Comparative Study of Soft Computing Techniques on Medical Datasets

Mangesh Metkari[1], M.A. Pradhan[2]

[1]Pune University, All India Shri Shivaji Memorial Society's College of Engineering, Kennedy Road, Pune-1, Maharashtra, India

[2]Professor, Pune University, All India Shri Shivaji Memorial Society's College of Engineering, Kennedy Road, Pune-1, Maharashtra, India

Abstract: ***Data classification is process of dividing dataset into two or more different classes where each class contains similar type of data items. In this paper we compare the different classification technique using the WEKA tool that will be helpful for decision making in medical diagnosis. WEKA is open source tool providing classification using soft computing technique for data mining process. Our goal is to analysis of the performance of different classifiers on different medical datasets. The analysis is done for five different medical datasets with four different classifiers in terms of the execution time, correctly classified, incorrectly classified and the mean absolute error. From the obtained results of classifiers we conclude that KNN is the effective classifier for medical dataset than other classifiers we used for the analysis.***

Keywords: Random Forest, KNN, Multilayer Perceptrons, Classifier**.**

1. Introduction

Now a day different soft computing technique is widely used in medical diagnosis. The problem in medical science is in evaluation the correct diagnosis as per available information from of data taken from patient. But the some soft computing methods and intelligence system and are helpful for classification. For better diagnosis of dieses so many test are needed, these test required classification of large scale data. Data classification [1] is method of dividing dataset into two or more different classes according to the datasets and data features. Features can be selected depend on datasets or application. There are so many classification algorithms in WEKA tool [20] like Nearest Neighbors, Random Forest, Multilayer Perceprtons, KNN and Bagging etc.

Initially the dataset divided in two parts called training and testing data for the classification. The training data is provided as input to classification algorithm for learning the classifier and results are stored. For learned classifier testing data gives as input for classification, based on the parameters compared and stored results classifier classifies testing data. Classification methods [2] [3] like Nearest Neighbors, Random Forest, Bagging, KNN and Multilayer Perceprtons etc extract model to perform the classification on different datasets. In section 3, these classification methods are described in detail.

2. Literature Survey

Data mining technique are plays important role for derive required information from large scale medical data [4]. Working of various techniques are analyzed and identified on the available datasets from different areas. Accuracy of soft computing technique depends on the behavior, algorithm used and the domain of data set. Most of the time medical data mining focuses on pattern extraction. For improving the performance of medical dataset Amalgam model was developed. In A this model performance is increased by combining the pre-processing, KNN, [5] k-means. Some attributes values are missing in medical data set that affects on classification process. We are having one approach for solving the problem of missing values [6]. By preprocessing of dataset is we can improve the performance of mining techniques because it removing noisy data [5] [6]. For finding incorrect classification k-means algorithms is useful. By applying SSI, KD, RB and L measures on different medical data sets they analyzed that SSI produced maximum classification accuracy [6] [7]. Applying N fold cross validation was on dataset we can analyze the states of classifiers based on nature of dataset [8]. Hybrid approaches are also used for classifying the high scale data [9] [10] [11]. By using combining different data mining technique or ensemble technique quality and accuracy can be increased.

3. Classification Algorithm

Classification is a process of data analysis used for extracting a model for learning and making the classes of given data objects, based on that prediction will be made for objects whose class label is unknown. The classification is done in main tow steps: training data and testing data. Classification model also represented using mathematical modeling, random forest, KNN etc. Some methods of classification are described below.

3.1 Nearest Neighbors

Nearest neighbors is simple and statistical learning and easy to implementation. In detail for training phase it store presented data points and depend on that it predicates class for unknown data point in testing phase accordingly to some distance metric[12]. Example of Nearest Neighbor is shown in figure1. The distance metric used in nearest neighbor or in KNN calculated using simply Euclidean distance [13]. Euclidean distance is calculated using this formula

$$\text{Euclidean distance}\,(X, Y) = \sqrt{\sum_{i=1}^{k}(Xi - Yi)2}$$

Formula shows the Euclidean distance between point X and point Y.

3.2 KNN: K- Nearest Neighbors

K-Nearest Neighbors is data mining technique mostly used in classification problems because it is simple and gives faster result [14]. K nearest neighbor is instance based classification method and it belongs to lazy learning algorithms. It classifies the objects based on the similarity measures. The object that will get minimum distance from will be selected as member of that class. Classes for the relative neighboring objects are identified based on rules decided for classification.
Working of K nearest neighbor [15] algorithm where K is an integer and where k=1 means that algorithm is Nearest Neighbors algorithm.
1. For the unknown instance find nearby K training instances.
2. Form these K instances select the most commonly occurring class.

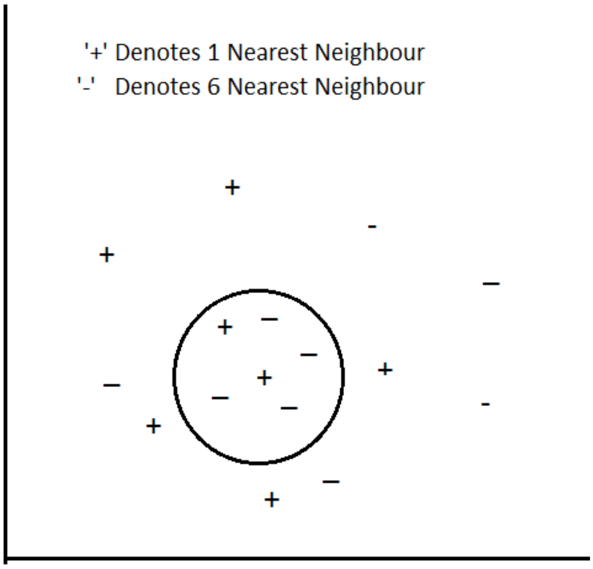
'+' Denotes 1 Nearest Neighbour
'-' Denotes 6 Nearest Neighbour

Figure 1: Nearest Neighbor and KNN

3.3 Random Forest Algorithm

Random Forest algorithm is ensemble machine learning technique and belongs from decision tree family [16]. In this technique data classification is done by constructing the forest of trees. In training phase it builds the forest of trees based on data and attributes. After that in testing phase it classifies data by voting [17] from all trees in that forest. Its gives quick results and computation as per number of inputs increased. Figure 2: shows working of Random Forest

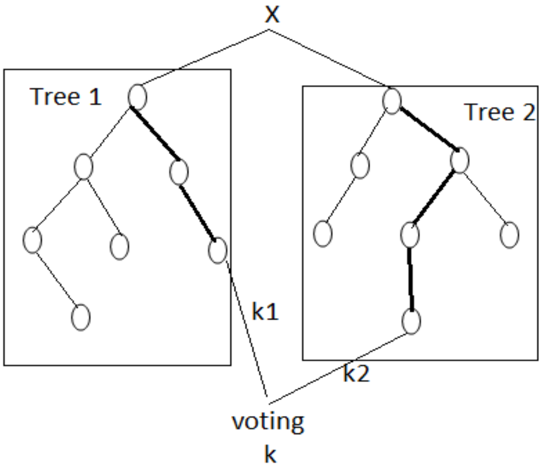

Figure 2: Working of Random Forest

3.4 Multi-layer Perceptrons

Multilayer Perceprtons classifier belongs from the Neural Network, it also known as feed forward neural network because it is extension of ANN model of classification. The ANN have only two layer for classification and number of layers are of extended to many layers but they are hidden .In training time of data weights at each layer id stored, based on the stored weights and the testing data final classification is done. Figure 3: shows Multilayer Perceprtons.

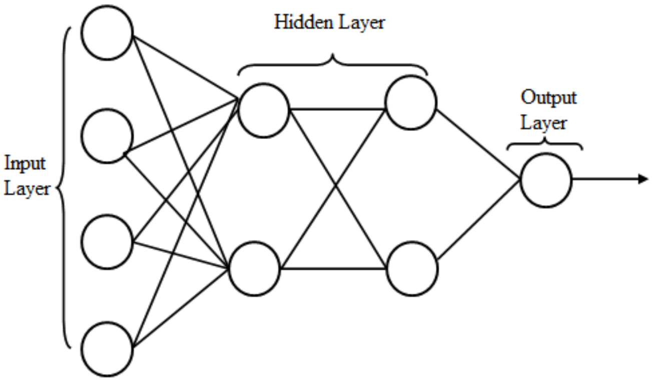

Figure 3: Multilayer Perceptrons

3.5 Bagging

Bagging is machine leaning classifier also called as Bootstrap aggregating, used for classification and feature selection purpose of data. In training it takes random subset for given dataset and predict classes and store result, Based on predications and by averaging or voting it make final classification in testing data.

4. Experimental Results and Analysis

For the four different datasets of medical data performance is carried out using five different classification algorithms. The classification algorithm used for the performance analysis is described in the previous section. The Current work is implemented using WEKA (Waikato Environment for Knowledge Analysis) tool for classification [19] and results taken for the different parameters namely: Accuracy, Execution Time and Mean Absolute Error. The results got in this experiment for different techniques are described in the form of table as well as graph.

4.1 Datasets

For this experiment, three medical datasets are taken into consideration, which are collected from the UCI Repository in ARFF format. These datasets are sufficient for classification process. Datasets are analyzed under different classification technique and parameters. All detailed information about datasets like instances and attributes are given in Table 1.

Table 1: Datasets

Sr.No.	Dataset	Number of Instances	Number of Attributes
1.	diabetes.arff	768	9
2.	contact-lenses.arff	24	5
3.	breast-cancer.arff	286	10
4.	liver-disorders.arff	345	7

Every dataset contain the different types of data like numbers, sex, text, age, dates and other domain data points. Every dataset needed to explore it explicitly because they having unique attributes and Attribute may be in have values in discrete form or continuous form. WEKA tool is used for the analysis of different datasets for classification under the different classification techniques. For this experiment 10 fold cross validation is adopted in work. WEKA is open source tool in java for data mining. WEKA itself providing number of inbuilt data mining technique, so different operation likes preprocessing, clustering can be performed easily. In this paper, we have used WEKA to perform the comparative study of different classification methods on medical datasets.

4.2 Accuracy

Accuracy of a classification algorithm or classifier is nothing but the number of correctly classified instances. Table 2 shows the accuracy of different classification algorithm. Accuracy of classifier is given by

$$Accuracy = \frac{Correctly\ Classified\ Objects}{Total\ Number\ of\ Objects} \times 100$$

Higher the accuracy, that classifier is more effective.

Table 2: Accuracy of Classification Algorithms

	Classification Algorithm				
Dataset	Nearest Neighbors	Bagging	Multilayer Perceptrons	Random Forest	KNN
diabetes.arff	70.5729 %	76.4323 %	75.9115 %	75.7813 %	77.0833 %
contact-lenses.arff	62.5 %	75 %	70.8333 %	79.1667 %	79.1667 %
breast-cancer.arff	65.7343 %	68.8811 %	64.6853 %	67.8322 %	72.3776 %
liver-disorders.arff	62.6087 %	65.5072 %	64.058 %	65.5072 %	67.5362 %

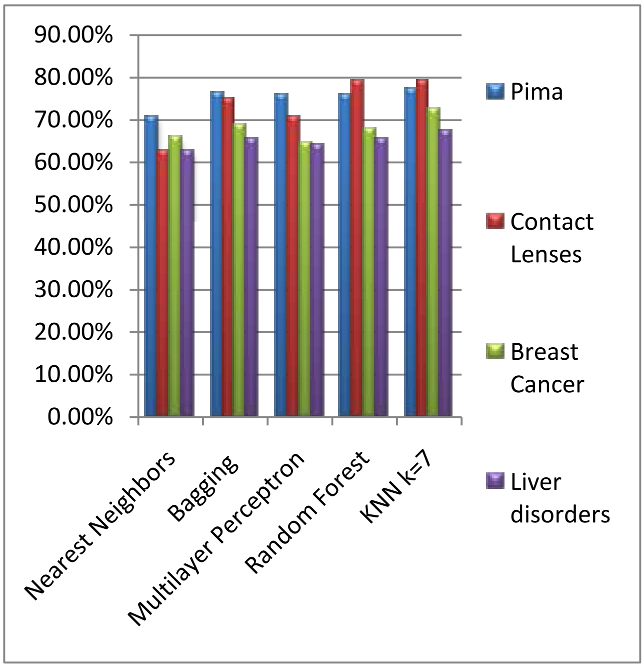

Figure 4: Graph of Accuracy

As per figure 4 is showing the comparison of accuracy for different algorithms. Figure shows that the KNN is most effective classifier for different datasets, because KNN gives the higher accuracy among all algorithms.

4.3 Execution Time

In the analysis of any algorithmic approach, the execution time is also one important parameter. In this current work, we have observed the execution time for identifying the efficient classification algorithm. Table 3 is showing the execution time of different classification algorithms.

Table 3: Execution Time of Classification Algorithms

	Classification Algorithm				
Dataset	Nearest Neighbors	Bagging	Multilayer Perceptrons	Random Forest	KNN
diabetes.arff	0.0	0.6	3.0	0.3	0.0
contactlenses.arff	0.0	0.0	0.6	0.2	0.0
breastcancer.arff	0.0	0.3	5.0	0.3	0.0
liver-disorders.arff	0.0	0.5	6.5	0.3	0.0

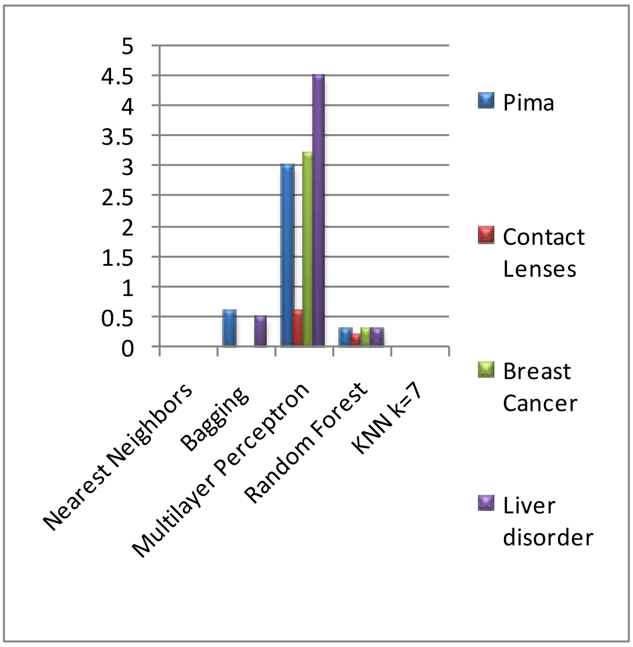

Figure 5: Execution Time of Different Algorithms

Figure 5, is showing the execution time of different classification algorithms. As figure shows KNN is efficient classification algorithm and as per results the Multilayer Perceptron algorithm is worst among the all other algorithms.

4.4 Mean Absolute Error (MAE)

MAE actually shows the efficiency capability of an algorithm. If the value of MAE is less, then classifier is efficient and capable. Table 4, is shows the mean absolute error of different algorithms in classification

Table 4: Mean Absolute Error for Classification Algorithms

	Classification Algorithm				
Dataset	Nearest Neighbors	Bagging	Multilayer Perceptrons	Random Forest	KNN
diabetes.arff	0.2934	0.3089	0.283	0.2913	0.2819
contactlenses. arff	0.25	0.2602	0.2072	0.2087	0.22
breastcancer. arff	0.3427	0.3806	0.3552	0.3632	0.3257
liver-disorders.arff	0.3739	0.4379	0.3648	0.3995	0.3211

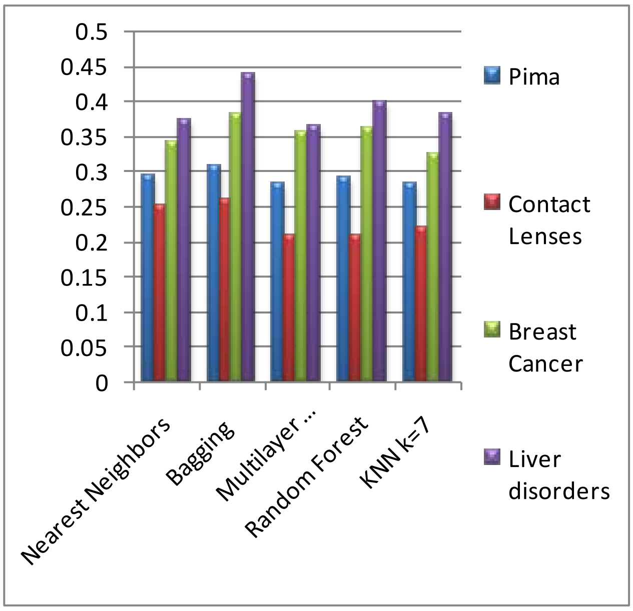

Figure 6: Mean Absolute Error of classification Algorithm

Figure 6, is showing the mean absolute error of different algorithms. As the MAE value in case of KNN algorithm is minimum that shows the accurate prediction capability of the algorithm and as per results the bagging algorithm is worst.

5. Conclusion

In this paper, the analysis the performance of different classification algorithms are Nearest Neighbors, Random Forest, Bagging, Multilayer Perceptrons and KNN based on three different parameters called accuracy, mean absolute error, execution time and analysis for four different medical datasets. The results we got in analysis show that the KNN is the efficient classification algorithm and Bagging is the worst algorithm for the classification.

References

[1] Gupta, M., and Aggarwal, N. “Performance Analysis of Classification Techniques on XML Dataset”, International Journal of Computer Science and Technology IJCST Vol. 1, Issue 1, pp. 76-79, 2010.

[2] Justin, T., Gajsek, R., Struc, V., and Dobrisek, S., “Comparison of Different Classification Methods for Emotion Recognition”, MIPRO 2010, Opatija, Croatia, pp. 700-703, 2010.

[3] Gupta, S., Kumar, D., and Sharma, A., “Data Mining Classification Techniques applied for Breast Cancer Diagnosis and Prognosis”, Indian Journal of Computer Science and Engineering (IJCSE) Vol. 2 No. 2, pp. 188-195, 2011

[4] Jacob S.G. , Ramani R.G, “ Mining of Classification Patterns in Clinical Data through Data Mining Algorithms”, ICACCI’12 –ACM 978-1-4503-1196-0/12/08

[5] NirmalaDevi M, Balamurugan S, Swathi U V, “ An amalgam KNN to predict Diabetes Mellitus”, 2013 IEEE International Conference on Emerging Trends in Computing, Communication and Nanotechnology

[6] Davis D N, Zhang Y, Kambhampati C, Goode K, Cleland J.G.F, “ A Comparative study of missing value imputation with multi class classification for clinical heart failure data”, 2012 IEEE, 9th International Conference on Fuzzy Systems and Knowledge Discovery.

[7] Saastamoinen K, Ketola J, “Medical Data Classification using Logical Similarity based Measures”, 1-4244-0023-6/06 2006 IEEE.

[8] Aslandogan Y.A, Mahajani G. A, “Evidence Combination in Data Mining”, 2004 IEEE Proceedings of the International Conference on Information Technology: Coding and Computing.

[9] Kumar S U, Inbarani H, Senthil Kumar, “ Bijective Soft Set Based Classification of Medical Data”, 2013 IEEE, International Conference on Pattern Recognition, Informatics and Mobile Engineering.

[10] Hassan S. Z, Verma B, “A Hybrid Data Mining Approach for Knowledge Extraction and Classification in Medical Databases”, 2007 IEEE 7th International Conference on Intelligent Systems Design and Applications.

[11] Michelakos I, Papageorgiou E, Vasilakopoulos M, “ A hybrid classification algorithm evaluated on medical data”, 2010 IEEE Workshops on Enabling Technologies: Infrastructure for Collaborative Enterprises.

[12] T. Darrell and P. Indyk and G. Shakhnarovich (2006). Nearest Neighbor Methods in Learning and Vision: Theory and Practice. MIT Press.

[13] K. ROY+, C.” Comparison of the Multi Layer Perceptron and the Nearest Neighbor Classifier for Handwritten Numeral Recognition” JOURNAL OF INFORMATION SCIENCE AND ENGINEERING 21, 1247-1259 (2005)

[14] F. Moreno-Seco, L. Mico, and J. A. Oncina, "Modification of the LAESA Algorithm for Approximated k-NN Classification," *Pattern Recognition Letters*, pp. 47–53, 2003.

[15] M. Akhil jabbar, “Classification of Heart Disease Using K- Nearest Neighbor and Genetic Algorithm”, International Conference on Computational Intelligence: Modeling Techniques and Applications (CIMTA) 2013.

[16] Leo Breiman, “Random Forests”, Machine Learning, 45, 5–32, 2001 © 2001 Kluwer Academic Publishers.

[17] Dr. B. G. Prasad,” Random Forest Based Classification of CT scan Brain Images using Statistical Texture Features”, International Conference on Computational Techniques and Artificial Intelligence (ICCTAI'2011).

[18] Luís M. Silva,” Data classification with multilayer perceptrons using a generalized error function”, Neural Networks 21 (2008) 1302_1310

[19] Pushpa, “Comparision of Clustering Techniques using WEKA”, M. Tech. Thesis, Guru Jambheshwar University of Science and Technology, Hisar, India,2010

[20] Desai, A., and Rai., S., “Analysis of Machine Learning Algorithms using WEKA”, International Conference & Workshop on Recent Trends in Technology, (TCET) 2012 Proceedings published in International Journal of Computer Applications (IJCA) 27, pp.27-32, 2012.

Author Profile

Mr. Mangesh Metkari received the B.E. degrees in Information Technology from Padmabhushan Vasantraodada Patil Institute of Technology, Budhgon Sangli in 2012. Now student of M.E. (Computer Engineering) at AISSMS’s CoE from Savitribai Phule University of Pune.

Mrs. Madhavi Pradhan is an Assistant Professor at AISSMS’s CoE, Pune. She received M.Tech (Software Engineering) from S.J.C.E. Mysore and B.E. (Computer Engineering) from Government College of Engineering, Amravati.