RESEARCH PAPER | Science | Volume : 4 | Issue : 10 | October 2014 | ISSN - 2249-555X

Comparison of Different Classification Techniques Using WEKA for Hepatitis

KEYWORDS: Sequential Minimal Optimisation, Multilayer Perceptron, Root Mean-Squared Error, Mean Absolute Error

Jaikee Kumar Singh
Assistant Professor in Life Sciences Dept., Jaipur National University, Jaipur, Rajasthan, India.

Shruti Pareek
Assistant Professor in Life Sciences Dept., Jaipur National University, Jaipur, Rajsthan, India.

ABSTRACT *In today's world, large amount of data is available in science, industry, business and many other areas. This data can provide valuable information which can be used by management for making important decisions. But problem is that how can find valuable information. The answer is data mining. Data Mining is popular topic among researchers. There is lot of work that cannot be explored till now. But, this paper focuses on the fundamental concept of the Data mining i.e. Classification Techniques. In this paper Sequential Minimal Optimisation, Multilayer Perceptron classifiers are used for the classification of data set. The performance of these classifiers analyzed with the help of Mean Absolute Error, Root Mean-Squared Error and Time Taken to build the model and the result can be shown statistical as well as graphically. For this purpose the WEKA data mining tool is used.*

1. INTRODUCTION

Bioinformatics is the application of computer technology to the management of biological information. Computers are used to gather, store, analyse and integrate biological and genetic information, which can be applied to gene-based drug discovery and development and other such application. The need for Bioinformatics capabilities has been precipitated by the explosion of publicly available genomic information resulting from different projects.

The pattern classification problem can be defined such that for a given set of training examples, construct an algorithm, which will do a labelling task on a test dataset. Classifier assigns a class label to a test sample. It describes a decision boundary. Boundary can be linear or nonlinear. If dataset is linearly separable, a linear machine can classify all samples correctly (e.g., Perceptron, to be discussed later).

Hepatitis is inflammation of the liver from any cause. Generally, hepatitis results from a virus, particularly, one of five hepatitis viruses: A, B, C, D and E. However, hepatitis may also result from other viral infections such as infectious mononucleosis and cytomegalovirus infection. The major non-viral causes of hepatitis are alcohol and drugs. Hepatitis can be acute that may persist up to 6 months; or chronic, which occurs commonly throughout the world.

The aim of this small case study is to assess the effectiveness of classifiers (computational models) to help an oncology doctor for prediction of Hepatitis occurrence. The Waikato Environment for Knowledge Analysis (WEKA) data mining tools are used for this purpose.

2. CLASSIFICATION TECHNIQUES

Classification of data is very typical task in data mining. There are large numbers of classifiers that are used to classify the data such as Bayes classifier, function (Multi Layer Perceptron, Sequential Minimal Optimisation, etc.) classifiers, rules and trees classifiers, meta classifiers, etc. The goal of classification is to correctly predict the value of a designated discrete class variable, given a vector of predictors or attributes.

Artificial Neural Network

Artificial Neural Network (ANN) models consist of the following three principal elements:

a) Topology – the way an ANN model is organised into layers and the manner in which these layers are interconnected;

b) Learning – the technique by which information is stored in the network; and

c) Recall – how the stored information is retrieved from the network.

The basic structure of an ANN model consists of artificial neurons (Fig.1). The neurons are also sometimes referred to as processing elements (PEs), nodes, neurodes, units, etc., and are analogous to biological neurons in the human brain, which are grouped into layers (also called slabs). The most common ANN structure consists of an input layer, one or more hidden layers and an output layer.

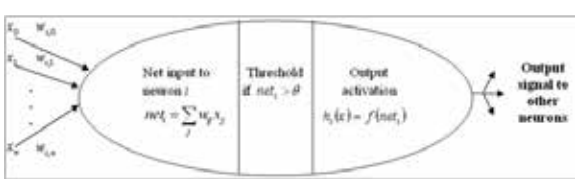


Fig.1: Schematic representation of general ANN model.

Let the input dimension be $n(n \in Z_+)$ and let the number of hidden neurons be $m(m \in Z_+)$. The training pairs are represented by $D = \{\mathbf{x}^{(p)}, t^{(p)}\}_*$, where $p = 1, 2, ..., P$; $P \in Z_+$, is the number of training exemplars; and the index p is always assumed to be present implicitly. The matrix w denotes the input to the hidden neurons connection strength, w_{ij} is the (i, j)th element of the matrix w representing the connection strength between the jth input and the ith hidden layer neuron. With this nomenclature, the net input to the ith hidden layer neuron is given by:

Volume : 4 | Issue : 10 | October 2014 | ISSN - 2249-555X

$$net_i = \sum_{j=1}^{n} w_j x_j + \theta_i^{(1)} = \mathbf{w}_i \cdot \mathbf{x} + \theta_i^{(1)} \quad \ldots (1)$$

where $\theta_i^{(1)}$ is the bias of the i[th] hidden layer neuron. The output from the i[th] hidden layer neuron is given by:

$$h_i(\mathbf{x}) = f^{(1)}(net_i) \quad \ldots (2)$$

where $f^{(1)}(\cdot)$ is a nonlinear activation function.

The activation function determines the output from a summation of the weighted inputs of a neuron. The activation functions for neurons in the hidden layer are often nonlinear and they provide the nonlinearities for the network. The choice of activation functions may strongly influence complexity and performance of ANN models. Although sigmoidal activation functions are most commonly used, there is no a priori reason why models based on such functions should always provide optimal decision borders. A number of alternative activation functions have been surveyed by some researchers.

The net input to the output neuron may be defined similarly as Eq. (1) as follows:

$$net = \sum_{i=1}^{m} v_i h_i + \theta^{(2)} = \mathbf{v} \cdot \mathbf{h} + \theta^{(2)} \quad \ldots (3)$$

where v_i represents the connection strength between the i[th] hidden layer neuron and the output neuron, while $\theta^{(2)}$ is the bias of the output neuron.

Adding a bias neuron x_0 with input value as +1, Eq. (1) can be rewritten as:

$$net_i = \sum_{j=0}^{n} w_j x_j = \mathbf{W}_i \cdot \mathbf{x} \quad \ldots (4)$$

where $w_{i0} = W_{i0} \equiv \theta_i^{(1)}$ and $\mathbf{W}_i$ is the weight vector $\mathbf{W}_i$ (associated with the i[th] hidden neuron) augmented by the 0[th] column corresponding to the bias. Similarly, introducing an auxiliary hidden neuron (i = 0) such that $h_0 = +1$, allows us to redefine Eq. (3) as:

$$net = \sum_{i=0}^{m} v_i h_i = \mathbf{V} \cdot \mathbf{h} \quad \ldots (5)$$

where $v_0 \equiv \theta^{(2)}$.

The equation for the network output neuron is given by:

$$net_o = f^{(2)}(net) = net \quad \ldots (6)$$

where $f^{(2)}(\cdot)$ is a linear function.

The notations are diagrammatically exemplified in Fig.2. This figure represents an n-input, m-hidden neuron and one-output feedforward ANN model. Such an ANN model is trained to fit a dataset D by minimising an error function (or performance function) as:

$$F = E_D(\mathbf{W}) = \frac{1}{P}\sum_{p=1}^{P} \varepsilon^2 = \frac{1}{P}\sum_{p=1}^{P} \left(net_o^{(p)} - t^{(p)}\right)^2 \quad \ldots (7)$$

This function is minimised using standard optimisation method.

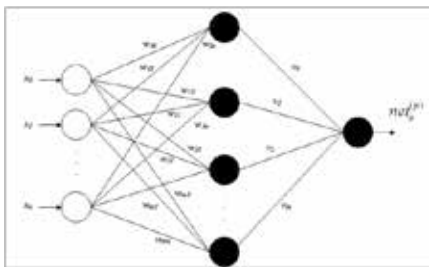


Fig.2: Schematic of a feedforward ANN model.

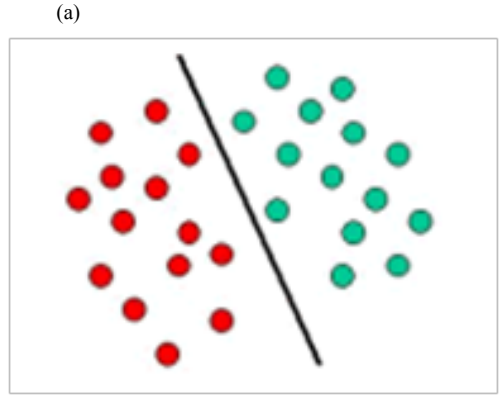
An ANN model is also known as Multi Layer Perceptron (MLP) especially in WEKA perspective. Hence, the terms ANN and MLP are used synonymously in subsequent discussion. As stated earlier, the MLP usually uses nonlinear activation functions in its neurons to define the outputs; and produces a nonlinear relationship between inputs and outputs across the network. Therefore, the MLP can be seen as a nonlinear pattern recognition technique.

Support Vector Machine

The Support Vector Machine (SVM) is a new and promising technique for data classification and regression. After the development in the past five years, it has become an important topic in machine learning and pattern recognition. Besides better theoretical foundation, it is practically competitive with existing methods such as ANN models and decision trees also. The SVM application in Bioinformatics is increasing.

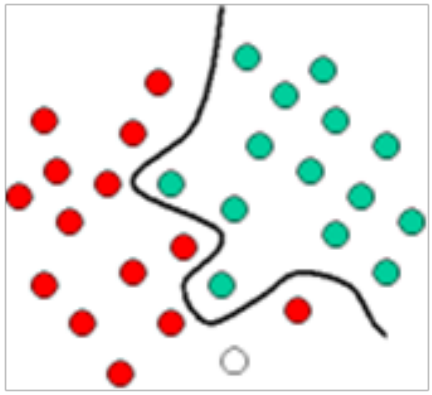
The SVM technique was first developed by Vapnik and his group at erstwhile AT&T Bell Laboratories. The original idea is to use a linear separating hyperplane, which maximises the distance between two classes to create a classifier as shown in Fig.3 (a). For problems, which cannot be linearly separated in the original input space as shown in Fig.3(b); SVM's employ two techniques to deal such a case. First technique is soft margin hyperplane; and in second technique, the original input space is nonlinearly transformed into a higher dimension feature space. Then in this new feature space, it is more possible to find a linear optimal separating hyperplane as shown in Fig.3(c).

(a)



(b)

Volume : 4 | Issue : 10 | October 2014 | ISSN - 2249-555X



(c

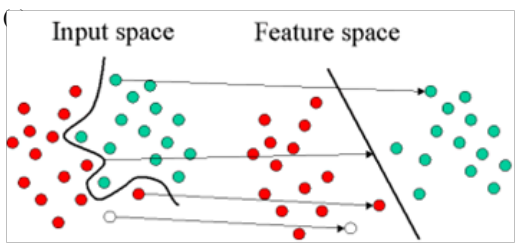


Fig.3: (a) Linearly separable objects; (b) Linearly non-separable objects; and (c) Nonlinear objects rearranged as linear objects using set of mathematical functions called kernels (basis of SVM).

The working of SVM is equivalent to a statistical learning machine that maps points of different categories from n-dimensional space into a higher dimensional space where the two categories are more separable. It tries to find an optimal hyperplane in that high dimensional space that best separates the two categories of points.

Essentially, the hyperplane is learned by the points that are located closest to the hyperplane, which are called support vectors. There can be more than one support vector on each side of the plane. Fig.4 shows an example of two categories of points separated by a hyperplane.

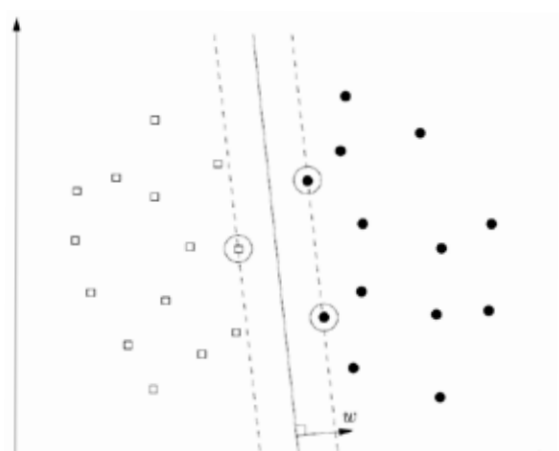


Fig.4: Separating hyperplane for feature selection where circles indicate the support vectors.

Identification of a separating hyperplane is a popular method for pattern classification. SVM looks for the separating hyperplane with largest margin. For two-category problem, (using conventional notations):

$x_i \cdot w + b \geq +1$, for $y_i = +1$ ...(8)

$x_i \cdot w + b \leq -1$ for $y_i = -1$ ...(9)

If problem is linearly separable, there will exist a and , which will satisfy these equations. Combining these into one inequality:

$y_i (x_i \cdot w + b) - 1 \geq 0, \forall i$ ...(10)

The vectors, which satisfy the equality in the previous equation, are called support vectors as shown in Fig.5.

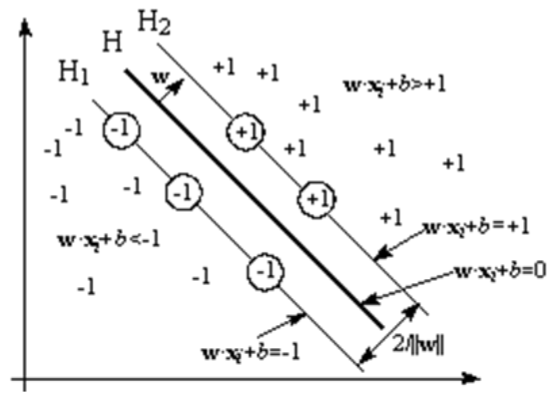


Fig.5: The optimum separation hyperplane.

The limitations of SVM are the selection of a suitable kernel, speed and size, both in training and testing. Another limitation of SVM model is that it classifies only two classes at a time. Practically, there are multiple classes to classify, which results in requirement of multiclass SVM. Multiclass SVM involves the construction of binary SVM classifiers for all pairs of classes.

Sequential Minimal Optimisation in WEKA Perspective

Sequential Minimal Optimisation (SMO) is an algorithm for efficiently solving the optimisation problem, which arises during the training of SVMs. SMO can quickly solve the SVM problem without any extra matrix storage. It was invented by John Platt in 1998 at Microsoft Research Organisation. SMO is widely used for training SVMs. SMO is an iterative algorithm for solving the optimisation problem described above. SMO breaks this problem into a series of smallest possible sub-problems, which are then solved analytically.

3. WEKA TOOLS

The WEKA toolkit is used to analyse the dataset with the data mining algorithms. WEKA is an assembly of tools of data classification, regression, clustering, association rules and visualisation techniques. The toolkit has been developed in Java programming environment and is open source software issued under the GNU General Public License. The WEKA tool incorporates the four applications within it:

- WEKA Explorer,
- WEKA Experiment,
- WEKA Knowledge Flow and
- Simple CLI.

Volume : 4 | Issue : 10 | October 2014 | ISSN - 2249-555X

For the classification of dataset, WEKA Explorer is used to generate the result or statistics. Weak Explorer incorporates the following features within it:

- **Pre-process:** It is used to process the input data. For this purpose, the filters are used that can transform the data from one form to another form. Basically, two types of filter are used, *i.e.,* supervised and unsupervised.

- **Classify:** Classify tab is used for the classification purpose. A large number of classifiers are used in WEKA such as Bayes, rule, tree and meta, *etc*. Four types of test options are mentioned within it.

- **Cluster:** It is used for the clustering of the data.

- **Associate:** Establish the association rules for the data.

- **Select attributes:** It is used to select the most relevant attributes in the data.

- **Visualize:** View an interactive 2D plot of the data.

4. DATA

Dataset used in WEKA is arranged in Attribute-Relation File Format (ARFF) that consists of special tags to indicate different components in the dataset such as attribute names, attribute types, attribute values and the data. This case study uses the dataset on hepatitis occurrence taken from the UCI repository (http://repository.seasr.org/Datasets/UCI/csv/), *i.e.,* real time multivariate dataset (please refer to the 'Appendix' for the original dataset). The dataset comprises 156 hepatitis patients' records on several attributes as summarised in Table-1.

Table 1: Attributes used for developing the classifiers.

Serial Number	Attribute	Attribute Description	Units of Measurements
1.	Age	Age of the patient at the time of diagnosis	Years
2.	Sex	Patient Gender	Male and Female
3.	Steroid	Drugs that mimics the effects of hormone in the body	Yes or No
4.	Antiviral	Used specifically for treating viral infections	Yes or No
5.	Fatigue	Muscle weakness	Yes or No
6.	Malaise	Feeling of general discomfort or uneasiness	Yes or No
7.	Anorexia	Symptom of poor appetite	Yes or No
8.	Liver-Big	Liver Size	Yes or No
9.	Liver-Firm	Enlargement of Liver	Yes or No
10.	Spleen-Palpable	Enlargement of the spleen	Yes or No.
11.	Spiders	Blood vessels near the skin surface	Yes or No
12.	Ascites	Accumulation of fluid in the peritoneal cavity	Yes or No
13.	Varices	Veins in the lower third of the esophagus.	Yes or No
14.	Bilirubin	Diagnosis or monitor liver disease such as Hepatitis	mg/dl
15.	Alkaline Phosphate	Enzymes with low substrate specificity	units/l
16.	Serum Glutamic Oxaloacetic Transaminase	Measure the amount of protein enzyme called Glutamic Oxaloacetic Transaminase occurring in our blood	units/l
17.	Albumin	Water soluble protein	g/l
18.	Protime	Test used to determine the clotting tendency of blood	Second
19.	Histology	Study of tissues for the functional morphology of man and animals	Yes or No
20.	Class	Hepatitis stage	Live or Die

These data are used to determine occurrence events for new patients. Hepatitis dataset is originally in the form of text file. Firstly, these are converted into the MS-Excel format (*.xls*); *.xls* format to *.csv* format and then *.csv* format converted into the *.arff* format.

In this case study, two above discussed data mining techniques, *viz.*, ANN and SVM algorithms are explored to predict instances of occurrence of hepatitis using the aforementioned dataset. The classification potential of these techniques has been compared to find the most suitable one for predicting the hepatitis occurrence. The following steps are pursued to train and test the ANN classifier and SVM classifier as depicted in Fig.6 and Fig.7, respectively.

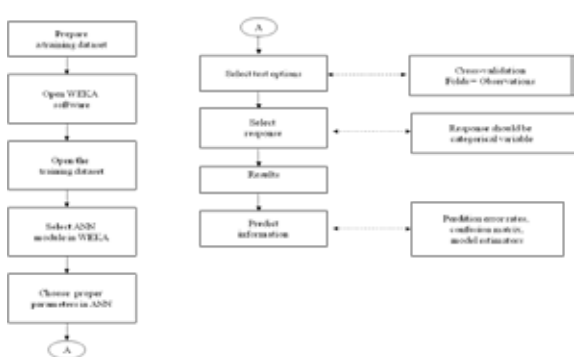


Fig.6: Flowchart showing steps involved in constructing an ANN classifier model under WEKA environment to predict hepatitis occurrences.

Volume : 4 | Issue : 10 | October 2014 | ISSN - 2249-555X

Fig.7: Flowchart showing steps involved in developing the SVM classifier model under WEKA environment to predict hepatitis occurrences.

5. EXPERIMENT

MLP classifier development

The MLP classifier algorithm under WEKA environment was employed to the hepatitis dataset as described earlier. As a result, WEKA generates a Confusion Matrix (Contingency Table), which includes four measures: *a)* the number of samples classified as true while they were true (True Positive:); *b)* the number of samples classified as false while they were false (True Negative:); *c)* the number of samples classified as false while they were actually true (False Negative:();and *d)* the number of samples classified as true while they were actually false (False Positive:). This process is delineated through a WEKA screenshot in Fig.8. This process involves opening the dataset in the ARFF format followed by selecting the MLP classifier algorithm. Algorithm parameters are set in the parameter box as described here. The GUI parameter is set as 'True'. Then the MLP classifier is run through WEKA Explorer for training the dataset. Hence, MLP classifier model is generated as shown through screenshot in Fig.9. The default values are used for the parameters, *viz.*, number of epochs as 500; error per epoch as 0; learning rate as 0.3; momentum as 0.2; and validation threshold as 20. With these settings, the classifier is trained with the training dataset.

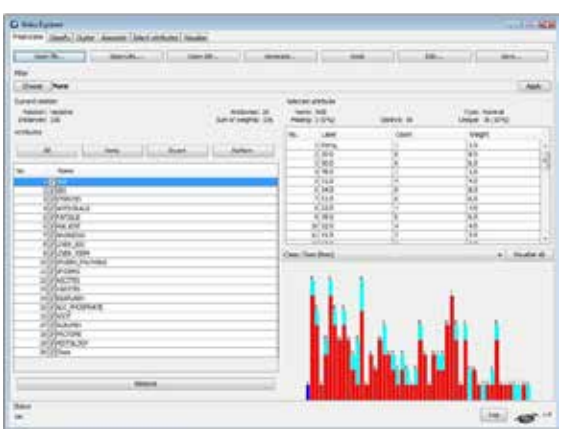


Fig. 8: Dataset is open for training in WEKA Explorer.

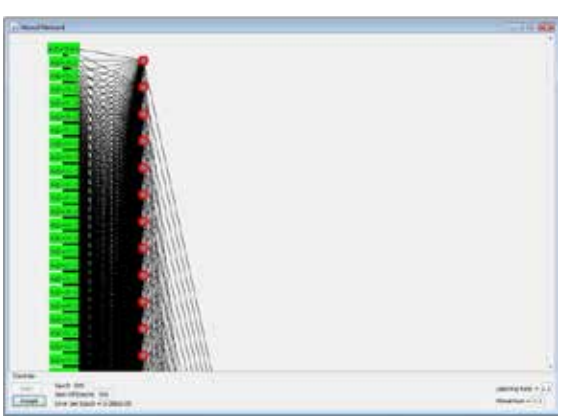


Fig. 9: Schematic diagram of the MLP classifier generated by WEKA.

The error per epoch achieved was as 0.0001. Thereafter, WEKA produces a confusion matrix that is shown through screenshot in Fig.10.

Fig.10: MLP classifier confusion matrix produced by WEKA.

Once the confusion matrix is constructed, the Precision, Recall and F–Measure are calculated as follows:

Precision=TP/(TP+FP) ...(11)

Recall=TP/(TP+FN) ...(12)

F-Measure=2×TP/(2×TP+FP+FN) ...(13)

Precision measures percentage of actual patients () among patients that get declared disease. Recall measures percentage of actual patients that were discovered. F–Measure balances between Precision and Recall. A Receiver Operator Characteristic (ROC) space is defined by rate () and rate () as and axes, respectively.

TPR= TP/(TP+FN) ...(14)

FPR=FP/(FP+TN) ...(15)

Accuracy=(TP+TN)/(TP+TN+FP+FN) ...(16)

In this case study, the confusion matrix thus produced as a result of MLP based classifier is shown in Table-2.

Table 2: Classification of hepatitis recurrence data by MLP model.

TP Rate	FP Rate	Precision	Recall	F–Measure	ROC Area	Class
1	0	1	1	1	1	Live
1	0.01	0.97	1	0.99	1	Die

SMO classifier development

The SMO classifier algorithm under WEKA environment is employed to the hepatitis dataset as already shown through WEKA screenshot in Fig.8 above. Then selecting

Volume : 4 | Issue : 10 | October 2014 | ISSN - 2249-555X

the SMO classifier in WEKA and keeping default values for the SMO parameters, *viz.*, complexity as 1.0; tolerance power as 0.001; number of folds as 10, and seed is set to unity. Now, SMO is run through WEKA Explorer for training the dataset and confusion matrix is produced by WEKA as a result (Fig.11).

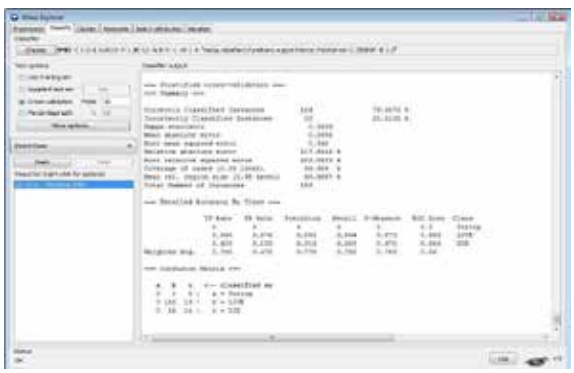


Fig.11: SMO classifier confusion matrix produced by WEKA

Further, the confusion matrices produced as a result of SMO based classifier is shown in Table-3.

Table 3: Classification of hepatitis recurrence data by SMO model.

TP Rate	FP Rate	Precision	Recall	F–Measure	ROC Area	Class
0.89	0.57	0.85	0.89	0.87	0.66	Live
0.44	0.16	0.52	0.44	0.47	0.67	Die

The performance of above mentioned classification techniques have been evaluated using following methods:

Mean Absolute Error (MAE): A statistical measure to assess as to how far an estimate is from actual values, *i.e.,* the average of the absolute magnitude of the individual errors. It is usually similar in magnitude but slightly smaller than the Root Mean Squared Error (RMSE).

RMSE: The RMSE calculates the differences between values predicted by a model/an estimator and the values actually observed from the phenomenon/process being modelled/ estimated. RMSE is used to measure the accuracy. It is considered as ideal if it is small.

Time: The amount of time required to build the model.

Comparative Analysis of MLP and SMO classifiers

A comparative analysis of the MLP and SMO classifiers is summarised in Table- 4. Evidently, the time taken by the SMO classifiers to train from the data is 0.11 seconds whereas the time taken by the MLP classifier to train from the data is relatively larger, *i.e.*, 31.85 seconds. Thus, in terms of time taken, the SMO classifier seems to be the better than MLP classifier. However, the analysis of the other two measures, *i.e.*, MAE and RMSE revealed that the model based on MLP algorithm seems to perform relatively better. The MLP classifier classified the instances more correctly as compared to the SMO classifier.

Table 4: Training and Simulation results of ANN and SVM.

Algorithm	Correctly Classified	Incorrectly Classified	Time Taken (in Seconds)	Mean Absolute Error	Root Mean Square Error	Relative Absolute Error (%)	Root Relative Square Error (%)
Neural Network	155	1	31.85	0.01	0.06	3.88	17.44
Support Vector Machine	124	32	0.11	0.27	0.385	117.64	103.86

Further, the classification accuracy of the two classifiers, *i.e.*, ANN and SVM models run against the Hepatitis dataset is shown in Table-5. Evidently, neural network classifier seems to classify the instances of occurrence of hepatitis in patients better than that with the support vector machine classifier.

Table 5: Classification accuracy of ANN and SVM models.

Dataset	Classifier Algorithm	
	ANN	SVM
Hepatitis	99.36%	79.49%

5. CONCLUSION AND FUTURE WORK

Two classifiers based on ANN and SVM algorithms (under WEKA environment) have been developed for the classification of data pertaining to occurrence of hepatitis. The classification potential of the two models has been compared. On the basis of this small case study, it seems that the ANN based classifier performed better with accuracy as 99.36% than that of the SVM based classifier, which achieved accuracy of 79.49%.

WEKA supports various classifiers such as Fuzzy rules, REP tree, Random tree, Gaussian Function, Regression, *etc*. Recently, Proximal Support Vector Machine (PSVM) technique has emerged that classifies points by assigning them to the nearest of two parallel planes unlike a standard SVM, which classifies points by assigning them to one of two disjoint half-spaces. Thus, the future work will be based on these classifiers, *i.e.,* applying these classifiers on the dataset so as to analyse their performance for classifying the occurrence of hepatitis instances in the patients. In this case study, only two statistical measures have been used to assess performance of the classifiers. However, in future, many more validation measures can also be explored to achieve better results.

Volume : 4 | Issue : 10 | October 2014 | ISSN - 2249-555X

REFERENCE

[1] Agrawal, R., Imielinski, T., Swami, A., 1993. Data mining: a performance perspective. IEEE Transactions on Knowledge and Data Engineering 5, 914–925.
[2] Bouckaert, R. R., Frank, E., Hall, M., Kirkby, R., Reutemann, P., Seewald, A., Scuse, D., 2010. WEKA Manual for Version 3-6-2. University of Waikato, Hamilton, New Zealand. URL: http://www.gnu.org/copyleft/gpl.html.
[3] Collins, M., Duffy, N., 2002. New ranking algorithms for parsing and tagging: kernels over discrete structures and the voted perceptron. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, July 2002, pp. 263–270.
[4] Desouza, K. C., 2001. Artificial intelligence for healthcare management. In: Proceedings of the 1st International Conference on Management of Healthcare and Medical Technology, Enschede, Netherlands, April 22-27.
[5] Feelders, A., Daniels, H., Holsheimer, M., 2000. Methodological and practical aspects of data mining. Information and Management 5, 271–281.
[6] Freund, Y., Schapire, R. E., 1999. Large margin classification using the perceptron algorithm. Machine Learning 37, 277–296.
[7] Grossman, D., Domingos, P., 2004. Learning Bayesian network classifiers by maximizing conditional likelihood. In: Proceedings of the 21st International Conference on Machine Learning, Banff, Canada.
[8] Han, J., Kamber, M., 2000. Data Mining: Concepts and Techniques. Morgan Kaufmann, San Francisco, CA.
[9] Hassoun, M. H., 1999. Fundamentals of Artificial Neural Networks. MIT Press, Cambridge, MA.
[10] Hu, Y., Li, H., Cao, Y., Teng, L., Meyerzon, D., Zheng, Q., 2006. Automatic extraction of titles from general documents using machine learning. Information Processing and Management 42, 1276–1293.
[11] Peksen, Y., Canbaz, S., Leblebicioglu, H., Sunbul, M., Esen, S., Sunter, A. T., 2004. Primary care physicians' approach to diagnosis and treatment of hepatitis B and hepatitis C patients. BMC Gastroenterology 4, 1-6. Available online at: http://www.biomedcentral.com/1471-230X/4/3.
[12] Ridgeway, G., Madigan, D., Richardson, T., 1998. Interpretable boosted naive Bayes classification. In: Agrawal, R., et al. (Eds). In: Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining. AAAI Press, Menlo Park, pp. 101–104.
[13] Witten, I. H., Frank, E., 2005. Data Mining Practical Machine Learning Tools and Techniques. Second Edition. Morgan Kaufmann, San Francisco, CA.
[14] Zak, S.H., 2003. Systems and Control. Oxford University Press, NY.