Performance Evaluation of Machine Learning Techniques for Medical Data Mining

Aakshi Mahajan
Department of Computer Science and Engineering
Amritsar College of Engineering & Technology,
Amritsar

Ajay Sharma
Department of Computer Science and Engineering
Amritsar College of Engineering & Technolgy,
Amritsar

Abstract-Machine learning based techniques are becoming popular day by day due to its wide range of applications. Due to great change in environmental conditions, many people are suffering from various kinds of diseases. This paper has considered different machine learning algorithms to efficiently mine the medical data set. By implementing various popular machine learning techniques in MATLAB environment, it has been found that the LAD-Tree outperforms over the J48 and random forest based machine learning algorithms.

Keywords: J48, LAD Tree, Machine Learning, MATLAB

I. INTRODUCTION

A. Data Mining

Stages of Data Mining

Data mining is carried out in different stages:

Exploration

This stage starts with data preparation which may involve cleaning data, data transformations, selecting subsets of records. This first stage of the process of data mining may involve anywhere between a simple choice of straightforward predictors for a regression model, to elaborate exploratory analyses using a wide variety of graphical and statistical methods in order to identify the most relevant variables and determine the complexity and the general nature of models that can be taken into account in the next stage.

Model building and validation

This stage involves considering various models and choosing the best one based on their predictive performance. There are a variety of techniques developed to achieve that goal - many of which are based on so-called "competitive evaluation of models," that is, applying different models to the same data set and then comparing their performance to choose the best

Deployment

This final stage involves using the model selected as best in the previous stage and applying it to new data in order to generate predictions or estimates of the expected outcome.

The key properties of data mining are:

- Automatic discovery of patterns
- Prediction of likely outcomes
- Creation of actionable information

Amritsar College of Engineering & Technolgy,
Amritsar

Data Mining Process-

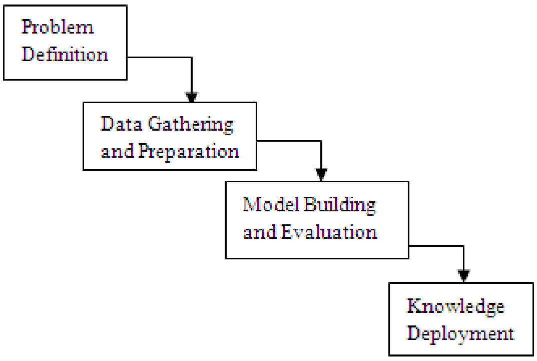

Figure 1: Data Mining Process

Problem Definition

This phase of data mining project focuses on understanding the project objectives and requirements.

Data Gathering and Preparation

The data understanding phase involves data collection and exploration. The data preparation phase covers all the tasks involved in creating the tables which will be used to build the model.

Model Building and Evaluation

In this phase, you select and apply various modeling techniques and evaluate the parameters to optimal values. This stage determines that how well the model satisfies the originally stated business goal

Knowledge Deployment

Knowledge deployment is the use of data mining within a target environment. In the deployment phase, insight and actionable information can be derived from data.

B.Multiclass SVM

Support Vector Machines are the supervised learning models with associated learning algorithms which are used to analyze the data which is used for classification as well as regression purposes. Along with performing *linear classification, SVM can also perform non linear* classification.

Linear SVM performs the classification of the data set into two classes but the Multiclass SVM can be used to perform the classification of the data set into more than two classes due to which it is widely used in many applications. The success of SVMs solving classification tasks in a wide variety of fields, such as text or image processing and medical informatics, has stimulated practitioners to do research on the execution performance and scalability of the training phase of serial versions of the algorithm

Applications of SVM

- SVMs are useful for text and hypertext categorization.
- SVMs are also used to perform classification of images.
- They are used to recognize the hand written characters.
- The SVM algorithms are also applied in biological and other sciences.

II. EXISTING TECHNIQUES

A. Bayes Net Classifier

In machine learning, Bayes Net classifiers are a family of simple probabilistic classifiers based on applying Bayes' theorem with strong (naive) independence assumptions between the features. These classifiers are highly scalable, requiring a number of parameters linear in the number of variables in a learning problem.

B. Logistic Classifier

The logistic function is useful because it can take input as any value from negative to positive infinity and output gives value between zero and one which can be interpreted as probability. The Logistic function is given as

$$F(x) = \frac{et}{et+1} \qquad (1)$$

Where t is a linear function of variable x.

C. Random Forest Tree

A random forest is a classifier which consists of many decision trees and outputs the class which is made up of the classes that are made by individual trees. There are many different advantages of using Random Forest Tree some of which are:

- It can handle large no. of input variables.
- For many different data sets, it produces maximum accuracy.
- It is helpful in balancing error of unbalanced data sets.
- It is easy to learn.
- It is useful in classification, clustering and visualizing the data.

D. J48 Decision Tree

J48 Decision Tree is used for te dataset which have list of predictors or independent variables and a list of target or dependent variables. Decision Tree J48 is the implementation of the algorithm ID3 (Iterative Dichotomiser 3) developed by WEKA project team. In J48 decision tree, the classification is done in two classes as Recurrence events and Non-Recurrence Events.

E. Zero R Classifier

Zero R is the simplest classification which is entirely based on the targets and ignores all the predictors. It only predicts the majority class. It is just used for determining the baseline performance as the benchmark for other classification methods.

F. Multilayer Perceptron Classifier

Multilayer Perceptron classifier is a feed forward artificial neural network in which the nodes of the input layer are connected to the nodes of the output layer with the help of hidden layer between them.

In this classifier, except for the input nodes, each node is a neuron with the non-linear activation function.

III. PERFORMANCE METRICS

The different measures which were used to evaluate the performance of classifiers are described below:

Correctly Classified Instances

It depicts the total of instances which are correctly classified among the total no. of instances being taken into consideration.

Incorrectly Classified Instances

It determines the no. of instances which are not correctly classified amog the total no. of instances.

Accuracy

Accuracy is a measure of how well the model correlates an outcome with the attributes in the data that has been provided. Accuracy is calculated as:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \qquad (2)$$

Kappa Statistics

This measure is used to evaluate the agreement between two different qualitative items. Cohen's kappa measures the agreement between two rates which each classify *N* items into *C* mutually exclusive categories. It is calculated as:

$$K = \frac{po-pe}{1-pe} \qquad (3)$$

where p_o is Observed accuracy and p_e is the expected accuracy.

F-Measure

The measure that combines precision and recall is called as F-measure. It is the harmonic mean of precision and recall.

$$F - Measure = \frac{2.Precision.Recall}{Precision+Recall} \qquad (4)$$

where Precision is the fraction of retrieved instances that are relevant and Recall is the fraction of relevant instances that are retrieved.

Confusion Matrix

A confusion matrix contains information about actual and predicted classifications done by a classification system. Performance of such systems is commonly evaluated using the data in the matrix. The following table shows the confusion matrix for a two class classifier.

The entries in the confusion matrix have the following meaning in the context of our study:

- *a* is the number of **correct** predictions that an instance is **negative**,
- *b* is the number of **incorrect** predictions that an instance is **positive**,

- *c* is the number of **incorrect** of predictions that an instance **negative**, and
- *d* is the number of **correct** predictions that an instance is **positive**.

	Predictive Negative	Predictive Positive
Actual Negative	A	B
Actual Positive	C	D

Figure 2: Confusion Matrix

Several standard terms have been defined for the 2 class matrix:

- The *accuracy* (*AC*) is the proportion of the total number of predictions that were correct. It is determined using the equation:

$$AC = \frac{a+d}{a+b+c+d} \qquad (5)$$

- The *recall* or *true positive rate* (*TP*) is the proportion of positive cases that were correctly identified, as calculated using the equation:

$$TP = \frac{d}{c+d} \qquad (6)$$

- The *false positive rate* (*FP*) is the proportion of negatives cases that were incorrectly classified as positive, as calculated using the equation:

$$FP = \frac{b}{a+b} \qquad (7)$$

- The *true negative rate* (*TN*) is defined as the proportion of negatives cases that were classified correctly, as calculated using the equation:

$$TN = \frac{a}{a+b} \qquad (8)$$

- The *false negative rate* (*FN*) is the proportion of positives cases that were incorrectly classified as negative, as calculated using the equation:

$$FN = \frac{c}{c+d} \qquad (9)$$

- Finally, *precision* (*P*) is the proportion of the predicted positive cases that were correct, as calculated using the equation

$$P = \frac{d}{b+d} \qquad (10)$$

IV. EXPERIMENTAL RESULTS

WEKA

Weka is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset. Weka contains tools for data preprocessing, classification, regression, clustering, association, rules, and visualization. It is also well suited for developing new machine learning schemes.

The different measures are tabulated below for Wisconsin Breast cancer, Diabetes, E-colii and Heart Disease. The following subsections give the detailed discussion of results obtained for the mentioned data sets under experimental work.

A. Discussion for Wisconsin Breast Cancer

Table 1 The detailed accuracy for Breast Cancer for different classifiers.

	Bayes Net	Logistic	Multilayer Perceptron	Multiclass Classifier	Classification via Regression	Decision Table	Zero R	J48	Random Forest	REP Tree
CCI	217	197	72	197	206	66	64	66	65	202
ICI	69	89	25	89	80	31	33	31	32	84
KS	0.3958	0.1979	0.3441	0.1979	0.2488	0.1435	0	0.2001	0.1674	0.1601
TP Rate	0.759	0.689	0.742	0.689	0.72	0.68	0.66	0.68	0.67	0.706
FP Rate	0.38	0.505	0.441	0.505	0.499	0.561	0.66	0.562	0.522	0.572
Precision	0.75	0.668	0.744	0.668	0.697	0.659	0.435	0.657	0.642	0.669
Recall	0.759	0.689	0.742	0.689	0.72	0.68	0.66	0.68	0.67	0.706
F-Measure	0.753	0.675	0.713	0.675	0.698	0.619	0.525	0.65	0.636	0.664
ROC Area	0.759	0.646	0.589	0.646	0.661	0.717	0.5	0.603	0.644	0.621

The accuracy calculated from the above tabulated measures is 75.8741%,76.2238%,96.5035%,76.2238%,75.8741%,78.6713%,70.2797%,75.8741%,97.9021%,74.1259% for Bayes Net, Logistic, Multilayer Perceptron, Multilayer Classification, Classification via Regression, Decision Table, Zero R,J48, Random Forest, LAD Tree respectively for Wisconsin Breast Cancer data set. It is evident from the above analysis that accuracy of Random Forest tree algorithm is maximum.

B.Discussion for Diabetes

Table 2 The detailed accuracy for Diabetes for different classifiers

	Bayes Net	Logistic	Multilayer Perceptron	Multilayer Classification	Classification via Regression	Decision Table	Zero R	J48	Random Forest	REP Tree
CCI	571	593	579	593	589	214	178	199	189	578
ICI	197	175	189	175	179	47	**83**	62	72	190
KS	0.429	0.4734	0.4484	0.4734	0.4565	0.5608	0	0.4342	0.3951	0.438
TP Rate	0.743	0.772	0.754	0.772	0.767	0.82	0.682	0.762	0.724	0.753
FP Rate	0.319	0.321	0.314	0.321	0.336	0.29	0.682	0.342	0.309	0.328
Precision	0.741	0.767	0.75	0.767	0.761	0.816	0.465	0.756	0.741	0.747
Recall	0.743	0.772	0.754	0.772	0.767	0.82	0.682	0.762	0.724	0.753
F-Measure	0.742	0.765	0.751	0.765	0.758	0.814	0.553	0.758	0.73	0.748
ROC Area	0.806	0.832	0.793	0.832	0.827	0.847	0.5	0.796	0.728	0.766

The accuracy calculated from the above tabulated measures is74.349%,77.2135%,75.3906%,77.2135%,76.6927%, 81.9923%,68.1992%,76.2452%,77.0115%,75.4789% for BayesNet, Logistic, Multilayer Perceptron, Multilayer Classification, Classification via Regression, Decision Table, Zero R,J48,Random Forest, LAD Tree respectively for Diabetes data set. It is evident from the above analysis that accuracy of Decision Table algorithm is maximum.

C.. Discussion for E-coli

Table 3 The detailed accuracy for E-coli disease for different classifiers

	Bayes Net	Logistic	Multilayer Perceptron	Multiclass Classifier	Classification via Regression	Decision Table	Zero R	J48	Random Forest	REP Tree
CCI	293	291	98	94	95	89	44	90	97	274
ICI	43	45	16	20	19	25	70	24	17	62
KS	0.8236	0.8152	0.8123	0.7666	0.7764	0.6996	0	0.7185	0.8	0.7428
TP Rate	0.872	0.866	0.86	0.825	0.833	0.781	0.386	0.789	0.851	0.815
FP Rate	0.034	0.036	0.038	0.043	0.045	0.074	0.386	0.071	0.04	0.056
Precision	0.864	0.86	0.859	0.822	0.828	0.782	0.149	0.794	0.863	0.796
Recall	0.872	0.866	0.86	0.825	0.833	0.781	0.386	0.789	0.851	0.815
F-Measure	0.863	0.862	0.855	0.82	0.819	0.747	0.215	0.788	0.844	0.805
ROC Area	0.983	0.957	0.969	0.96	0.961	0.951	0.5	0.88	0.956	0.926

The accuracy calculated from the above tabulated measures is 81.25%,86.6071%,86.0119%,86.0119%85.7143%,76.7857%,42.5595%,84.2262%,83.631%,81.5476% for BayesNet, Logistic, Multilayer Perceptron, Multiclass Classifier, Classification via Regression, Decision Table, Zero R, J48, Random Forest, LAD Tree respectively for E-coli data set. It is evident from the above analysis that accuracy of Logistic classifier is maximum.

D. Discussion for Heart Disease

Table 4 The detailed accuracy for heart disease for different classifiers

	Bayes Net	Logistic	Multilayer Perceptron	Multilayer Classifier	Classification via Regression	Decision Table	Zero R	J48	Random Forest	REP Tree
CCI	249	249	250	249	243	79	60	78	80	228
ICI	45	45	44	45	51	21	40	22	20	66
KS	0.6589	0.6618	0.6727	0.6618	0.6135	0.5494	0	0.5259	0.569	0.4034

TP Rate	0.847	0.847	0.85	0.847	0.827	0.79	0.6	0.78	0.8	0.776
FP Rate	0.206	0.197	0.183	0.197	0.229	0.257	0.6	0.272	0.25	0.32
Precision	0.846	0.845	0.847	0.845	0.825	0.79	0.36	0.78	0.801	0.773
Recall	0.847	0.847	0.85	0.847	0.827	0.79	0.6	0.78	0.8	0.776
F-Measure	0.844	0.845	0.85	0.85	0.824	0.786	0.45	0.775	0.795	0.766
ROC Area	0.925	0.908	0.895	0.908	0.89	0.864	0.5	0.824	0.870	0.829

The accuracy calculated from the above tabulated measures is 85.034%,84.6939%,85.034%,84.6939%,82.6531%,80.6122%,63.9456%,80.9524%,78.9116%,77.551% for Bayes Net, Logistic, Multilayer Perceptron, Multiclass Classifier, Classification via Regression, Decision Table, Zero R,J48,Random Forest, LAD Tree respectively for E-colii data set. It is evident from the above analysis that accuracy of Multilayer Perceptron classifier is maximum.

where CCI-Correctly Classified Instances
ICI-Incorrectly Classified Instances
KS-Kappa Statistics

V. CONCLUSION AND FUTURE WORK

From the above experimental results, we conclude that no single classifier for different data sets provide the maximum accuracy. Each data set has the maximum accuracy of different classifiers. In case of Breast Cancer data set, the maximum accuracy is of Random Forest algorithm i.e. 97.9021%.In case of Diabetes data set, the maximum accuracy is of Decision Table algorithm i.e. 81.9923%.In case of E-colii data set, the maximum accuracy is of Logistic algorithm i.e. 86.6071%.In case of Hepatitis data set, the maximum accuracy is of Multilayer Perceptron i.e. 85.034%.

Therefore the future work will focus on building the hybrid model which will combine two different classifiers i.e. Multilayer Perceptron algorithm and LAD tree algorithm in order to get the maximum accuracy and performance from all of the classifiers discussed above.

REFERENCES

[1] Guo, L., Ma, Y., Cukic, B., & Singh, H. (2004, November). Robust prediction of fault-proneness by random forests. In *Software Reliability Engineering, 2004. ISSRE 2004. 15th International Symposium on* (pp. 417-428). IEEE.

[2] Rish, I. (2001, August). An empirical study of the naive Bayes classifier. In *IJCAI 2001 workshop on empirical methods in artificial intelligence* (Vol. 3, No. 22, pp. 41-46). IBM New York.

[3] Dreiseitl, S., & Ohno-Machado, L. (2002). Logistic regression and artificial neural network classification models: a methodology review. *Journal of biomedical informatics*, *35*(5), 352-359.

[4] Suykens, J. A., & Vandewalle, J. (1999). Training multilayer perceptron classifiers based on a modified support vector method. *IEEE Transactions on Neural Networks*, *10*(4), 907-911.

[5] Koc, L., Mazzuchi, T. A., & Sarkani, S. (2012). A network intrusion detection system based on a Hidden Naïve Bayes multiclass classifier. *Expert Systems with Applications*, *39*(18), 13492-13500.

[6] Patil, T. R., & Sherekar, S. S. (2013). Performance analysis of Naive Bayes and J48 classification algorithm for data classification. *International Journal of Computer Science and Applications*, *6*(2), 256-261.

[7] Pawlak, Z. (2002). Rough sets, decision algorithms and Bayes' theorem. *European Journal of Operational Research*, *136*(1), 181-189.

[8] Xing, Y., Wang, J., & Zhao, Z. (2007, November). Combination data mining methods with new medical data to predicting outcome of coronary heart disease. In *Convergence Information Technology, 2007. International Conference on* (pp. 868-872). IEEE.

[9] Fan, C. Y., Chang, P. C., Lin, J. J., & Hsieh, J. C. (2011). A hybrid model combining case-based reasoning and fuzzy decision tree for medical data classification. *Applied Soft Computing*, *11*(1), 632-644.

[10] Li, D. C., Liu, C. W., & Hu, S. C. (2011). A fuzzy-based data transformation for feature extraction to increase classification performance with small medical data sets. *Artificial Intelligence in Medicine*, *52*(1), 45-52.

[11] Al Jarullah, A. A. (2011, April). Decision tree discovery for the diagnosis of type II diabetes. In *Innovations in Information Technology (IIT), 2011 International Conference on* (pp. 303-307). IEEE.

[12] Jerez-Aragonés, J. M., Gómez-Ruiz, J. A., Ramos-Jiménez, G., Muñoz-Pérez, J., & Alba-Conejo, E. (2003). A combined neural network and decision trees model for prognosis of breast cancer relapse. *Artificial intelligence in medicine*, *27*(1), 45-63.

[13] Majhi, B., Panda, G., & Choubey, A. (2006, September). On the development of a new adaptive channel equalizer using bacterial foraging optimization technique. In *2006 Annual IEEE India Conference* (pp. 1-6). IEEE.

[14] Wu, Y., & Wang, C. (2006). Linear least-squares fusion of multilayer perceptrons for protein localization sites prediction. In *Proceedings of the IEEE 32nd Annual Northeast Bioengineering Conference* (pp. 157-158). IEEE.

[15] Palaniappan, S., & Awang, R. (2008, March). Intelligent heart disease prediction system using data mining techniques. In *2008 IEEE/ACS International Conference on Computer Systems and Applications* (pp. 108-115). IEEE.

[16] Kukar, M., Kononenko, I., Grošelj, C., Kralj, K., & Fettich, J. (1999). Analysing and improving the diagnosis of ischaemic heart disease with machine learning. *Artificial intelligence in medicine*, *16*(1), 25-50.

[17] Su, C. T., Yang, C. H., Hsu, K. H., & Chiu, W. K. (2006). Data mining for the diagnosis of type II diabetes from three-dimensional body surface anthropometrical scanning data. *Computers & Mathematics with Applications*, *51*(6), 1075-1092.

[18] Barriga, K. J., Hamman, R. F., Hoag, S., Marshall, J. A., & Shetterly, S. M. (1996). Population screening for glucose intolerant subjects using decision tree analyses. *Diabetes research and clinical practice*, *34*, S17-S29.

[19] Garcia, M., Sanchez, C. I., Lopez, M. I., Diez, A., & Hornero, R. (2008, August). Automatic detection of red lesions in retinal images using a multilayer perceptron neural network. In *2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (pp. 5425-5428). IEEE.

[20] Sinthanayothin, C., Kongbunkiat, V., Phoojaruenchanachai, S., & Singalavanija, A. (2003, September). Automated screening system for diabetic retinopathy. In *Image and Signal Processing and Analysis, 2003. ISPA 2003. Proceedings of the 3rd International Symposium on* (Vol. 2, pp. 915-920). IEEE.