information extraction and automatic identification of diseases [2].

To analyze these huge amount of data easily one of the data mining techniques we can use is Classification. Classification is a popular data mining technique used to categorize unlabeled data points using a training data set of pre-labeled data points by developing a classification model to classify new data points [3]. Classification techniques are used to accurately classify the patient records into appreciate class categories. Several major kinds of classification algorithms including C4.5 [4], ID3 [5], K-nearest neighbor classifier [6], Naive Bayes [7], Support Vector Machine (SVM) [8] and Artificial Neural Networks (ANN) [9] are used for classification.

II. ISSUES RELATED TO THE TRADITIONAL DATA CLASSIFICATION TECHNIQUES

Medical data mining is the process of exploration of large quantities of data in order to discover meaningful patterns. Real world medical data involve huge databases usually contain missing, inaccurate, noisy or inconsistent data.

Imperfect data [14] may:

- negatively impact on pattern discovery, results are not trustworthy [24]
- Increases the chances of discovering spurious patterns. [25][26]

May lead to over fitting of the models [27].

Traditional classification techniques don't handle these data imperfections, noises, outliers etc. in the data and thus lead to less accurate results. One of the most common uses of medical data mining is to predict what disease a patient will suffer from based on his lifestyle. The traditional techniques will classify a patient to only one group indicating that he is at risk of suffering from only one kind of disease. But in fact, he may suffer from multiple diseases (belonging to different groups of classification at the same

Revised Manuscript Received on July 10, 2019.

Zahid Ansari, Department of Computer Science and Engineering, P A College of Engineering Nadupadavu, Mangalore. Karnataka India (zahid_cs@pace.edu.in)

Quazi Mateenuddin H., Faculty of Electronics and Communication Engineering, Indian Naval Acadamy, Ezhimala. Kerala, India. (mateen@rediffmail.com)

Ansari Abdullah, Department of Computer Science and Engineering, Bearys Institute of Technology, Mangalore.Karnataka, India. (ansaridx99@gmail.com)

time) and not only one disease. To handle these issues soft computing based classification techniques are used.

III. SOFT COMPUTING TECHNIQUES

Soft computing techniques are a collection of methodologies that

- Exploit the tolerance for imperfection and uncertainty
- Provide capability to handle real life ambiguous situations
- Try to achieve robustness against imperfection.

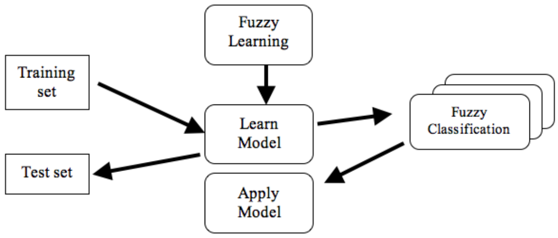

Fig. 1 Fuzzy Classification Block diagram

One of the most popular soft computing based classification techniques is Fuzzy classification. Fuzzy classes can better represent transitional areas than hard classification, as class membership is not binary but instead one location can belong to a few classes. In fuzzy set based systems membership values of data items range between 0 and 1, where 1 indicates full membership and 0 indicates no membership. Fig. 1 shows a block diagram of fuzzy classification technique.

IV. ARCHITECTURAL DESIGN

This section outlines various layers of analysis framework. The analysis framework is divided into user-interface layer and processing layer. The user interface layer is responsible for taking input from the user and processing. The processing layer is responsible for classification and comparison. The data access layer is responsible for connecting the application to the database. The figure shows the system architecture and the interaction between various components. Each layers are implemented using class files which will implement interfaces and processing of the data.

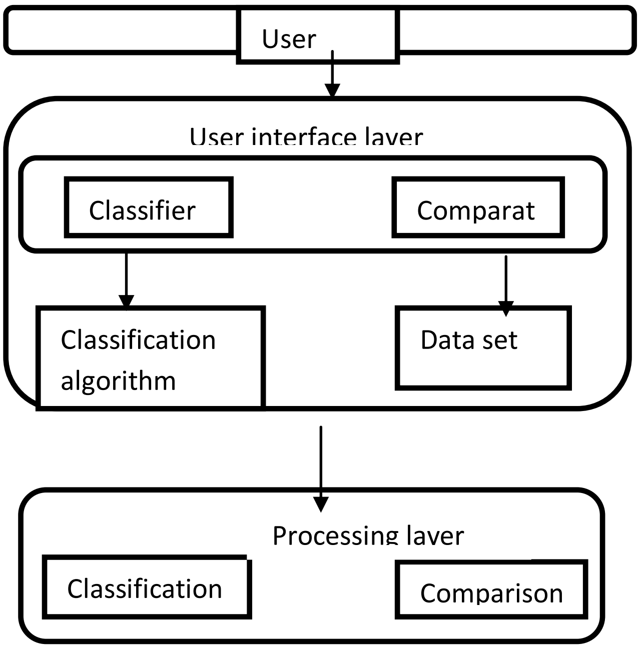

Fig. 2 Analysis Framework Architecture

Fig. 2 depicts that analysis framework consists of two layers one is User interface layer which allows the user to select the desired datasets and algorithms. In the processing layer the user chosen algorithms are applied on the selected data sets and the results are compared.

V. ALGORITHMS

K Nearest Neighbour

One of the most straightforward classifier in machine learning is the Nearest Neighbor Classifier. Classification is achieved by identifying the nearest neighbors to a query example and using those neighbors to determine the class [18] of given data point. Here K indicates the number of neighbors which can be chosen arbitrarily. Varying the value of K may result in the formation of different cluster sets. Various algorithmic steps are described below:

```
Classify (X, Y, x) // X: Training Data,   //Y: class labels of X, x: unknown sample
1. for i=1 to m do
   Compute distance d(X, x)
   end for
2. Compute a set of indices of k nearest neighbor data points having the smallest distances $d(X_i, x)$.
3. return majority label for $\{Y_i, i \in I\}$
```

Decision Tree

One of the most popular and widely used classification technique is Decision tree based classification [19]. In decision tree classifications a test record undergoes tests against a number questions related to its attributes. After each answer next question is asked related to some other important attribute. This testing continues till the class of category of the given test record is identified.

The algorithm is as follows:

```
1. t = createNode()
```

DOI : 10.35940/ijrte.B1185.0782S319

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

```
2. label(t) = mostCommonClass(D, Target)
3. IF ∀(x, c(x)) ∈ D : c(x) = c THEN return(t) ENDIF
4. IF Attributes = ∅ THEN return(t) ENDIF
5. A∗ = $\text{argmax}_{A\in\text{Attributes}}$ (informationGain(D, A))
6. FOREACH a ∈ A∗ DO
        $D_a$ = {(x, c(x)) ∈ D : $x|_{A*}$ = a}
        IF $D_a$ = ∅ THEN
                t’ = createNode()
                label(t’) = mostCommonClass(D, Target)
                createEdge(t, a, t’)
        ELSE createEdge(t, a, ID3($D_a$, Attributes {A∗}, Target))
        ENDIF
ENDDO
7. return(t)
```

Naive Bayes

In a machine learning classification problem [20], there are multiple features and classes, say, C_1. The main aim in the Naive Bayes algorithm is to calculate the conditional probability of an object with a feature vector x_1 belongs to a particular class C.

$$\mathrm{P}(\mathrm{C_i}|\mathrm{x_1},\mathrm{x_2},\ldots\ldots,\mathrm{x_n}) = \frac{\mathrm{P}(\mathrm{x_1},\mathrm{x_2},\ldots\ldots,\mathrm{x_n}|\mathrm{C_i}).\mathrm{P}(\mathrm{C_i})}{\mathrm{P}(\mathrm{x_1},\mathrm{x_2},\ldots\ldots,\mathrm{x_n})} \quad \text{for } 1 \leq i \leq k \qquad (1)$$

Now, the numerator of the fraction on right-hand side of the equation above is

$$P(x_1,x_2,\ldots\ldots,x_n|C_i).P(C_i) = P(x_1,x_2,\ldots\ldots,x_n,C_i) \qquad (2)$$

$$\begin{aligned} P(x_1,x_2,\ldots\ldots,x_n,C_i) &= P(x_1|x_2,\ldots.,x_n,C_i)\,.\,P(x_2,\ldots\ldots,x_n,C_i) \\ &= P(x_1|x_2,\ldots.,x_n,C_i)\,.\,P(x_2|x_3,\ldots.,x_n,C_i)P(x_3,\ldots\ldots,x_n,C_i \\ &= \ldots.. \\ &= P(x_1|x_2,\ldots.,x_n,C_i)\,.\,P(x_2|x_3,\ldots.,x_n,C_i)\ldots.P(x_{n-1}|x_n,C_i).P(x_n|C_i).P(C_i) \end{aligned} \qquad (3)$$

The conditional probability term, $P(\,x_j|x_{j+1},\ldots.,x_n,C_i)$ becomes $P(\,x_j|C_i)$ because of the assumption that features are independent. Assumption of independence and above computations results in the following expression:

$$P(C_i|x_1,x_2,\ldots\ldots,x_n) = \left(\prod_{j=1}^{j=n} P(x_j|C_i)\right).\frac{P(C_i)}{P(x_1,x_2,\ldots\ldots,x_n)} \qquad (4)$$

For all classes, the expression $P(\boldsymbol{x_1},\boldsymbol{x_2},\ldots\ldots,\boldsymbol{x_n})$ is constant, therefore:

$$P(C_i|x_1,x_2,\ldots\ldots,x_n) \propto \left(\prod_{j=1}^{j=n} P(x_j|C_i)\right).P(C_i) \qquad (5)$$

Support Vector Machine (SVM)

An SVM classifier performs classification with the help of hyperplane separating data points into class categories as shown in Fig. 3. Consider the case of 2-dimensional data points corresponding to two classes, the task of SVM classifier is to find the optimal separating line. A good separating line should not pass too close to the data points rather it should pass as far as possible from all these data points. SVM algorithm discovers the separating hyperplane that has the largest minimum distance from all the data points in the training set. This results in discovery of the hyper plane for which margin of the training data points is maximized [21].

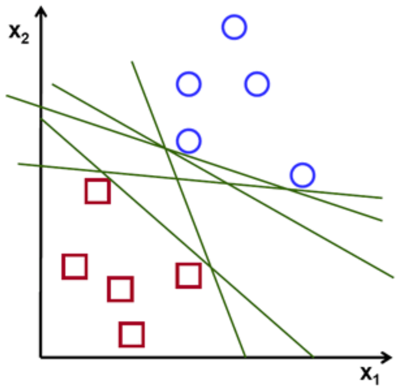

Fig. 3 Hyperplane Example

Algorithm performs the following steps:

1) Select and set up data set for training
2) Set up the SVM parameters (SVM type, Kernel type, Termination criteria of etc.)
3) Perform the Training of SVM
4) Classify the new input based on the regions

Artificial Neural Networks

Artificial neural networks are relatively crude electronic networks of neurons based on the neural structure of the brain [22]. They process records one at a time, and learn by comparing their classification of the record with the known actual classification of the record. The classification errors of previous data points are feedback for adjusting the network weights for subsequent iterations.

The algorithm is as follows:

1) Initialize the weights to small random values.
2) Randomly choose an input pattern $x^{(\mu)}$
3) Propagate the signal forward through the network.
4) Compute $\delta_i^L = g'(h_i^L)[d_i^L - y_i^L]$,

Where h_i^L represents the net input to the i^{th} unit to the lth layer, and $g'$is the derivative of the activation function g.

5) Propagate the error backwards in order to adjust the values of deltas corresponding to the preceding layers;

$$\delta_i^L = g'(h_i^l)\sum_j w_{ij}^{l+1}\delta_j^{l+1}$$

for I=(L-1),…..,1

6) Modify weights using

$$\Delta w_{ij}^l = \eta\delta_i^l y_j^{l-1}$$

7) Repeat from step 2 for next data points until convergence or number of iterations cross the specified limit.

Fuzzy K Means Based Clustering

Unlike traditional K Means clustering in which a each data point belongs to only a single cluster, in fuzzy counterpart part of K Means clustering, a data point may belong to multiple clusters with a certain degree of probability. Fuzzy k-means tries to deal with the situation where data points are somewhat in between centers of multiple clusters. Fuzzy k-means computes the weighted centroid for each cluster based on those probabilities of belongingness of the data points for that cluster [23]. The best way of assessing the resulting clusters is with the help probabilistic distributions. Fuzzy k-means algorithm is generalization of k-means where the probability value of each data point for each cluster lies between 0 and 1. The various steps of fuzzy k-means algorithm are described below:

1. ***Assume*** a fixed number of clusters k
2. ***Initialization***: Randomly initialize the k-means μ_k associated with the clusters and compute the probability that each data point x_i is a member of a given cluster k, $P(point\ x_i\ has\ label\ k|x_i, k\ )$.
3. ***Iteration***: Recalculate the centroid of the cluster as the weighted centroid given the probabilities of membership of all data points x_i:

$$\mu_k(n+1) = \frac{\sum_{x_i \epsilon k} x_i \times P(\mu_k|x_i)^b}{\sum_{x_i \epsilon k} P(\mu_k|x_i)^b}$$

4. ***Termination***: Iterate until convergence or until a user-specified number of iterations has been reached.

VI. DATASETS

The Breast Cancer Wisconsin (Original) Dataset

This dataset is extracted from UCI machine learning repository. It is used to classify the records corresponding to various breast cancer cases. These cases corresponding to types of breast cancer are classes viz. benign and malignant. The data set consists of 699 breast cancer records.

Haberman's Survival Dataset

This dataset contains survival of patients who are operated for breast cancer. It consists of a total 306 records.

Contraceptive Method Choice Dataset

The data set consists of records of the women who were either not pregnant or unsure about their pregnancy. The data set is used to predict the choice of a contraceptive method of a woman based on her other attributes. The data set consists of a total 1473 records.

VII. RESULTS AND ANALYSIS

Results on Haberman's Survival Dataset

All the six classification algorithms described above have been applied on the Haberman's survival dataset. The results are summarized in the Table I:

TABLE I CLASSIFICATION OF HABERMAN DATASET

Performance Measures Algorithms	Accuracy	Sensitivity	Specificity	Precision
K Nearest Neighbor	76.92	78.31	62.5	78.31
Decision Tree	71.73	77.92	57.14	77.92
Naïve Bayes	74.51	76.53	55.17	76.53
Support Vector Machine	73.63	77.22	50	77.22
Artificial Neural Networks	82.16	81.24	65.47	79.23
Fuzzy Clustering	98.04	100	92.86	100

Fig. 4 provides the comparative results of various classification algorithms when applied on Haberman dataset. From Table I, it is obvious that Classification based on Fuzzy clustering algorithm outperforms all the other algorithms in terms of performance measures such as accuracy, sensitivity, specificity and precision of the classification. Artificial Neural Network based classification also produces better results that all other algorithms except fuzzy clustering.

Results on Contraceptive Method Choice Dataset

Table II shows the classification results of all the six algorithms when applied on Contraceptive Method Choice (CMC) dataset.

Fig. 5 provides the comparative results of various classification algorithms when applied on CMC dataset. From Table II clearly shows that Classification based on Fuzzy clustering algorithm outperforms all the other algorithms in terms of performance measures such as accuracy, sensitivity, specificity and precision of the classification. Artificial Neural Network based classification is second to fuzzy clustering in performance.

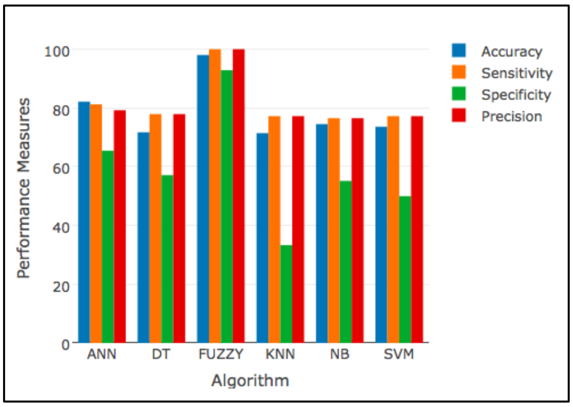

Fig.4 Comparison of Algorithm Performance (Haberman Dataset)

Table II Classification of CMC Dataset

Performance Measures **Algorithms**	**Accuracy**	**Sensitivity**	**Specificity**	**Precision**
K Nearest Neighbor	76.92	78.31	62.5	78.31
Decision Tree	71.73	77.92	57.14	77.92
Naïve Bayes	74.51	76.53	55.17	76.53
Support Vector Machine	73.63	77.22	50	77.22
Artificial Neural Networks	82.16	81.24	65.47	79.23
Fuzzy Clustering	98.04	100	92.86	100

Results on Breast Cancer Wisconsin (Original) Dataset

All the six algorithms have been applied on the Breast Cancer dataset. The results are summarized in the Table III:

Fig. 6 provides the comparative results of various classification algorithms when applied on CMC dataset. From Table III clearly shows that Classification based on Fuzzy clustering algorithm outperforms all the other algorithms in terms of performance measures such as accuracy, sensitivity, specificity and precision of the classification. Artificial Neural Network based classification is second to fuzzy clustering in performance.

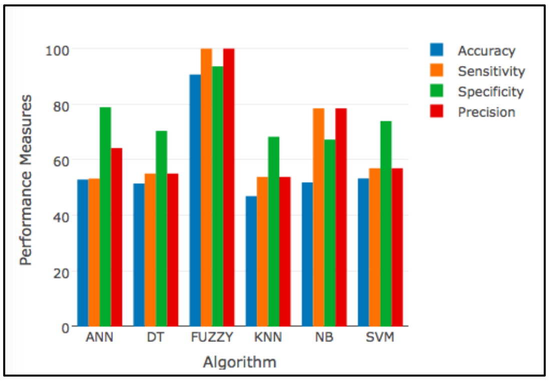

Fig.5 Comparison of Algorithm Performance (CMC Dataset)

Table III Classification of Breast Cancer Dataset

Performance Measures **Algorithms**	**Accuracy**	**Sensitivity**	**Specificity**	**Precision**
K Nearest Neighbor	76.92	78.31	62.5	78.31
Decision Tree	71.73	77.92	57.14	77.92
Naïve Bayes	74.51	76.53	55.17	76.53
Support Vector Machine	73.63	77.22	50	77.22
Artificial Neural Networks	82.16	81.24	65.47	79.23
Fuzzy Clustering	98.04	100	92.86	100

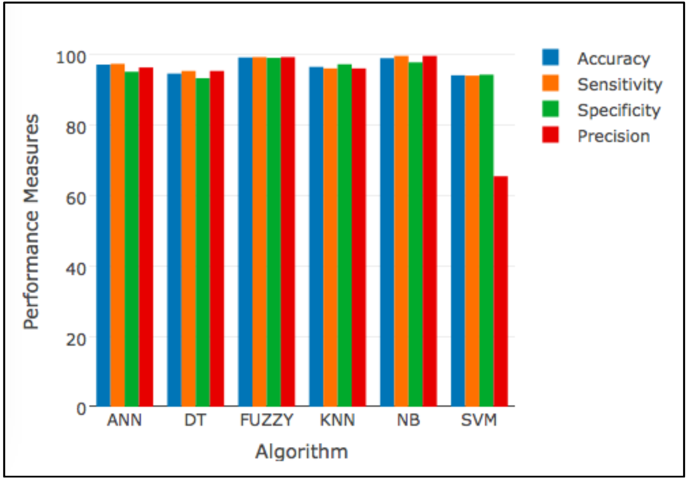

Fig 6 Comparison of Algorithm Performance (Breast Cancer Dataset)

VIII. CONCLUSION

This study deals with the performance analysis of medical data classification using various soft and hard computing based classification algorithms. Experiments were performance one three difference medical data sets. From the experimental results following conclusions may be drawn:

- The soft computing based classifications algorithms such as ANN, SVM and Fuzzy Clustering performed better than tradition or hard classifications algorithm in terms of various classification performance measures such as accuracy, sensitivity, specificity and precision.
- Among the soft computing based classification algorithms, Fuzzy clustering outperformed almost all other algorithms used.
- The performance traditional classification algorithms was inferior to their soft computing in terms of most of the performance measures.
- Also, the results vary from one dataset to another dataset. More the number of records in a dataset for training the model, better is the performance of the classification model.

ACKNOWLEDGEMENT

This study was facilitated by Government of Karnataka (GoK) Vision Group on Science and Technology (VGST) CISEE (2015-16) scheme grant GRD No-461.

REFERENCES

1. Wu, Xindong, et al. "Data mining with big data." *IEEE transactions on knowledge and data engineering* 26.1 (2014): 97-107.
2. Acharya, U. Rajendra, and Wenwei Yu. "Data mining techniques in medical informatics." *The open medical informatics journal* 4 (2010): 21.
3. Deulkar, Miss Deepa S., and R. R. Deshmukh. "Data Mining Classification." *Imperial Journal of Interdisciplinary Research* 2.4 (2016).
4. Ruggieri, Salvatore. "Efficient C4. 5 [classification algorithm]." *IEEE transactions on knowledge and data engineering* 14.2 (2002): 438-444.
5.
6. Cios, Krzysztof J., and Ning Liu. "A machine learning method for generation of a neural network architecture: A continuous ID3 algorithm." *IEEE Transactions on Neural Networks* 3.2 (1992): 280-291.
7. Dudani, Sahibsingh A. "The distance-weighted k-nearest-neighbour rule." *IEEE Transactions on Systems, Man, and Cybernetics* 4 (1976): 325-327.
8. Kim, Sang-Bum, et al. "Some effective techniques for Naive Bayes text classification." *IEEE transactions on knowledge and data engineering* 18.11 (2006): 1457-1466.
9. Chapelle, Olivier. "Training a support vector machine in the primal." *Neural computation* 19.5 (2007): 1155-1178.
10. Mili, Faissal, and Manel Hamdi. "A comparative study of expansion functions for evolutionary hybrid functional link artificial neural networks for data mining and classification." *Computer Applications Technology (ICCAT), 2013 International Conference on*. IEEE, 2013.
11. Fatima, M. and Pasha, M. (2017), “Survey of Machine Learning Algorithms for Disease Diagnostic”, Journal of Intelligent Learning Systems and Applications, 9, 1-16.
12. Sarvestani, A. Soltani, et al. "Predicting breast cancer survivability using data mining techniques." *Software technology and Engineering (ICSTE), 2010 2nd international Conference on*. Vol. 2. IEEE, 2010.
13. Shiv Shakti Shrivastava, Dr. V.K.Choubey, Dr. Anjali Sant,” Classification Based Pattern Analysis on the Medical Data in Health Care Environment”, IJSRSET Volume 2, Issue 1, ISSN : 2395-1990,2016
14. Divya Tomar and Sonali Agarwal,” A survey on Data Mining approaches for Healthcare” International Journal of Bio-Science and Bio-Technology Vol.5, No.5 (2013), pp. 241-266
15. K.K.R. Hewawasam, K. Premaratne, Mei-Ling Shyu, Rule Mining: A Belief-Theoretic Approach for Handling Data Imperfections in IEEE Transactions on Systems, Man and Cybernetics,Part B, vol 37(6),2007
16. Sushmita Mitra, Ranajit Das, Yoichi Hayashi, Genetic networks and soft computing, in IEEE Transaction on Computational Bioinformatics, vol. 8, 2011
17. Ravi Jain, Ajith. Abraham,”A Comparative study of Fuzzy Classification Methods on Breast Cancer Data”, Australas Phys Eng Sci Med 27: 147–15
18. Bezdek, J. C., 1987, Some non-standard clustering algorithms. In: Developments in numerical ecology, P. and L. Legendre, eds. Berlin: Springer-Verlag. pp. 225-87
19. Cunningham, Padraig, and Sarah Jane Delany. "k-Nearest neighbour classifiers." *Multiple Classifier Systems* 34 (2007): 1-17.
20. Bogaert, Jan, Reinhart Ceulemans, and David Salvador-Van Eysenrode. "Decision tree algorithm for detection of spatial processes in landscape transformation." *Environmental management* 33, no. 1 (2004): 62-73.
21. Stewart, B. "Predicting project delivery rates using the Naive–Bayes classifier." *Journal of Software Maintenance and Evolution: Research and Practice* 14, no. 3 (2002): 161-179.
22. Scholkopf, Bernhard, and Alexander J. Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2001.
23. Jain, Anil K., Jianchang Mao, and K. Moidin Mohiuddin. "Artificial neural networks: A tutorial." *Computer* 29, no. 3 (1996): 31-44.
24. Huang, Zhexue, and Michael K. Ng. "A fuzzy k-modes algorithm for clustering categorical data." *IEEE Transactions on Fuzzy Systems* 7, no. 4 (1999): 446-452.
25. Zahid Ansari, M.F.Azeem, A. Vinaya Babu and Waseem Ahmed. “A Fuzzy Approach for Feature Evaluation and Dimensionality Reduction to Improve the Quality of Web Usage Mining Results”. International Journal on Advanced Science Engineering and Information Technology (IJASEIT), ISSN: 2088-5334, vol. 2 no. 6, pp. 67-73. 2012.
26. Zahid Ansari, M. F. Azeem, A. Vinaya Babu and Waseem Ahmed, “A Fuzzy Clustering Based Approach for Mining Usage Profiles from Web Log Data”, International Journal of Computer Science and Information Security, ISSN 1947-5500, vol. 9, no. 6, pp.70-79 , June 2011,
27. Tanvir Sardar and Zahid Ansari, “Detection and Confirmation of Web Robot Requests for Cleaning the Voluminous Web Log Data”, Proceedings of the IEEE International Conference on Impact of E-Technology on US (IC-IMPETUS), Bangalore, India. January 10-11. 2014 pp. 13-19
28. Zahid Ansari, Syed Abdul Sattr and A.Vinaya Babu, “A Fuzzy Neural Network Based Framework to Discover User Access Patterns from Web Log Data”, Advances in Data Analysis and Classification (ADAC), Springer Berlin Heidelberg, ISSN: 1862-5347, (ISI Indexed Journal, Impact Factor 2.326) vol. 9, pp. 1-28, December 2015. doi: 10.1007/s11634-015-0228-4.