IOSR Journal of Dental and Medical Sciences (IOSR-JDMS)
e-ISSN: 2279-0853, p-ISSN: 2279-0861.Volume 15, Issue 4 Ver. I (Apr. 2016), PP 105-111
www.iosrjournals.org

Performance Accuracy between Classifiers in Sustain of Disease Conversion for Clinical Trial Tuberculosis Data: Data Mining Approach

Chinnaiyan Ponnuraja [1], Babu C Lakshmanan[2], Valarmathi Srinivasan[3]
[1]*Department of Statistics, National Institute for Research in Tuberculosis (ICMR), Chennai, India*
[2] *Cognizant Technology Solutions (Chennai), India*
[3] *Department of Epidemiology, The TamilNadu Dr.MGR Medical University, Chennai, India*..

Abstract: *Data mining has been used intensively and extensively by many organizations. In healthcare, data mining is becoming increasingly popular, if not increasingly essential. Data mining applications can greatly benefit all parties involved in the healthcare industry. For example, data mining can help physicians to identify risk factors, effective treatments and best practices in health care industry which helps patients to receive better and affordable healthcare services. The huge amounts of data generated by healthcare transactions are too complex and voluminous to be processed and analyzed by traditional methods. Data mining provides the methodology and technology to transform these mounds of data into useful information for decision making. This work will explore data mining applications within healthcare in one of the major area such as the evaluation of treatment effectiveness as well as disease conversion in tuberculosis patients; it aims to compare various classification techniques and its performance accuracy to build a predictive modelling with controlled clinical trial tuberculosis (TB) data. The classification of tuberculosis patients is of substantial importance in TB disease conversion. During the last few years, many algorithms have been proposed for this task. In this paper, we review different supervised machine learning techniques and its performance accuracy for classification with an original dataset and carry out a methodological comparison using "WEKA" software (Wang, 2010). We used the C4.5 (J48) tree classifier, Iterative Dichotomiser- 3 (ID3), a Multilayer Perceptron (MLP) and a naive Bayes 7classifier over a large set of TB data. It is found that Multilayer Perceptron achieves a competitive performance than naive Bayes, and when the number of features to be classified is reduced naive bayes performs well.*

Keywords: *C4.5 (J48) tree classifier, ID3, Multilayer Perceptron, naive Bayes, Tuberculosis, WEKA, Data Mining*

I. Introduction

Data mining is the process of exploring large amounts of data for finding unknown patterns and building models based on the information. This process has become a progressively more pervasive activity in all areas of medical science research. Data mining has resulted in the discovery of useful hidden patterns from massive databases. Data mining problems are repeatedly resolved using different approaches from both computer sciences and statistics. In computer science there are numerous approaches, such as multi-dimensional databases, machine learning, soft computing and data visualization; and in statistics, it includes hypothesis testing, clustering, classification, and regression techniques. In this paper clustering, classification, and regression techniques in the area of medical research data irrespective of all major diseases is explored.

Four different classifiers of supervised machine learning, namely the naive Bayes algorithm, the J48, ID3 decision tree and the Multilayer Perceptron (MLP) are compared. Chronological order of classifiers are arranged accordingly ID3 (1979), C4.5 (1993), C4.8 (1996) and C5.0 (purely for commercial). J48 is a direct implementation of C4.8. In general, Machine Learning Algorithms need to be trained for supervised learning tasks in the vein of classification, prediction etc. or for unsupervised learning tasks similar to clustering.

The same machine learning techniques were already used in literature: in particular, Bellaachia and Guven in [1], revising a study of Delen et al. [3], used the above methods to find the most suitable one for predicting survivability rate of breast cancer patients. Our study is instead motivated by the necessity to find a programmed and robust method to validate classification of Tuberculosis Disease conversion.

II. Classification Prediction

Main mission of Data mining is Classification. Data mining, is a machine learning discipline motivated by pattern recognition, which is a branch of science, of which one of its goals is to classify objects into a number of categories referred to as classes(Witten et al(2011); Haiyang 2011; Han and Kamber (2006) . Objects refer to compact data units specific to a particular problem, which is in general, known as patterns. Classification

prediction encompasses two levels: classifier construction and the usage of the classifier constructed. The former is concerned with the building of a classification model by describing a set of predetermined classes from a training set as a result of learning from that dataset. Each sample in the training set is assumed to belong to a predefined class, as determined by the class attribute label. The model is represented as classification rules, decision trees, or mathematical formula. The later involves the use of a classifier built to predict or classify unknown objects based on the patterns observed in the training set. The entire process begins with collection of evidence acquired from various data sources or warehouses. In the ideal situation, the data should be of low dimensionality, independent and discriminative so that its values are very similar to characteristics in the same class but very different in features from different classes. Raw data hardly satisfies these conditions and therefore a set of procedures called *feature generation, extraction and selection* is required to provide a relevant input for classification system (Michie et al, 1994).

Multi Layer Perceptron (MLP)

MLP is a feed forward artificial neural network model that maps sets of input data onto a set of appropriate outputs. This type of neural network is known as a supervised network because it requires a desired output in order to learn. The goal of this type of network is to create a model that correctly maps the input to the output using historical data so that the model can then be used to produce the output when the desired output is unknown. As its name advocates, it consists of multiple layers of nodes in a directed graph, with each layer fully connected to the next one. The architecture of this class of networks, besides having the input and the output layers, also have one or more intermediary layers called the hidden layers(Lu et al(1996); Han and Kamber (2006); Witten et al(2011)). The hidden layer does intermediate computation before directing the input to output layer. A graphical representation of an MLP is shown below:

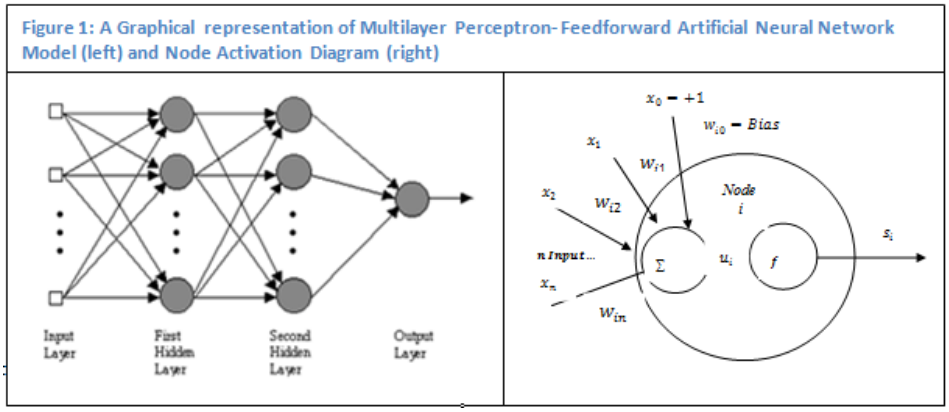

Figure 1: A Graphical representation of Multilayer Perceptron-Feedforward Artificial Neural Network Model (left) and Node Activation Diagram (right)

From the above figure (right) the node activation is explained like each node ***outputs*** an activation function applied over the weighted sum of ***inputs***

$$s_i = f(w_{i,0} + \sum_{j \in l} w_{i,j} * s_j)$$

ID3 (Iterative Dichotomiser 3)

ID3 algorithm was introduced in the year 1986 by Quinlan Ross. It is the precursor to the C4.5 algorithm. The ID3 is inductive inference algorithm which is successfully applied to diagnose medical cases. It learns decision trees by constructing them top down and the strategy is dividing and conquers which is based on Hunts algorithm. The tree is normally constructed in two phases namely tree building and pruning. The measure Information gain is used to select the best attribute at each step in growing the tree. The basic idea behind ID3 is that each node of the tree corresponds to a non-categorical attribute and each arc to a possible value of that attribute. A leaf of the tree specifies the expected value of the categorical attribute for the records described by the path from the root to that leaf. Ultimately it mimics the definition of the Decision Tree. Secondly, in the decision tree each node should be associated to a non-categorical attribute which is *most informative* among the attributes which are not yet considered in the path from the root and finally it establishes a "Good" decision tree. Thirdly, *Entropy* is used to measure how informative is a node which defines what we mean by "Good". The major feature of ID3 is, it only accepts categorical attributes in building a tree model. However it does not give accurate result when there is noise and also it does not support. The ID3 algorithm can be handled continuous attributes also by discretizing or directly, only by considering the values to find the best split point by taking a threshold on the attribute values. ID3 algorithm is used in knowledge acquisition for tolerance design, applied to calculate logistic performance and applicable in the field of computer crime forensics.

J48 Classifier

J48 is an implementation of C4.5 classification algorithm (known in "Weka" as J48: J for Java), is producing decision tree based on information theory and based on Hunt's algorithm as well. ID3 is an implementation of Quinlan's which the precursor to J48 is. ID3 mainly operates on nominal attributes. But J48 handles both categorical and continuous attributes to build a decision tree. J48 is serially implemented like ID3. Using this algorithm, pruning can take place that is it replaces the internal node with a leaf node thereby reducing the error rate unlike ID3. In order to handle continuous attributes, it splits the attribute values into two divisions based on the selected threshold such that all the values above the threshold as one child and the remaining as another child. It very well handles missing attribute values. J48 uses gain ratio impurity method to evaluate the splitting attribute that is to build the decision tree (Quinlan (1986)). It removes the biasness of information gain when there are many outcome values of an attribute. J48 is used in classification problems and it is the most used algorithm for building Decision Tree. It is suitable for real world problems as it deals with numeric attributes and missing values.

Naive Bayes classifier

A naive Bayes classifier is a probabilistic classifier based on Bayes' theorem with independence assumption (Caruana and Niculescu-Mizil (2006); Bermejo et al (2014)). The objective is to predict the class of test instances as accurately as possible. This kind of classifier is termed naive because it is based on two simplifying common assumptions: firstly, it assumes that the predictive attributes are conditionally independent given the class and secondly, the values of numeric attributes are normally distributed within each class. Naive Bayes treats discrete and continuous attributes somewhat differently. For each discrete attribute, the probability that the attribute X will take on the particular x when the class c is modelled by a single real number between 0 and 1. In contrast, each continuous attribute is modelled by some continuous probability distribution over a range of that attribute's values. Let C be the random variable denoting the class of an instance and X be a vector of random variables denoting the observed attribute values. Let c be a particular class label and x represent a particular observed attribute value. If we have a test case x to classify, the probability of each class given the vector of observed values for the predictive attributes may be obtained using the Bayes' theorem:

Clinical Trial Tuberculosis (TB) DATA

The data used in this work is the randomized controlled clinical trial pulmonary tuberculosis (*pTB*) data from National Institute for Research in Tuberculosis (ICMR). The eligible patients were randomly allocated into three different regimens including a standard revised form of RNTCP(Revised National Tuberculosis Control Programme) regimen as a control treatment with two more trial regimen of each six months duration (Tuberculosis Research Centre (2004)). All patients were assessed clinically and bacteriologically every month up to six months. In this application there were 1237 patients with five variables such as *age* (years) and *weight* (kg) at baseline as a continuous, *sex* (male, female), *drug susceptibility test* (DST: sensitive to all drugs and resistant any one drug) and *treatment group* (treatment A, treatment B, Control) as categorical, included with an outcome variable of *status* of two levels (sputum culture conversion: "converted" and "not converted").

Table1: Attributes respond to sputum culture conversion of Randomized Controlled Clinical Trial pulmonary tuberculosis(PTB) data

Attributes	Converted	Not Converted
	1062(86%)	175(14%)
Regimens		
Treatment A	369	46
Treatment B	338	77
Control	355	52
Sex		
Female	281	32
Male	781	143
Pre-Treatment DST		
sensitive	916	91
resistant	146	84
AGE_Group		
Less & Equal 24	263	35
25-34	344	57
35-45	289	60
More & Equal to46	166	23
Weight_Group		
Less & Equal to 38	336	73
39-44	404	58
More & Equal to44	322	44

III. Results and Discussion

In this paper, Weka software was used for implementing large collection of machine leaning algorithms and is extensively used in data mining applications. The dataset was loaded into WEKA explorer. The classify panel enables the user to apply classification and regression algorithms to estimate the accuracy of the resulting predictive model. The ID3, J48 (C4.5) decision tree algorithm and Multilayer Perceptron, naive Bayes algorithm were implemented in WEKA. The data were transformed into Weka data mining software as acceptable formats as listed. The data file was in the Comma Separated Value (CSV) file format in Microsoft excel and later it was converted into Attribute Relation File Format (ARFF) using the ARFF converter and then classified using WEKA and the result is produced. The 10-fold cross-validation is selected under "Test options" for evaluation approach. After applying the pre-processing and training methods, we make an effort to analyze the data visually and figure out the distribution of values.

Figure 2: Visualization of all attributes are involved in this application as per the act in response with sputum culture conversion

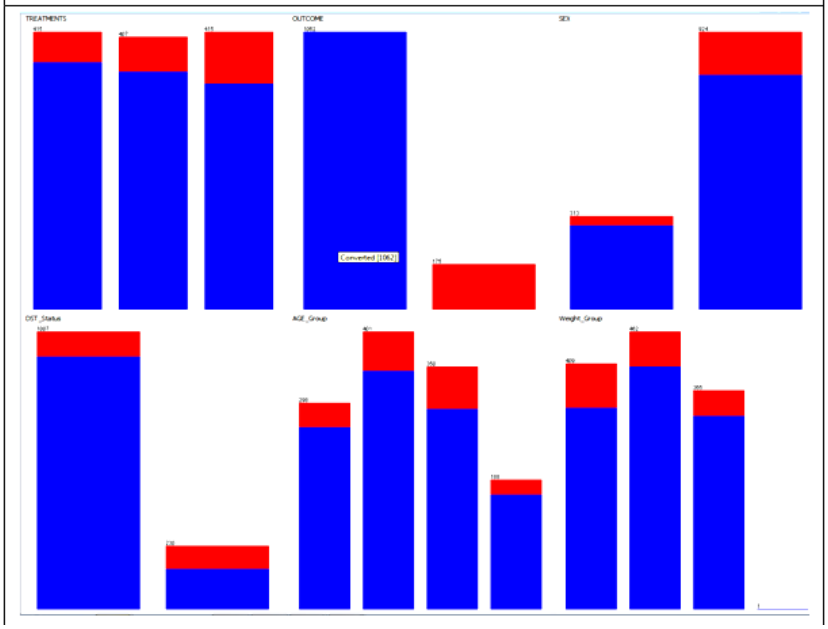

Figure 1 illustrates the visualization chart for all attributes which are involved in this application and they have been scheduled as per the act in response with outcome of point in time to sputum culture conversion.

Some experiments were carried out in order to evaluate the performance and usefulness of different classification algorithms for predicting TB disease conversion based on sputum culture conversion. The following metrics were used to verify the performance of the model: Total number of Instances, Prediction accuracy based on correctly classified Instances (%), Kappa statistics, Mean absolute Error (MAE) which measures the average magnitude of the errors in a set of forecasts, without considering their direction and it also measures accuracy for continuous variables, Root Mean Squared Error(RMSE) is a quadratic scoring rule which measures the average magnitude of the error, Root Relative Absolute Error (RRAE(%)) and Root Relative Squared Error (RRSE(%)).

Table 2: Overall Performance Accuracy between classifiers

Stratified cross-validation Summary	**NaiveBayes**	**Multilayer Perceptron**	**J48**	**Id3**
Total Number of Instances	1237	1237	1237	1237
Correctly Classified Instances	**86.2 %**	**86.0%**	**84.8%**	**84.3%**
Kappa statistic	0.08	0.07	0.16	0.09
Mean absolute error	0.2154	0.2151	0.2096	0.214
Root mean squared error	0.329	0.328	0.348	0.351
Root relative Absolute error	88.48 %	88.38 %	86.13 %	88.25 %
Root relative squared error	94.46 %	94.35 %	100.01 %	101.09%

The percentage of correctly classified instances is called as model accuracy and the accuracy is obviously called of a model performance. As a result from the Table2, The naive Bayes classifier has more accuracy compared to ID3 and J48 classifiers but MLP is almost closely to naive Bayes. From the same table, it is seen that ID3 has the lowest model performance accuracy than compared to all others classifiers. The Kappa

statistic value is very closer between naive Bayes and MLP. The table 2 as well shows the error values derived for each algorithm based on the performance errors. The values of Mean absolute error, Root Mean Squared Error, Root Absolute Error, Root Relative Squared Error for naive Bayes has comparatively lower when compared to the values of MLP, ID3 and J48. This result reveals that the naive Bayes and MLP algorithms are suitable for the prediction of disease conversion for TB. However the other two algorithms are likely to have closer value with naive Bayes and MLP algorithms. . In view of the fact that the lesser the error value the better the prediction. As per the concern of ROC area naive Bayes and MLP algorithms are having similar values. The higher the ROC Area is better in the aspect of model prediction.

Confusion Matrix

The confusion matrices have been arrived for all methods using *Weka* in accordance with (Provost and Fawcett (2001); Provost, et al (1998)) and it matches the predicted and actual values according to *Weka's criteria for how to decide the matrix is as* follows.

The first criterion of *Weka* Confusion Matrix is, if "a" is taken to be the positive class; that is, it has "disease":

	a	b
actual a=0	TP	FN
actual b=1	FP	TN

The second criterion for *Weka* Confusion Matrix is if "a" is taken to be the negative class; that is, it has "no disease":

	a	b
actual a=0	TN	FP
actual b=1	FN	TP

The second criterion is the choice for our data which looks a like the following table

	Predicted (→)	
Actual (↓)	**Negative**	**Positive**
Negative	True Negative(**TN**)	False Positive(**FP**)
Positive	False Negative(**FN**)	True Positive(**TP**)

The table of confusion matrix is as like as 2 × 2 table. The left column is meant for actual (↓) and the right top row is for predicted (→) as well. Three accuracy measures can be defined like Sensitivity, Specificity and Accuracy.

The Sensitivity is as well as called Type II Error, "recall" in *Weka* = (TP/ (FN+TP)) × 100 (%);

$$\text{The Sensitivity is as well as called Type II Error} = \frac{\text{TP}}{\text{TP}+\text{FN}} * 100\%$$

The Specificity (Type I Error) = (TN/ (TN+FP)) × 100 (%);

$$\text{The Specificity (Type I Error)} = \frac{\text{TN}}{\text{TN}+\text{FP}} * 100\%$$

The Accuracy = (TN+TP)/ (TN+FP+FN+TP) × 100 (%);

$$\text{The Accuracy} = \frac{\text{TN}+\text{TP}}{\text{TN}+\text{FP}+\text{FN}+\text{TP}} * 100\%$$

Table3: Performances of accuracy between classifiers with their Confusion Matrices for the status of sputum culture conversion

Sputum Culture Status	***Predicted (→)***		***TP(%)***	***FP(%)***	***ROC Area***	***Precision%***
Actual (↓)	Converted **(Negative)**	Not Converted **(Positive)**				
Naive Bayes						
Converted**(Negative)**	**TN**=1055	**FP**=7	99.3	94.3	71.8	86.5
Not Converted**(Positive)**	**FN**=165	**TP**=10	5.7	7.0	71.8	58.8
Weighted Average			86.1	81.0	71.8	**82.6**
Multilayer Perceptron						
Converted**(Negative)**	**TN**=1022	**FP**=40	96.2	86.3	64.6	87.1
Not Converted**(Positive)**	**FN**=151	**TP**=24	13.7	3.8	64.6	37.5
Weighted Average			84.6	74.6	64.6	**80.1**
J48						
Converted**(Negative)**	**TN**=1058	**FP**=4	99.6	99.4	52.4	85.9
Not Converted**(Positive)**	**FN**=174	**TP**=1	6.0	4.0	52.4	20.0
Weighted Average			85.6	85.4	52.4	**76.6**
ID3						
Converted**(Negative)**	**TN**=1029	**FP**=33	96.9	89.1	61.6	86.8

Not Converted**(Positive)**	**FN**=156	**TP**=19	10.9	3.1	61.6	36.5
Weighted Average			84.7	77.0	61.6	**79.7**

Table 3 demonstrates precision based on sputum culture conversion status through confusion matrices. The confusion matrix in predictive analytics is a table with two rows and two columns that reports the number of false positives (FP), false negatives (FN), true positives (TP), and true negatives (TN). Stehman and Stephen (1997) have expressed as the confusion matrix is also known as a contingency table or an error matrix and this is also a specific table layout that allows visualization of the performance of an algorithm, typically in a supervised learning scenario. The row of confusion matrix indicates the true class and the column indicates the classifier construction. Each entry, subsequently gives the number of instances of row that were classified as column. This was conveyed equivalently by Powers and David (2011) as each column of the matrix represents the instances in a predicted class whereas each row represents the instances in an actual class. In this table all "negative" cases have real conversion of their sputum culture status and "not converted" cases are still positive throughout the treatment period. *TP* is the proportion of positive cases that are correctly identified and TN is the proportion of negative cases that are correctly identified. FP and FN are also called type I error and type II error respectively. TP (%) and FP (%) are the rates of TP and FP respectively. ROC examines the performance of classifiers (Swets, 1988) as an additional way besides the confusion matrix. It examines with the false positive rate and the true positive rate. A non-parametric classifier is represented by a single ROC point, corresponding to its pair of (*FP*,*TP*). The point (0,1) is the perfect classifier: that classifies all positive cases and negative cases correctly and the point (1,0) is the classifier that is incorrect for all classifications. Each parameter setting provides a (*FP*, *TP*) pair and a series of such pairs can be used to plot an ROC curve. The higher the rate or ROC area is better the performance. Precision or positive predictive value is an assessor in the confusion matrix. The higher the precision is the better the option. "Weighted Average" is weighting the results based on the sample sizes for each class. Naive Bayes and Multilayer Perceptran are identified as the better classifier based on the percentage of precision (82.6% and 80.1% respectively) and ROC area (71.8 and 64.6 respectively) on the criteria of "Weighted Average" approach.

IV. Conclusions

In this manuscript we studied the performance of four different classifiers. The study is done with a dataset of patients infected with tuberculosis; we got different results for each classifier. Experiments performed mainly to identify the best classifier for predicting the patient with tuberculosis. These classifiers for instance Naive Bayes, ID3, J48 and MLP were used for providing the incomparable impression as well as to identify relatively appropriate classifiers among the four. In which Naive Bayes and MLP classifiers are performing well especially in correctly classifying the instance for sputum culture conversion. It is also observed that these two approaches perform better in many ways when compared to other two methods. In the aspects of correctly classified instances, accuracy, precision and area of ROC are the evidence for identifying better performance among classifiers. According to these criteria, Naive Bayes and MLP are clearly defined their consistency as well as individuality in all aspects. However, the Naive Bayes and Multilayer Perceptron classifiers are remarkably effective and showing a good performance for this dataset. We propose these two approaches are the better to predict the performance based on sputum culture conversion among patients with tuberculosis by means of data mining classification technique.

References

[1]. Quinlan J.R., (1986).Induction of Decision Trees, Machine Learning, pp81-106.
[2]. Bishop, C. Neural Networks for Pattern Recognition. Oxford University Press.1995
[3]. Sarle, W. Neural Network Frequently Asked Questions. Available from ftp://ftp.sas.com/pub/neural/FAQ.html.(2003)
[4]. Quinlan, J. R. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, 1993.
[5]. Kotsiantis S.B., Supervised Machine Learning: A Review of Classification Techniques, Informatica 31(2007) 249-268, 2007
[6]. Weka 3: Data mining with open source machine learning software in java. http://www.cs.waikato.ac.nz/ ml/weka/.
[7]. Fawcett, Tom (2006). "An Introduction to ROC Analysis". Pattern Recognition Letters **27** (8): 861 – 874. doi:10.1016/j.patrec.2005.10.010.
[8]. Powers, David M. W. (2011). "Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation" (PDF). Journal of Machine Learning Technologies **2** (1): 37–63.
[9]. Stehman, Stephen V. (1997). "Selecting and interpreting measures of thematic classification accuracy". Remote Sensing of Environment **62** (1): 77–89. doi:10.1016/S0034-4257(97)00083-7.
[10]. Swets, J. (1988). Measuring the accuracy of diagnostic systems. Science, 240, 1285–1293
[11]. Han. J. and Kamber .M., "Data Mining: Concepts and Techniques", Morgan Kaufmann, 2nd , 2006
[12]. Witten I. H., Frank E., and Hall M. A., Data Mining: Practical Machine Learning Tools and Techniques, 3rd ed. Morgan Kaufmann, 2011
[13]. Lu H., Setiono R., and Liu H., "Effective Data Mining Using Neural Networks", IEEE, 1996
[14]. Michie D., Spiegelhalter D.J., and Taylor C.C., "Machine Learning, Neural and Statistical Classification", Ellis Horwood Series in Artificial Intelligence, 1994
[15]. Haiyang Z., "A Short Introduction to Data Mining and Its Applications", IEEE, 2011

[16]. Caruana, R.; Niculescu-Mizil, A. (2006). "An empirical comparison of supervised learning algorithms". Proceedings of the 23rd international conference on Machine learning, pp. 161-168, 2006
[17]. Bermejo P., Gámez J. A., Puerta J.M., "Speeding up incremental wrapper feature subset selection with Naive Bayes classifier", Knowledge Based Systems, vol. 55, pp. 140–147, 2014
[18]. Wang W. A tutorial in WEKA. Data Mining & Statistics within the Health Services, University of East Anglia; 2010
[19]. Provost, F., & Fawcett, T. (2001). Robust classification for imprecise environments. Machine Learning, 42 (3), 203–231
[20]. Provost, F., Fawcett, T., & Kohavi, R. (1998). The case against accuracy estimation for comparing induction algorithms. In Shavlik, J. (Ed.), Proceedings of the Fifteenth International Conference on Machine Learning, pp. 445–453 San Francisco, CA. Morgan Kaufmann