



Performance Analysis of Machine Learning Techniques in Micro Array Data Classification

Sankhadeep Roy¹, Anjali Mohapatra²

¹ Research Scholar, ² Assistant Professor, Department of Computer Science Engineering, IIIT, Bhubaneswar, Malipada, Orissa. Pin-751003, INDIA

Abstract: The development of data-mining applications such as classification has shown the need for Machine Learning algorithms to be applied on large scale data. This paper presents the comparison of different classification techniques and investigates the performance of different classifiers for a set of Micro Array data. The algorithm or methods tested are Naive Bayes Classifier, Support Vector Machines and an ensemble SVM-KNN model. The study demonstrates the advantage of the new method and provides insights on the preferable combinations of classifier and feature selection technique. The evaluation results indicates better classification accuracy with reduced training time and implementation complexity compared to earlier implemented models.

Keywords: Machine Learning, Data Mining, Naive Bayes, Support Vector Machine (SVM), Nearest Neighbour.

I. Introduction

A major problem in bioinformatics analysis is attaining the correct diagnosis of certain important information such as gene selection, protein structure prediction and Microarray data classification. The key challenge in recent years is the classification of samples into categories such as individuals who carry some illness and others who do not. This is done by learning how to classify, based on a training set containing labelled samples from the two populations, and then predicting the label of new samples. The use of gene expression Micro Arrays, a high throughput technology allows simultaneous measuring of expression levels of thousands or tens of thousands of genes on a silicon chip. In the past few years gene expression data resulting from microarray technology is extensively used in clustering and classification of human diseases [16]. Accurate classification of these diseases is of great importance for diagnosis and managing treatment of its patients. Hence the objective of this study is to attain a classification system, capable of classifying different types of diseases using different microarray datasets.

In medicine, some machine learning techniques such as neural network, decision tree and naive bayes, support vector machine are actively used for meeting the high accuracy criteria for classification. These techniques are used as base classifier but in ensemble methods the classifiers are combined to maximize the results of the base classifiers. Diverse classifiers make different errors on different samples. Combination of such classifiers might lead to more accurate decisions. Keeping in mind the unique characteristics of ensemble classifiers the paper proposes a novel ensemble classification technique to improve the classification accuracy. Here Naive Bayes classifier, Support Vector Machine algorithm with different kernel functions is used as base classifier and a SVM-KNN combined classifying method ensemble for classification of various diseases. Each one of these techniques is validated and evaluated on three different microarray datasets.

II. Methods

A. Data Pre-processing

Data pre-processing is a data mining technique that involves transforming raw data into an understandable format. Real-world data is often incomplete, inconsistent, and/or lacking in certain behaviours or trends, and is likely to contain many errors. Data pre-processing is a proven method of resolving such issues preparing raw data for further processing. Once in a while one has numeric data but wants to use a classifier that handles only nominal values. In that case one needs to discretize the data. The number of values for a given continuous attribute is reduced by dividing the range of the attribute into intervals. Interval labels can then be used to replace actual data values. Also feature Subset Selection allows us to investigate which (subsets of) attributes are the most predictive and useful ones, finding a feature subset that is a good substitute to all features resulting in better accuracy especially on new data[7].

B. Classifiers

Classification is the problem of identifying to which of a set of categories (sub-populations) a new observation belongs, on the basis of a training set of data containing observations (or instances) whose category membership

is known. An algorithm that implements classification, especially in a concrete implementation, is known as a classifier.

1. *Naive Bayes Classifier*

The Naive Bayesian classifier uses the naive Bayes formula to calculate the probability of each class given the values of all the attributes and assuming the conditional independence of the attributes[12]. The attributes are usually defined by a human (often in medicine), and are therefore relatively independent. The independence assumptions often do not have an impact on reality. Therefore they are considered as naive. This is the reason why the naive Bayesian formula often performs well on real-world problems.

2. *Support Vector Machine (SVM)*

Support vector machine is a machine learning algorithm used for classification and regression analysis. SVM map the problem in high dimensional space. It is a kernel based technique and linearity and non-linearity of the classifier is based on the selection of the kernel [18]. Support Vector Machine uses different kernel functions, polynomial kernel (SVM-P), radial basis kernel (SVM-RBF), and linear kernel function (SVM-L) to classify the various diseases because SVM scales the problems in to high dimensional space to divide the data between two classes correctly.

3. *Nearest Neighbours Algorithm*

Nearest neighbours algorithm is considered as statistical learning algorithm and it is extremely simple to implement and leaves itself open to a wide variety of variations. In brief, the training portion of nearest-neighbour does little more than store the data points presented to it. When asked to make a prediction about an unknown point, the nearest neighbour classifier finds the closest training-point to the unknown point and predicts the category of that training point accordingly to some distance metric. The distance metric used in nearest neighbour methods for numerical attributes can be simple Euclidean distance [3].

4. *Ensemble of classifiers*

Ensemble classifier constructs to combine the decision of the individual classifiers in such a way to make a final decision more accurate. Combining a number of trained classifiers generates better performance than any single classifier because errors produce by one classifier may be correctly classified by other classifier[5]. Support vector machine (SVM) and K-Nearest Neighbour (KNN) ensemble classifier is a combined classifying method, having an excellent performance on various Microarray datasets[13]. In classification phase, Support vector machine classifier is used with different kernels: Linear, Polynomial and Radial Basis Kernel. The classification performance of SVM-KNN classifier is evaluated and compared to the one that obtained by support vector machine and Naive Bayes. Experimental results show that SVM-KNN model has achieved a remarkable performance with better classification accuracy on testing subset.

III. Classifier testing and evaluation

A. *Cross validation*

A classifier usually learns from the available data. The problem is that the resulting classifier may fit on the training data, but might fail to predict unseen data. Cross validation is a technique for assessing the generalization performance of a given classifier. It can be used for estimating the performance of a given classifier as well as for tuning the model parameters. In K-Fold Cross Validation, the available data is divided into k equally sized folds. Subsequently, k iterations of training and validation are performed such that, within each iteration a different fold of the data is held-out for validation while the remaining (k-1) folds are used for training the classification model. Data is usually stratified prior to being split into k folds i.e. data is rearranged to ensure that each fold contains instances of all the classes in the problem at hand[15].

IV. Datasets

Two different datasets have been used in the investigation namely Diabetes and Heart Statlog selected from University of California Irvine (UCI) Machine Learning Repository [2].

Table 1:Dataset Description

Dataset	Instances	Attributes	Missing	Classes
PIMA Indian Diabetes	768	13	No	2
Heart Statlog	270	9	No	2

V. Results

Various different classification techniques have been applied on the two different healthcare datasets taken from UCI Machine Learning Repository. In this study WEKA, Machine learning tool [17] for data mining is used to achieve the proposed objectives. The percentage of accuracy rate and the value of Kappa Statistic for classification techniques are used as the measurement parameters for analysis. A high value of accuracy rate for a classification technique applied on a dataset shows that the dataset is correctly classified by the obtained classifier. Kappa statistic is used to assess the accuracy of any particular measuring cases, it is usual to distinguish between the reliability of the data collected and their validity. When Kappa statistics K equals 1, it means that there is complete agreement between the classifier output and the expected output (real world output). Kappa is always less than or equal to 1. A value less than 1 implies less than perfect agreement between the classifier output and the real world expected output. [11]

The results are shown as below:

Table 2: Results of Naive Bayes Classifier obtained in Weka

Dataset Name	Cross Validation	Kappa Statistic	Accuracy
Pima Indian Diabetes	3 Fold	0.5153	77.9948 %
Pima Indian Diabetes	5 Fold	0.5161	77.9952 %
Pima Indian Diabetes	10 Fold	0.5128	77.8646 %
Heart Statlog	3 Fold	0.699	85.1852 %
Heart Statlog	5 Fold	0.6772	84.0741 %
Heart Statlog	10 Fold	0.6611	83.3333 %

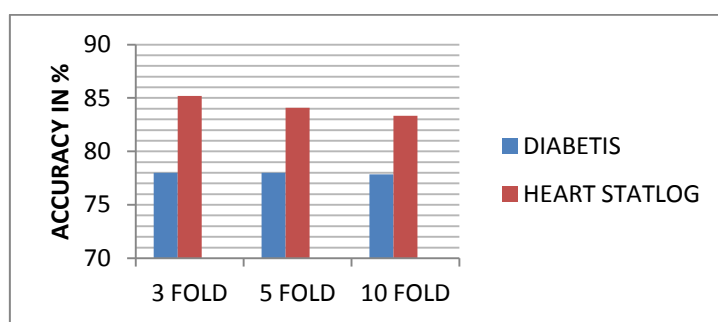


Figure 1: Results of Naive Bayes Classifier

Table 3: Results of SVM obtained in Weka

Dataset Name	Kernel type	Kappa Statistic	Cross Validation Accuracy
PIMA Indian Diabetes	Linear	0.5267	79.1667 %
PIMA Indian Diabetes	Polynomial	0.5526	80.7292 %
PIMA Indian Diabetes	RBF	0.5777	81.5104%
Heart Statlog	Linear	0.6917	84.8148 %
Heart Statlog	Polynomial	0.7213	86.2963%
Heart Statlog	RBF	0.8492	92.5926 %

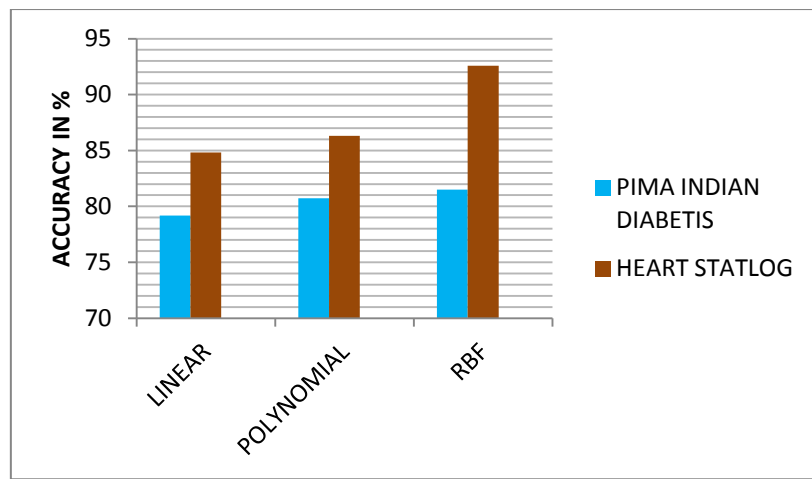


Figure 2: Results of SVM

Table 4: Results of SVM-KNN Ensemble Classifier obtained in Weka

Dataset Name	Kernel type	Kappa Statistic	Cross Validation Accuracy
PIMA Indian Diabetes	Linear	0.5506	80.3385 %
PIMA Indian Diabetes	Polynomial	0.5587	80.9896 %
PIMA Indian Diabetes	RBF	0.5785	81.5104 %
Heart Statlog	Linear	0.8497	92.5921 %
Heart Statlog	Polynomial	0.872	93.7037 %
Heart Statlog	RBF	0.887	94.4444 %

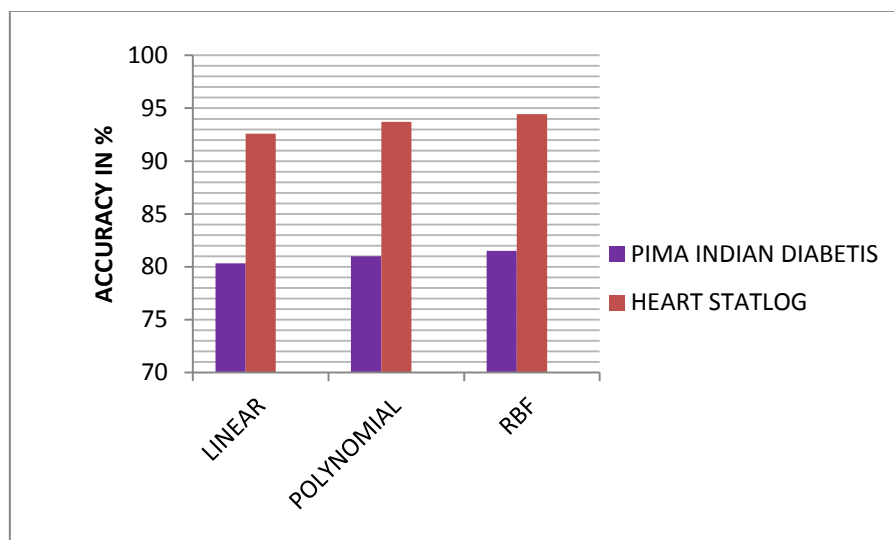
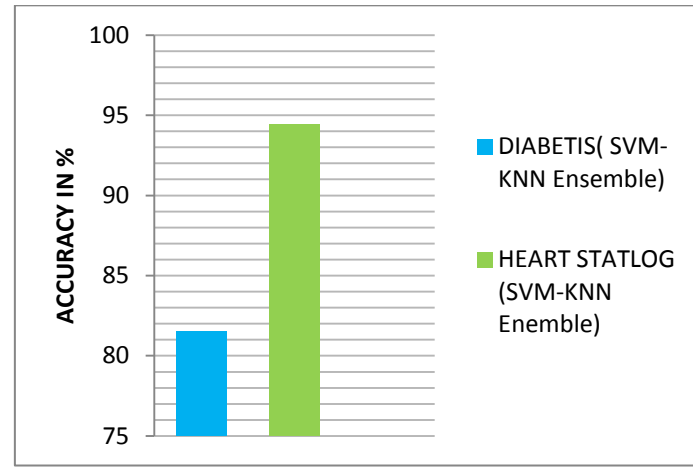


Figure 3: Results of SVM-KNN Ensemble Classifier

Table 5: Results Showing the Best classification techniques over given datasets

Datasets Used	Technique Applied	Accuracy Rate	Tools Used
Pima Indian Diabetes	SVM_KNN Ensemble	81.5104 %	Weka,LibSVM[4]
Heart Statlog	SVM_KNN Ensemble	94.4444 %	Weka,LibSVM[4]

**Fig 4: Best case of Classification Accuracy on the two datasets**

VI. Discussion

Based on the above Figures 1, 2, 3 and Tables 2, 3, 4, we can clearly see that the highest accuracy achieved is the for Heart Statlog dataset. Considerable performance has also been achieved on Diabetes dataset by using Ensemble based optimizing technique. From Table 4, we can see that SVM-KNN Ensemble model has a higher diagnosis accuracy of 94.444% for the Heart Statlog dataset compared to that of 92.5926 % for SVM and 85.1852 % for the Naive Bayes model. Also the SVM and SVM-KNN Ensemble model has a higher diagnosis accuracy of 81.5104% for the Diabetes dataset compared to that of 77.9952 % for the Naive Bayes model respectively. In fact, the highest accuracy belongs to the SVM_KNN Ensemble Classifier, followed by SVM Radial basis function and subsequently the Naïve Bayes Classifier. The declining of classification accuracy was of unbalanced dataset was solved by data discretization and feature selection techniques. Results of experiments show that the classification of RBF kernel function is better than others. Kappa statistics K value for SVM-KNN Ensemble Classifier model with feature selection is equal to 0.887 which clearly illustrates the match between the Ensemble classifier and the real world output.

VII. Conclusion and Future work

The experimental results have shown that different classification techniques behave differently on different datasets depending on the nature of their attributes and size. The classification technique which has shown the highest accuracy rate over a dataset has been selected as the best classification technique for that dataset. The best algorithm based on the breast cancer data is SVM-KNN Ensemble classifier with an accuracy of 94.444%.The proposed classification model exploits the use of powerful machine learning models such as SVMs and ensemble methods coupled with feature subset selection. These results suggest that among the machine learning algorithm tested, SVM-KNN algorithm Ensemble has the potential to significantly improve the conventional classification methods for use in medical or in general, bioinformatics field. For future work it is suggested varied techniques can be investigated for distributing the features among subsets. Higher numbers of base classifiers / numbers of feature subsets can be experimented with and the time complexity of the proposed models needs to be calculated. The uses of other ensemble classifiers are worth investigating. In addition, the use of the proposed models can be extended to other data sets and other domains in the bioinformatics field using different parameters and techniques.

Reference

- [1] E. Alpaydin, Introduction to Machine Learning, The MIT Press London, England, (2010)
- [2] K. Bache, and M. Lichman,, UCI Machine Learning Repository [<http://archive.ics.uci.edu/ml>]. Irvine, CA: University of California, School of Information and Computer Science, 2013.
- [3] Bin Othman, Mohd Fauzi, and Thomas Moh Shan Yau. "Comparison of different classification techniques using WEKA for breast cancer." 3rd Kuala Lumpur International Conference on Biomedical Engineering, 2006. Springer Berlin Heidelberg, 2007.
- [4] C.C. Chang and C.J. Lin, LIBSVM: a library for support vector machines, 2001.
- [5] B. Fida, M. Nazir, N. Naveed,.; S. Akram,, "Heart disease classification ensemble optimization using Genetic algorithm," Multitopic Conference (INMIC), IEEE 14th International , vol., no., pp.19,24, 22-24 Dec. 2011
- [6] S. Gupta, D. Kumar,, and A. Sharma, "Performance Analysis of Various Data Mining Classification Techniques on Healthcare Data." International Journal of Computer Science & Information Technology (IJCSIT), 3(4). 2011.
- [7] M. Gutkin,. Feature selection methods for classification of gene expression profiles. Tel-Aviv University. (2008)
- [8] H.L. Yu, G.C. Gu, H.B. Liu, and J. Shen, "Feature Subspace Ensemble Classifiers for Microarray Data", ICIC Express Letters, vol.4, no.1, pp.143-148, 2011.
- [9] T. Helmy,.; .Rasheed, "Multi-category bioinformatics dataset classification using extreme learning machine," Evolutionary Computation, 2009. CEC '09. IEEE Congress on , vol., no., pp.3234,3240, 18-21 May 2009
- [10] I. Guyon, J. Weston, S. Barnhill and V. Vapnik. "Gene selection for cancer classification using support vector machines. Machine Learning." 46(1-3):389-422, 2002.
- [11] Karegowda, Asha Gowda, M. A. Jayaram, and A. S. Manjunath. "Cascading K-means Clustering and K-Nearest Neighbor Classifier for Categorization of Diabetic Patients."International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249 – 8958, Volume-1, Issue-3, February 2012
- [12] M. Kukar, C. Groselj, I. Kononenko and J.J. Fettich. "An application of machine learning in the diagnosis of ischaemic heart disease," Computer-Based Medical Systems. 1997. Proceedings., Tenth IEEE Symposium on, vol., no., pp.70-75, 11-1,3 Jun 1997
- [13] Li Rong and Sun Yuan, "Diagnosis of Breast Tumor Using SVM-KNN Classifier," Intelligent Systems (GCIS), Second WRI Global Congress on , vol.3, no., pp.95,97, 16-17 Dec. 2010 doi: 10.1109/GCIS.2010.278
- [14] Manaswini Pradhan and Dr. Ranjit Kumar Sahu, "Predict the onset of diabetes disease using Artificial Neural Network (ANN)", International Journal of Computer Science & Emerging Technologies, pp.303 -311, vol. 2, iss. 2, 2011.
- [15] Neamat El Gayar, Eman Ahmed and ImanEl Azab, "Novel Machine Learning Techniques for Micro-Array Data Classification, Bioinformatics -Trends and Methodologies", MahmoodA. Mahdavi (Ed.), 2011
- [16] D.A Salem, R.A.A.A Abul Seoud and H.A Ali, "K5. Merging genetic algorithm with different classifiers for cancer classification using microarrays," Radio Science Conference (NRSC), 29th National, vol., no., pp.659-666, 10-12, April 2012.
- [17] WEKA at <http://www.cs.waikato.ac.nz/~ml/weka>.
- [18] Yan Zhang, Fugui Liu, Zhigang Zhao, Dandan Li, Xiaoyan Zhou and Jingyuan Wang, "Studies on Application of Support Vector Machine in Diagnose of Coronary Heart Disease," Electromagnetic Field Problems and Applications (ICEF), Sixth International Conference on , vol., no., pp.1,4, 19-2,1 June 2012.