Volume 1, No. 5, December 2010

Journal of Global Research in Computer Science

ISSN-2229-371X

RESEARCH PAPER

Available Online at www.jgrcs.info

ERRORS OF SUPERVISED CLASSIFICATION TECHNIQUES ON REAL WORLD PROBLEMS

J. Ashok Kumar*[1] and P.R. Rao[2]

*[1]Central Institute of Brackishwater Aquaculture /Chennai, Tamilnadu, India
akjangam@yahoo.com[1]
[2]Department of Computer Science and Technology /Goa University/Taleigao Plateau, Goa, India
pralhadrrao@gmail.com[2]

Abstract: In supervised learning, classifiers are trained with data consisting class labels to solve real-world classification problems. Decision trees, random forest, naïve Bayes, Bayesian networks, K- Nearest neighbourhood logistic regression, artificial neural networks and support vector machines are some of the most popular classification techniques among the class of popular classifiers available for researchers. Although each one of these techniques has their own strengths in dealing varied real-world problems, they have inherent problems too. No classification technique can be universally applied for all real world applications. In the past, several researchers tried to understand the behaviour of these techniques by applying to different areas of research. In the present paper, we experimented with training and test datasets used by some of the researchers to get better understanding on the behaviour of the above classifiers. Study reveals the inadequacies of some of the techniques and superiority of support vector machines and logistic regression over other tools used.

Keywords: classification, errors, supervised learning

INTRODUCTION

In machine learning, supervised learning builds the model from labeled training data. These models often called as pattern classifiers are used to predict the class labels with attribute information of a given object. This art of classifying an object into one of the predefined classes is called classification [1-3]. Different steps involved in supervised learning from problem statement to classifier design include identification of data, data preprocessing, and definition of training set, algorithm selection, training and evaluation with test set [4]. If test set results are unsatisfactory we must go back to training step by tuning the parameters of the algorithm. Corrections are even required at data preprocessing, definition of training data and algorithm selection for obtaining satisfactory results.

THE INDICATOR FOR CLASSIFIER PERFORMANCE IS CLASSIFICATION ACCURACY, WHICH IS RATIO BETWEEN NUMBER OF CORRECTLY CLASSIFIED INSTANCES AND THE TOTAL NUMBER OF INSTANCES EXPRESSED IN PERCENTAGE [5].

$$A_i = \frac{t}{n}.100$$

A_i = classification accuracy; t = number of correctly classified instances; n = total number of input instances

Classification error is proportion of number of incorrectly classified instances in the total number of instances expressed in percentage.

$$E_i = \frac{f}{n}.100$$

E_i = classification error; f = number of incorrectly classified instances; n = total number of input instances

Ideally the classification errors should be as low as possible for training and testing data. But many a times classifiers which perform with high accuracy on training data fail to predict test data with the same accuracy. Receiver operating characteristic curve (ROC) is used as performance measure for graphical representation of classifier accuracy [6]. One of the main reasons for this is model over fitting [7] during training. Hence the models with less generalization errors on both training and test data are preferred.

There are several reasons for errors of classification. Sampling errors of training and test data, inherent problems with chosen classification algorithms and model parameters are some important factors that results in high classification errors. In the present study we tried understand the inherent problems of most popular classification algorithms available today.

Research efforts in recent time led to development of an array of classification algorithms which are capable of handling different real-world classification applications. Decision trees, random forest, naïve bayes method, Bayesian networks, logistic regression, artificial neural networks and support vector machines are the popular classifiers in use for several research works. No single classifier is universally applicable for all real world classification problems. Some classifiers are capable of handling some problems and fail in others. One of the major research areas in design and analysis of classification algorithms is to find the applicability and evaluation of different algorithms to real world problems. The present study is intended to provide in depth understanding on the behavior of most popular classification algorithms of the day. We also provide brief overview on the capabilities of different classifiers for the benefit of reader.

REVIEW OF CLASSIFICATION TECHNIQUES

Decision Trees

Decision trees are most successful tool for classification and prediction [8]. The strength of the decision trees is easy interpretability and performance comparable with other state of art classifiers. Most popular algorithm of decision tree is C4.5 [9]. Decision tree model starts at the root node and branches pass through internal nodes towards terminal nodes. The terminal nodes called as leaf nodes represent different classes. Decision tree induction is a typical inductive approach to learn knowledge on classification. The principal requirements of decision tree are attribute value description, predefined classes, classes that are not predefined, and large number of instances

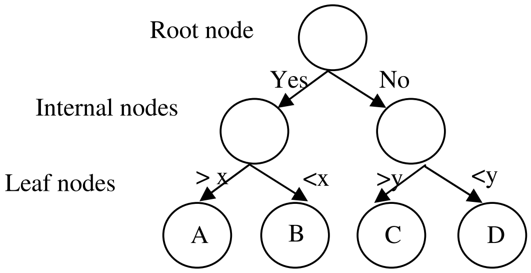

Figure 1: Schematic representation of decision tree with 4 classes namely A, B, C, D.

The key issues in decision tree are to determine how to find the best attributes for splitting, how to split and when to stop splitting. Splitting criteria depends on data type of attributes (nominal, ordinal, continuous) and number of ways to split (two way/multiway). Best split criteria are decided by using measures of impurities like gini index, entropy and misclassification errors [10-12]. Different indices for a given node r are

$$Gini(r) = 1 - \sum_i [p(i \mid r)]^2$$

$$Entropy(r) = -\sum_i p(i \mid r) \log p(i \mid r)$$

$$Error(r) = 1 - \max p(i \mid r)$$

P (i|r) is probability of i^th^ class at node r.

Nodes with homogeneous class distribution are preferred over heterogeneous distribution. Measures of impurities are on higher side (~0.5) when the instances are equally distributed for a splitting criteria and lowest (~0) when all instances belong to a single class.

The strengths of decision trees are easy interpretability, high speed classification, less computational requirement, ability to handle nominal, ordinal and continuous data, ability to indicate which attributes are more important for classification.

Weaknesses of the method include not suitable for parameter estimation/prediction, huge data requirement and erroneous results with less number of training instances, complex as the size of the tree grows, less tolerant to noise.

RANDOM FOREST

Random Forest is an ensemble classification method of multiple decision trees [13]. Each tree of random forest is constructed independently using bootstrap sample of training instances.

Randomly selected predictors are used for splitting criterion at each node by random subspace selection. Merits of decision tree include its suitability for different data types, robustness to missing observations and model over fitting, ability to handle large number of attributes, easy parameter handing. Computational speed is the main drawback of this method

BAYESIAN CLASSIFIERS

Bayesian classifiers predict the probability of a given instance belonging to a specified class. These classifiers are based on bayes theorem [14] on conditional probabilities. Bayes theorem relates conditional probabilities of events A and B

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

P(A) is probability of event A. P(B) is probability of event B. P(A|B) is conditional probability of event A give event B and P(B|A) is conditional probability of event B give event A. There are several algorithms developed and used for addressing classification problems using bayes theorem. Naïve bayes and bayes networks are discussed here under

NAÏVE BAYES

Naive Bayesian classifiers [15] assume the independency of the attribute values. The conditional independence stated mathematically as follows

$$P(X \mid Y = y) = \prod_{i=1}^{n} P(X_i \mid Y = y)$$

Attribute set X={X_1, X_2….X_n} consists n elements.

Naïve bayes classifier instead of calculating each combination of class conditional probabilities it calculates conditional probability of X_i given Y making it more practical as it does not require large training set to get reasonably good estimate. Naïve bayes classifier calculates posterior probabilities of class Y during testing of input instances.

$$P(Y \mid X) = \frac{P(Y)\prod_{i=1}^{n} P(X_i \mid Y)}{P(X)}$$

Naïve bayes models are faster in learning and classification. They are also robust to missing values and easy to interpret. But they are less accurate while dealing real-world problems and less tolerant redundancy in input instances.

BAYESIAN NETWORKS

Bayesian networks [16] are probabilistic graphical models (GMs). Each node in a graph represents a random attribute and edges represent dependencies between them. These conditional dependencies in the graph are estimated through probability theory. Bayesian networks does not assume class conditional independence of all variables under consideration for classification task like naïve bayes and allows to specify which variables are conditionally independent.

Direct acyclic graph encoding relationships among attributes and associated probability table of each node to its parent are key components of Bayesian networks. From the direct acyclic graph it is known that an attribute is independent from other or not. Once the topology of attributes and their relations are known probabilities are calculated based on dependence or independence of the attributes.

Bayesian networks are robust to missing data, model overfitting. But the network building is time consuming process.

K-NEAREST NEIGHBORHOOD CLASSIFIERS

In Nearest neighborhood method [17], classification is based directly on training instances. As the training instances are needed at runtime, these classifiers are also called memory based classifiers.

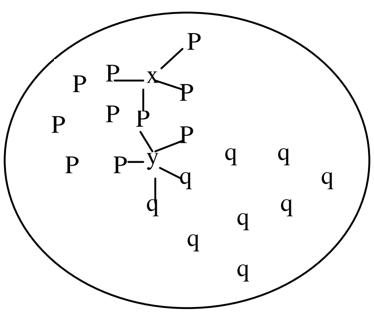

Fig. 2. Illustration of nearest neighborhood classification

Figure 5 illustrates two class classification problem with p and q as class labels. Object x will be given label directly as p as all its nearest neighbors belong to class p. But object y class will be decided based on majority vote or distance weighted vote by nearest neighbors. Hence k-nearest neighborhood classification involves identifying the nearest neighbors and predicting the class based on the neighbors.

K-nearest neighbor method advantages include easy interpretability, fast learning and easy parameter handling. On the flip side this method fails in many real-world classification problems, slow in classification not robust to missing /redundant attributes and noise.

LOGISTIC REGRESSION

In linear least square regression the dependent variable should necessarily be continuous. When the dependant variable is binary (1, 0/ yes, no) or ordinal (low, medium, high) logistic regression [18] is the choice. When applied to a classification problem logistic regression used to predict the class with the help of explanatory variables. Model building using logistic regression is an iterative process. Logistic regression is widely used technique for categorical data analysis having many applications in business, genetics etc[14]. With respect to classification it can be applied for two class or multiclass problems. Logistic regression does not predict the class directly. It only predicts the log odds, the ratio of the probability that an event occurs to the probability that it fails to occur which will be considered as an indicator for particular class based on threshold values set. Hence, the dependent variable of Logistic regression is

$$\ln(\frac{P}{1-P})$$

and the logit model is

$$\ln(\frac{P}{1-P}) = a + b_1 x_1 + b_2 x_2$$

Unlike the probability values, log odds ranges from negative infinity to positive infinity and symmetric around the log odds equals to zero. Figure 3 depicts logistic function curve, with log odd values in x axis (z) and equivalent function derivatives in 0-1 scale (f (z)). All the regression coefficients of linear least square regression are interpreted in the same way for logistic regression also[15]. Logistic regression is more robust as the dependent and independent variables need not be normally distributed and its interpretability is much easier than discriminant analysis and other neural methods.

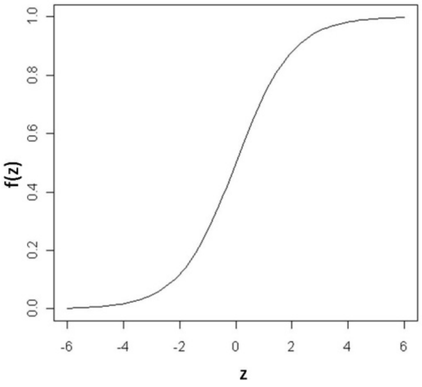

Fig 3. Logistic function curve

ARTIFICIAL NEURAL NETWORKS

Artificial neural networks (ANN) [19] are the mathematical models inspired by biological neural systems. Functions of the neurons, axons, dendrites and synapse are simulated in artificial neural networks. Neurologists have discovered that the human brain learns by changing the strength of the synaptic connection between neurons upon repeated stimulation by the same impulse. Similarly in ANN consists of interconnected assembly of nodes and direct links.

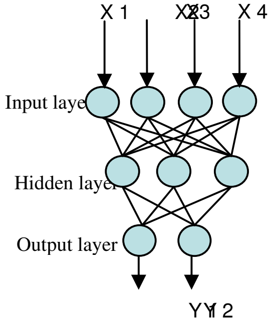

Fig. 4. Multi-layer neural network model

ANN model initiates at the input level, may contain some hidden layers and terminates at output layer. Complexity of the model depends on the number of hidden layers used. Nodes of one layer are connected to the nodes of subsequent layers in multi-layer feed forward artificial neural networks. The network may use different types of activation functions like sign, linear, sigmoid to produce the output from nodes of hidden and output layers. ANNs are powerful nonlinear classifiers and the models are fast to run. They can efficiently handle the redundant data. Training neural network classifiers is a time consuming process and complexity grows with the number of hidden layers. These models are also sensitive to noisy data.

SUPPORT VECTOR MACHINES (SVM) CLASSIFIERS

Support Vector Machine Classifiers [20] gained lot of importance during the recent past as this can be used in various applications. SVM can handle high-dimensional data very efficiently. Another unique aspect of this approach is that it represents the decision boundary using a subset of the training examples, known as the support vectors[21].

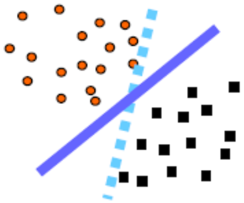

Fig. 5a. Two possible linear discriminant planes

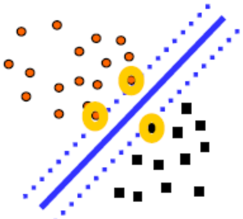

Fig. 5b. Best plane maximizes the margin

Support vector machines map input vectors to a higher dimensional space where a maximal separating hyperplane is constructed. Two parallel hyperplanes are constructed on each side of the hyperplane that separates the data. The separating hyperplane is the hyperplane that maximizes the distance between the two parallel hyperplanes. Generalization error of the classifier is better when the distance between these hyperplanes is larger. SVM classifiers are the extremely powerful non-linear classifiers with high accuracy and classification speed. These are also sensitive to noise in the data, complex to train and are prone to overfitting.

MATERIAL AND METHODS

Different classifiers like decision trees, random forest, naïve bayes method, bayesian networks, logistic regression, artificial neural networks and support vector machines obtained from open source software WEKA [21], used for the present experimental study. Training data sets used by in [22] are given in table no 1.

Table 1. Data used for training the classifiers

Class	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
C1	2	2	2	2	2	2	2	2
C2	4	4	4	4	4	4	4	4
C3	6	6	6	6	6	6	6	6
C4	8	8	8	8	8	8	8	8
C5	10	10	10	10	10	10	10	10

In [22], authors have used different test conditions which are more appropriate for artificial neural networks. Out of those test conditions we have taken four test conditions which are applicable for all the classificatory techniques.
Test condition 1. Positional changes in the attributes in input data causes classification errors
Test condition 2. Large input values results in loosing the ability to classify
Test condition 3. Classifiers can not discriminate sharply where there is significant difference in only one attribute value.
Test condition 4. Bigger the attribute value smaller the class it belongs to.
All the above test conditions are applied in [22] only for back propagation network model. In the present work we applied these test conditions for all the classifiers under the study for in depth understanding of the techniques in use.

RESULTS AND DISCUSSION

Different test conditions used in literature where classification errors are reported are taken to study the behavior of different classifiers. The classifiers used are Naïve bayes (NB), Bayesian networks (BN), Decision trees (DT), random forest (RF), logistic regression (LR), K-nearest neighborhood (KNN), artificial neural networks (ANN) and support vector machines (SVM). The training set contains five instances with eight attributes, each instance belongs one class (say c1 to c5). Among the classifiers used, Bayesian networks and Decision trees produced errors in training. Bayes networks recognized all five classes as class one (c1), whereas Decision trees recognized first two classes as class one (c1) and rest three classes as class three (c3). All other classifiers recognized training samples with 100 percent accuracy i.e, from class one to class five each instance belongs to one class. Hence Bayesian networks and Decision Tree algorithms were deleted from the experiment.

Test condition 1: The change of positions of attributes in input data causes the possible classification changes.

i1 = (4; 4; 4; 4; 10; 10; 10; 10)
i2 =(10; 10; 10; 10; 4; 4; 4; 4)

Literature says some of the classifiers cannot recognize the positional changes in the attributes. Hence they predict the instances with same attributes as different classes due to changes in positions.

Table 2. Variation in classifier prediction with positional changes in attribute values

NB	RF	LR	KNN	ANN	SVM
2-c4	1-c2 1-c5	2-c4	1-c2 1-c5	2-c4	2-c3

As per our experimental results this is true for Random forests and nearest neighborhood methods. Other classifiers used could recognize the positional changes in attribute values. Then we extended this test with all possible combinations of the 4 fours and 4 tens. There will be total 70 instances with all combinations with positional changes with values 4 fours and 4 tens. In this case even artificial neural networks lost its ability to recognize the positional variation among input instances.

Table 3. Performance with all possible combinations of positional changes in input instances.

NB	RF	LR	KNN	ANN	SVM
70-c4	25-c2 10-c3 35-c5	70-c4	70-c3	4-c3 66-c4	70-c3

Test condition 2. With large values in input instances classifiers losses its ability of classification.

i3 = (2; 2; 2; 100; 100; 100; 100; 100)

It is evident from the input values that the instance belongs to a higher class among the trained classes. But Naïve bayes, Random forest and nearest neighborhood models lost their ability to recognize the large unknown values so predicted as lower classes (c1 and c2). Logistic regression, artificial neural networks and Support vector machines proved to be efficient in recognizing the large values and produced better output.

Table 4. Performance of classifiers over large unknown values in input instance.

NB	RF	LR	KNN	ANN	SVM
1- c1	1-c1	1-c5	1-c2	1-c5	1-c5

Test condition 3. It is difficult for classifiers to discriminate sharply.

i4 =(6; 4; 4; 4; 4; 4; 4; 4)
i5 =(60; 4; 4; 4; 4; 4; 4; 4)

In the given instances s7 and s8 there is a huge difference in the first attribute value. But Naïve bayes and Nearest neighborhood algorithms could not recognize the difference and failed to classify the instances correctly. All other classifiers could recognize the difference in first attribute value and classified them into two different classes.

Table 5. Classifier performance over extreme attribute values

NB	RF	LR	KNN	ANN	SVM
2-c2	1-c2 1-c3	1-c2 1-c5	2- C2	1-c2 1-c5	1-c2 1-c5

Test condition 4. The bigger the attribute's value is, the lower the class to that it belongs.

i6 = (10; 5; 2; 2; 2; 2; 2; 2)
i7 = (30; 5; 2; 2; 2; 2; 2; 2)

In this test condition also naïve bayes, random forest and nearest neighborhood algorithms did could not discriminate difference in the input instances i6 and i7. And it is found none of the classifiers are in accordance with the test condition. All the classification algorithms used are capable of placing lower values in lower classes and higher values in higher classes or at least both in the same class.

Table 6: Performance over lower / higher attribute values.

NB	RF	LR	KNN	ANN	SVM
2-c1	2-c2	1-c2 1-c3	2-c2	1-c1 1-c3	1-c2 1-c3

Logistic regression and support vector machines have shown better performance over all the real world conditions tested in the experiment. Artificial neural networks are almost at par with the best classifiers but failed to recognize the positional changes in the values input instances. There are some serious concerns with the Naïve bayes, random forest and K-nearest neighborhood algorithms with regard to real world classification problems which need to be further researched and improved.

CONCLUSIONS

Classification is an art of classifying an object into one of several predefined classes. In supervised learning input instances along with the class labels are trained for model building. Later the model is used for predicting the class of unknown attribute values. Several classification algorithms are available these days for addressing classification problems. Each one of these classifiers has its own advantage while applying to real-world classification problems. As there is no classifier available which can universally applied to all the real world problems, it is important to evaluate classifiers and find suitable one for a given problem. We have taken some test conditions from previous research works and evaluated with some popular classification algorithms of recent times to have an in depth understanding of the classifiers. Out of the classifiers used in the study Bayesian networks and Decision tree algorithms failed even at model building (training) stage hence, deleted from the experiment. Naïve Bayes, Random forest and K-Nearest neighborhood have failed in addressing the test conditions applied in the study. Artificial neural networks have shown better performance with regard to test conditions except for one. Support Vector Machine and Logistic regression which are considered to be state of art models outperformed all the models used in the study. The present work provides in depth understanding on strengths and weaknesses of the different classifiers which need to be addressed in future research works.

REFERENCES

[1] Tan, P.N., Steinbach, M. and Kumar, V., Introduction to datamining. Addison-wesley publ. Co., 2006.

[2] Han, J. and Kamber, M., Data mining : Concepts and Techniques. 2nd ed., Elsevier publ., 2006.

[3] Pedro, L., Borja, C., Roberto S., Concha B., Josu G., Iñaki I., José, A. L., Rubén A., Guzmán, S., Aritz, P., Victor, R., Machine learning in bioinformatics.

Briefings in bioinformatics, vol 7, no. 1, 2005, 86-112.

[4] S. B. Kotsiantis, Supervised Machine Learning: A Review of Classification Techniques, Informatica 31 (2007) 249–268

[5] Pierre Baldi, Soren Brunak, Yves Chauvin, Claus A.F. Andersen and Henrik Nielsen, Assessing the accuracy of prediction algorithms for classification : an over view, Bioinformatics, Vol 16(5), 2000, 412-424

[6] Bradley AP. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognition 1997, 30(7):1145-59

[7] Gavin C. Cawley and Nicola L. C. Talbot, On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation, Journal of Machine Learning Research 11, 2010, 2079-2107

[8] Murthy, (1998), Automatic Construction of Decision Trees from Data: A Multi-Disciplinary Survey, *Data Mining and Knowledge Discovery* 2: 345–389.

[9] Quinlan, J. R. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, 1993.

[10] Breiman L., Friedman J.H., Olshen R.A., Stone C.J. (1984) Classification and Regression Trees, Wadsforth International Group

[11] Hunt E., Martin J & Stone P. (1966), Experiments in Induction, New York, Academic Press

[12] Jaynes, E.T. "Information Theory and Statistical Mechanics". *Physical Review* 1957, 106 (4): 620–630.

[13] Breiman, Leo. Random Forests. *Machine Learning, 2001* 45 (1): 5–32.

[14] Bradley P. Carlin and Thomas A. Louis, Bayesian Methods for Data Analysis, Third Edition, 2008, CRC press,

[15] Harry Zhang, The Optimality of Naive Bayes. The Florida Artificial Intelligence Research Society 2004 conference,

[16] Ben-Gal Irad Bayesian Networks. in Ruggeri, Fabrizio; Kennett, Ron S.; Faltin, Frederick W. *Encyclopedia of Statistics in Quality and Reliability*. 2007. John Wiley & Sons.

[17] Cover TM, Hart PE. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory* 1967, 13 (1): 21–27

[18] Agresti A. "Building and applying logistic regression models". *An Introduction to Categorical Data Analysis*. Hoboken, New Jersey: Wiley.2007. p. 138

[19] Kishan Mehrotra, Chilukuri K. Mohan and Sanjay Ranka Elements of Artificial Neural Networks, 1996, MIT press

[20] Corinna Cortes, Vladimir Vapnik Support-Vector Networks. Machine Leaming, 1995, 20, 273-297

[21] Witten, I. and Frank, E., Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann at http://www.cs.waikato.ac.nz/ml/weka/book.html, 2005.

[22] Lihua Feng and Weihu Hong (2009), Classification error of multilayer perceptron neural networks, Neural Computing & Applications 18:377–380