International Journal of Power Control Signal and Computation(IJPCSC)
Vol3. No1. Jan-Mar 2011 pp:123-126
ISSN: 0976-268X
www.ijcns.com

Comparison Study of Classification Algorithms

N.Hema [1],.Jha[2]
[1] Assistant Professor, Dept. Of Computer Science,
[2] Professor,Department of Computer Science,
[1]Vivekanandha College Of Arts And Sciences For Women, Elayampalayam, Tiruchengodu(Tk), Namakkal (Dt.) Tamilnadu, India.
[2]M.L.S.M College,Darbanga, Patna. India

Abstract:- Classification is the process of arranging a number of items into groups in such a manner that the members of the group have one or more characteristics in common. In this research paper, we propose to compare the accuracy of various AI and statistical methods on several classification tasks. There may be generalizations that can be drawn about the types of data sets for which certain methods are most appropriate. Statistical methods used in the comparison will include decision trees (CART, CHAID, and QUEST), discriminant analysis, and logistic regression. AI approaches will include various multi-layer perception neural networks, learning vector quantization (LVQ) neural networks and other related supervised learning methods. Real world datasets are used with the idea that good performance on them will generalize to similar performance on other real-world tasks. The main aim of this paper is to make a comparison of different classification algorithms and to find out the best algorithm which gives the most accurate result.

Keywords: CART, CHAID,QUEST,LVQ, Supervised Learning ,AI

1. INTRODUCTION

Data mining (DM) is the process of analyzing data from different angles and summarizing it into useful information that can be used for making intelligent business decision. It is not specific to any industry, applied in almost all areas to explore the possibility of hidden knowledge.

II. COMPARING THE ALGORITHMS

Classification as we shall use it in this chapter refers to establishing rules so that we can classify new observations into one of a set of existing classes. Observations have attributes. The task of the classifier is to assign an observation to a class given its set of attributes. The rules may be explicit or comprehensible, as in the case of decision trees. Or, as with neural networks, rules may not be capable of explicit formulation.

It is assumed that we have a number of sample observations from each class. The classifier is presented with a substantial set of the data from which it can associate known classes with attributes of the observations. This is known as training. When such guidance is given the process is known as supervised learning. The rules developed in the training process are tested on the remaining portion of the data and compared with the known classifications. This is known as the testing process. Here the response of the procedure to new observations is a prediction of the class to which the new observations belong. The proportion correct in the test set is an unbiased estimate of the accuracy of the rules implemented by the classifier.

III. DISCRIMINANT ANALYSIS

Discriminant analysis is the oldest statistical technique for classification. R.A. Fisher first published it in 1936. In it the difference between two classes is maximized by a linear combination of variables. This linear function acts as a hyper plane that partitions the observation space into classes. Which side of a hyper plane a point falls into determines its classification.

Discriminant analysis assumes that the predictor variables are normally distributed. We will use the implementation of discrminant analysis provided in SPSS Version 8.0.

LOGISTIC REGRESSION

Logistic regression is a version of linear regression used for predicting a classifying variable. Logistic regression builds up a linear model using the logarithm of the odds of occurrence of a class membership. In logistic regression the modeler must select the right variables and account for their possible interactions. There is no normality assumption imposed upon the data. We will use the implementation of logistic regression provided in SPSS Version 8.0.

DECISION TREES (CART,CHAID,QUEST)

Decision trees develop a series of rules that classify observations. We will use three types - CART (known as "C&RT" in SPSS's version), CHAID, and QUEST. In all decision trees an observation enters at the root node. A test is applied which is designed to best separate the observations into classes. This is referred to as making the groups "purer." The observation then passes along to the next node. The process of testing the observations to split them into classes continues until the observation reaches a leaf node.

Observations reaching a particular leaf node are classified the same way. Many leaves may make the same classification but they do so for different reasons. Decision trees differ from the classical statistical tests in that they do not draw lines through the data space to classify observations. Decision trees may be thought of as drawing boxes around similar observations. Several different paths may be followed for an observation to become part of a particular class. Criticisms of decision trees include that any decision on how to split at a node is made "locally." It does not take into account the effect the split may have on future splits. And the splits are "hard splits" that often may not reflect reality. Thus an attribute "years of age" may be split at "age > 40." Is someone thirty-nine so different than a forty-one year old? Also, splits are made considering only one attribute at a time (Two Crows Corporation, 1998).

Brieman, Friedman, Olshen, and Stone developed the CART algorithm in 1984. It builds a binary tree. Observations are split at each node by a function on one attribute. The split is selected which divides the observations at a node into subgroups in which a single class most predominates. When no split can be found that increases the class specificity at a node the tree has reached a leaf node. When all observations are in leaf nodes the tree has stopped growing. Each leaf can then be assigned a class and an error rate (not every observation in a leaf node is of the same class). Because the later splits have smaller and less representative samples to work with they may over fit the data. Therefore, the tree may be cut back to a size which allows effective generalization to new data. Branches of the tree that do not enhance predictive classification accuracy are eliminated in a process known as "pruning."

CHAID differs from CART in that it stops growing a tree before over fitting occurs. When no more splits are available that lead to a statistically significant improvement in classification the tree stops growing. Also, any continuously valued attributes must be redone as categorical variables. The implementations of CART and CHAID we will use are from SPSS's Answer Tree Version 2.0.

QUEST is another type of decision tree developed by Loh and Shih (1997). It is unique in that it performs approximately unbiased as to class membership variable selection to split nodes. We will use the implementation of QUEST with linear combination splits available from http://www.stat.wisc.edu/~loh/quest.html.

III. ASSESSING CLASSIFICATION TOOL PERFORMANCE

While we seek to determine the fitness of each algorithm the results obtained when a technique is applied to data may depend upon other factors. These include the implementation of the technique as a computer program and the skill of the user in getting the best out of the technique.

We will use several metrics to assess the performance of classification tools. The first is the traditional one of percentage of cases in the test set incorrectly classified (mean error rate). We will average this number across all datasets to give us a measure of a classifier's overall effectiveness. We will also examine the ranks of the classifiers within datasets. The classifiers with the lowest error rate will be assigned a rank of one, the one with the second lowest error rate will be assigned a rank of two, etc. The average ranks will be assigned in the case of ties.It has been shown that there are problems with using accuracy of classification estimation as a method of comparing algorithms (Provost, Fawcett, and Kohavi, 1998). It assumes that the classes are distributed in a constant and relatively balanced fashion. But class distributions may be skewed. For example, if your classification task is screening for a rare disease, calling all cases "negative" can lead to a spuriously and trivially high accuracy rate. If only .1 percent of patients has the disease a test that says no one has the disease will be correct 99.9% of the time. Accuracy percentage is affected by prevalence rates and there is no mathematical way to compensate for this.

Accuracy is also of limited usefulness as an index of a classifier's performance because it is insensitive to the types of errors made. Using classification accuracy as a measure assumes equal misclassification costs - a false positive has the same significance as a false negative. This assumption is rarely valid in real-world classification tasks. For example, one medical test may have as its mistakes almost all false negatives (misses). Another might err in the direction of false positives (false alarms). Yet these two tests can yield equal percentages of correctly classified cases. If the disease detected by the test is a deadly one a false negative may be much more serious than a false positive. Similarly, if the task is classifying credit card transactions as fraudulent the cost of misclassifying a transaction as fraudulent (false alarm) may be much less than missing a case of fraud.

The limitations of using classification accuracy can be overcome by an approach known as receiver operating characteristic (ROC) analysis (Metz, 1978; Swets, 1973). This is the second metric we shall use to evaluate classifier performance. We can begin our look at it by defining decision performance in terms of four categories:

True Positive Decisions = True Positive Fraction
Positive cases (TPF) Actually

False Positive Decisions = False Positive Fraction
Negative cases (FPF) Actually

True Negative Decisions = True Positive Fraction
Negative cases (TNF) Actually

False Negative Decisions= False Positive Fraction
Positive cases (FNF) Actually

Since all observations are classified as either positive or negative with respect to membership in a class the number of correct decisions plus the number of incorrect decisions equals the number of observations in that class. Thus, the above fractions are related by:

TPF+FNP=1 and TNF+FPF=1

ROC curve must be above the lower left to upper right diagonal. When this is so a decision to place an observation in a class when it actually is a member of that class is more probable. A ROC curve illustrates the tradeoffs that can be made between TPF and FPF (and hence all four of the decision fractions).

ROC analysis gives us another perspective on the performance of classifiers. An ROC curve shows the performance of a classifier across a range of possible threshold values. The area under the ROC curve is an important metric for evaluating classifiers because it is the average sensitivity across all possible specificities. One point in ROC space is better if it is to the upper left in the ROC chart. This means TPF is higher; FPF is lower, or both. A ROC graph permits an informal visual comparison of classifiers. If a classifier's ROC curve is shifted to the upper left across all decision thresholds it will perform better under all decision cutoffs. However, if the ROC curves cross then no classifier is best under all scenarios. There would then exist scenarios for which the model giving the highest percentage correctly classified does not have the minimum cost. The computer program we will use for figuring ROC curves was developed by Charles Metz, Ph.D. of the Department of Radiology at the University of Chicago (Metz, 1998).

Bradley (1997) investigated the use of the area under the ROC curve (AUC) as a measure of a classification algorithm's performance. He compared six learning algorithms on six real-world medical datasets using AUC and conventional overall accuracy. AUC showed increased sensitivity (a larger F value) in analysis of variance

DATA

Contraceptive Method Choice The data consists of nine demographic attributes for 1,473 married women. The data is modified slightly from the original dataset to include two classifications - does or does not use contraception.

EXPERIMENTAL PROCEDURE

Eighty percent of each dataset will be used for training the algorithms and twenty percent will be held back as a test set. For the back propagation, cascade correlation, and Levenberg-Marquardt neural networks ten percent of the training data (8 percent of the total) will be put into a file used to prevent overtraining (Masters, 1993).

Assignment of data to training, overtrading prevention, and trusting files will be randomized. Three hidden layers were used with the back propagation neural network and two with the Levenberg-Marquardt network to insure the ability to model complex relationships. Training of these neural networks stops when the error level on overtraining prevention file passed through the neural net model reaches its minimum and no improvement occurs for 10,000 iterations for back propagation networks. Tuning discriminant analysis using the stepwise technique to remove non-contributory variables was not done because this might have given an advantage over the other methods.Performance on the test sets using percentage accurately classified and ROC analysis forms the basis for comparing the algorithms.

RESULTS

Methods	Error Rate	Rank
Discriminant Analysis	37.97	9
CART	28.47	1
CHAID	30.85	5
Logistic Regression	34.76	8
QUEST	65.08	13

The attempt to increase classification accuracy by first clustering the training and testing data, and then developing and testing the classification model within the clusters failed. It was no more accurate than just developing one model by training the algorithm on all the data. Possibly clustering methods other than Ward's method could be tried. And it may be that this approach will work on datasets other than those included in the present study.

The error rates within clusters in training sets are highly predictive of error rates for those clusters in testing sets. The relative rankings of the accuracy of the clusters within the training data can be used to indicate a confidence level for predictions within those clusters from new data or a testing set. Thus, if cases are classified at the four cluster level predictions on new or test set cases could be ranked from 1 (most confidence) to 4 (least confidence). This would be based upon their membership in clusters that in the training set the classification model had greater or lesser success classifying correctly. This is a new use for cluster analysis that can be explored further.

REFERENCES

1. William I. Grosky. “Managing multimedia information in database systems,” Communications of the ACM, 40(12) pp 72–80, 1997
2. Wishart, David. (1999a)Personal communication, June 19, 1999.
3. Wishart, David. (1999b)Personal communication, February, 5, 1999.
4. U niversity of M a r y l a n d . Public health i n f o r m a t i c s , University of Maryland, 2000. [http://www. phi.
5. Sh avlik, J.W., Mooney, R.J. and Towell, G.G. (1991) "Symbolic and neural learning algorithms: An experimental comparison" Machine Learning 6, 111-143.
6. Metz, C. (1978) "Basic principles of ROC analysis" -Seminars in Nuclear Medicine 8, 283-298.
7. Pesonen, E. (1997) "Is neural network better than statistical methods in diagnosis of acute appendicitis?" In: Medical Informatics Europe '97 Pappas, C., Maglaveras, N., and Scherrer, J.R. (eds.) IOS Press, Amsterdam, Netherlands.
8. Lim, T.-S. and Loh, W.-Y. and Shih, Y.-S. (in press) "A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms" Machine Learning.
9. Jo hn, George (1997) ements to the Data Mining Process Doctoral dissertation, Stanford University.
10. Ho rrace, W. and Schmidt, P. (in press) "Multiple comparisons with the best, with economic applications" Journal of Applied Econometrics