International Journal of Fuzzy Logic and Intelligent Systems, vol. 8, no. 1, March 2008 pp. 1-5

A Comparison Study of Classification Algorithms in Data Mining

Seung-Joo Lee and Sung-Hae Jun

Department of Bioinformatics & Statistics, Cheongju University, 360-764 Chungbuk, Korea
{access, shjun}@cju.ac.kr

Abstract

Generally the analytical tools of data mining have two learning types which are supervised and unsupervised learning algorithms. Classification and prediction are main analysis tools for supervised learning. In this paper, we perform a comparison study of classification algorithms in data mining. We make comparative studies between popular classification algorithms which are LDA, QDA, kernel method, K-nearest neighbor, naïve Bayesian, SVM, and CART. Also, we use almost all classification data sets of UCI machine learning repository for our experiments. According to our results, we are able to select proper algorithms for given classification data sets.

Key words : Data Mining, Supervised Learning, Classification Algorithms

1. Introduction

Supervised and unsupervised learning tools are mainly learning approaches in analytical process of data mining[6]. Unsupervised learning is based on clustering algorithms[5],[6]. Classification and prediction algorithms are typical tools for supervised learning[5],[6]. In this paper, we show empirical results of classification performances between popular classification algorithms. Many researches of classification have been performed case by case according to given training data sets[2],[3],[5],[6],[9]. So, we have a necessity the comparisons among popular tools for classification.

We consider two types for classification algorithms. They are statistical methods and machine learning algorithms. Because statistical methods demand normality assumption and machine learning algorithms do not demand the limitation.

To verify the results of empirical comparisons, we make experiments among competitive methods using UCI machine learning repository[11]. In the experiments, we use most of the classification data sets in UCI machine learning repository.

2. Research Background

Classification is to find a model that distinguishes classes of data points. Also, using this model, we are able to predict the class of point whose class is unknown[5]. In general, classification model is able to be represented in diverse types which are based on statistical methods or machine learning algorithms. Statistical models support various forms to classification, such as linear discriminant analysis(LDA), quadratic discriminant analysis(QDA), naïve Bayes(NB) classifier, and so forth[3],[4],[8]. Kernel methods, K-nearest neighbor classifier, support vector machine(SVM), and classification and regression trees(CART) are popular classification tools which are based on machine learning algorithms[1],[3],[7],[10],[12],[13],[14]. On the whole, statistical methods are needed to normality assumption about given training data. Also, these methods provide good performance when the assumption is satisfied. But, real data sets are hard to be satisfied the normality assumption. On the other hand, the classification models based on machine learning are not needed the normality assumption of given data. So, these methods are used elastically in various data sets. In this paper, we perform empirical comparison among the popular classification algorithms which are based on statistical methods or machine learning algorithms.

Manuscript received Nov. 23, 2007; revised Mar. 11, 2008.

3. Empirical Comparisons of Classification Algorithms

In this section, we introduce popular classification algorithms which are based on statistical methods and machine learning algorithms. In addition, a hierarchical structure of current classification algorithms is proposed by statistical assumption and usability.

3.1 Statistical methods for classification

3.1.1 Discriminant analysis

To explain discriminant analysis, first of all, we consider logistic regression[5]. Logistic regression is a scoring model which assigns a score to each predicted value. Score can be used to classify the points into each class[4]. The objective of logistic regression is to predict scores which can be transform binary values. In general, they are 0 and 1. To do the transformation, a threshold value is needed. Discriminant analysis supports the theory of this predictive classification. To choose one of two classes, we need a probabilistic criterion which is able to select the class with the highest probability of occurrence. Discriminant analysis model represents as linear terms by assigning jth points to class 1 if the following condition is satisfied.

$$b_0 + b_1 x_{j1} + b_2 x_{j2} + \cdots b_k x_{jk} > 0 \quad (1)$$

This is the logistic discriminant. Of course, logistic discriminant is able to be extended to qualitative response with more than two classes[5]. In this paper, we consider an alternative to logistic regression called linear discriminant analysis. For each class, the input variables are assumed to be distributed as multivariate normal distribution[4]. In the following discriminant rule shows that the point i is assigned to class 1.

$$\log \frac{n_1}{n_0} - \frac{(\bar{x}_1 - \bar{x}_0)^2}{2s^2} + \frac{x_i(\bar{x}_1 - \bar{x}_0)}{s^2} > 0 \quad (2)$$

Where, n_1 and n_0 are the number of points in class 1 and 0 respectively. Also, s^2 and x_i are the variance and ith points of X respectively.

3.1.2 Bayesian models

Bayesian models are based on Bayes' rule in the following[8].

$$P(H \mid X) = \frac{P(X \mid H)P(H)}{P(X)} \quad (3)$$

Where, X and H are data and hypothesis. Also, X can be belonged a class C. Using Bayesian approach, we can consider recursive graphical models which are naïve Bayes and Bayesian belief network models. Naïve Bayes classifier has a simplified assumption that all variables are conditionally independent as the following.

$$P(c_j \mid X) = P(c_j) \prod_{i=1}^{n} P(x_i \mid c_j) \quad (4)$$

So, this classifier reduces the computation cost. When ith variable is categorical, $P(x_i|C)$ is estimated as the relative frequency having value x_i as ith variable in class C. also, this is estimated through Gaussian density function if variable is continuous. Bayesian belief network has a subset of the variables conditionally independent. Also, this model is a graphical model of causal relationship. Using Bayesian belief network, we can take both network structure and all the variables easily.

3.2 Machine learning algorithms for classification

3.2.1 Decision trees

Decision tree methods can be used to solve either classification or regression problems. Generally the methods are commonly used for classification. The goal of decision trees is to be used as input variables of the other analytical models as well as predict the value of a categorical variable. The split criteria of decision trees for classification are CHAID(Chi-squared automatic interaction detector), gain ratio, and Gini diversity index. CHAID is a impurity measure which is the distance between the observed and the expected frequencies. This is based on the Pearson χ^2 statistic. Gain ratio criterion of decision trees is based on information theory. This is defined as the following.

$$Gr(t) = -\sum_{i=1}^{k} p(i \mid t) \log_2 (p(i \mid t)) \quad (5)$$

Where, t and i are a node and ith class of node t. Also, k is the number of classes in node t. Another criterion of node impurity, Gini diversity index is defined in the next formular.

$$G(t) = \phi(p(\cdot \mid t)) \quad (6)$$

This is similar to gain ratio.

3.2.2 Support vector machine

SVM is good classification method that maps the training vectors to high dimensional feature space, labeling each vector by its class[12],[13],[14]. That is, SVM is able to be defined the classification problem as a quadratic optimization problem. It classifies data by determining a set of support vectors, which are members of the set of training inputs that outline a hyper plane in feature space. The main advantage of SVM is binary classification and regression that they provide to a classifier with a minimal VC dimension, which implies low expected probability of generalization errors. To make this hyper plane, we maximize the quadratic form (7) subject to constraints (8).

$$W(\alpha) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{l} \alpha_i \alpha_j K(x_i, x_j) y_i y_j \quad (7)$$

$$\sum_{i=1}^{l} y_i \alpha_i = 0, \quad 0 \le \alpha_i \le C, \quad i = 1,2,\ldots l \quad (8)$$

where, x and y are the input vector and output label. is a kernel function. The kernels for three common types of SVM are polynomial function, radial basis function(RBF), and sigmoid function[17].

3.2.3 K-nearest neighbor and Kernel methods

K-nearest neighbor(K-nn) is a efficient classification model which is based on a combination of local methods[5]. A data set of K-nn is formed as (x,y), input-output pairs. K-nn is needed to a distance measure between the x of the observations. In K-nn, the response value is predicted as the following.

$$\hat{y}_i = \frac{1}{k} \sum y_i \quad (9)$$

Where, the mean of all output values are computed. Also, k is a constant which is the number of points to be included in each neighborhood. Kernel method is a nonparametric classification model as well as an unsupervised learning model. In classification problems,

this separates class densities well for classifying data points accurately.

3.3 Hierarchical structure of classification algorithms

In this paper, we propose a hierarchical structure of classification algorithms. The following figure shows the hierarchical structure.

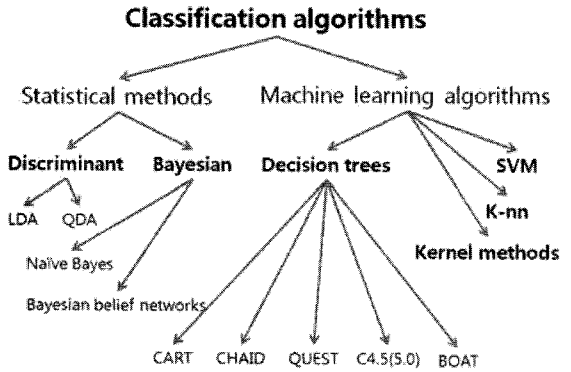

Fig. 1. Hierarchical Structure of Classification Models

In above figure, we show the hierarchical structure of popular classification algorithms by statistical methods and machine learning algorithms. In general the classification algorithms based on statistical methods are needed the normality assumptions. But, in the domains of practical application, the data sets are difficult to be satisfied the assumptions. On the other hand, machine learning algorithms are not needed the statistical assumptions.

4. Experimental Results

In our experiments, we use data sets from UCI machine learning repository[11]. These data sets are most data for classification. The data consist of diverse types according to the number of data points, the number of input variables, and the number of classes of target variable. The hardware system for the experiments has Microsoft Windows XP, Intel Pentium 4 CPU(3.6GHz), and 1GB RAM. Also, we use R language and packages of www.r-project.org for experimental computing. The following table shows summary of our training data sets.

Table 1. Summary of data sets

data	# of data points	# of input variables	# of classes
Bupa	345	6	2
Diabetes	768	8	2
Glass	214	10	7
Ionosphere	351	34	2
Iris	150	4	3
Letter	20000	16	26
Satellite	6435	36	6
Segmentation	2310	19	7
Shuttle	43500	9	7
Sonar	208	60	2
Spam	4601	57	2
Vehicle	846	18	4
Vowel	990	10	11
Wine	178	13	3

We are able to find the diversity of the experimental data sets. For example, the number of data points of Letter data is 20,000. This is big size compared with other data sets of UCI machine learning repository. Also, the number of data set of Wine is only 178. In the above figure, we show the number of input variables and classes of target variable about each data set. We use misclassification rate and cross-validation as measures of performing the result of each classification method[9]. Misclassification rate shows the percentage of the number of points which have been incorrectly classified for each class[5]. So, the result performance of classification algorithm is better according to decreasing the rate. We find the results of misclassification rates between statistical methods for classification.

Table 2. Misclassification rates of statistical methods

data	LDA	QDA	NB
Bupa	0.2957	0.3652	0.4406
Diabetes	0.2161	0.2357	0.2382
Glass	0.3271	0.2757	0.6449
Ionosphere	0.1054	0.0456	0.1709
Iris	0.0200	0.0200	0.0400
Letter	0.2951	0.1025	0.3548
Satellite	0.1554	0.1158	0.2026
Segmentation	0.0818	0.1121	0.2004
Shuttle	0.0556	0.1405	0.1550
Sonar	0.0962	0.0000	0.2692
Spam	0.1113	0.1671	0.2865
Vehicle	0.2021	0.0836	0.5272
Vowel	0.4303	0.1313	0.3919
Wine	0.0000	0.0056	0.0112

In above table, we know the results of discriminant analysis are better than Bayesian approach. That is, LDA and QDA show smaller misclassification rates than NB. Also, in the data sets with large size, the misclassification rates of LDA and QDA is smaller than NB. The following table shows the empirical results of misclassification rates among popular classification algorithms by machine learning approach.

Table 3. Misclassification rates of machine learning algorithms

data	Kernel	K-nn	SVM	CART
Bupa	0.0522	0.1942	0.2174	0.1826
Diabetes	0.0586	0.1406	0.1758	0.2057
Glass	0.0140	0.1869	0.2103	0.1542
Ionosphere	0.0171	0.0940	0.0370	0.6980
Iris	0.0000	0.0400	0.0267	0.0267
Letter	0.0000	0.0226	0.0376	0.6250
Satellite	0.0000	0.0468	0.0886	0.1829
Segmentation	0.0541	0.0229	0.0498	0.0528
Shuttle	0.0012	0.0012	0.0012	0.0047
Sonar	0.0000	0.1106	0.0192	0.0529
Spam	0.1026	0.1026	0.0526	0.0826
Vehicle	0.0000	0.2080	0.1537	0.2352
Vowel	0.0000	0.0061	0.0374	0.4586
Wine	0.0000	0.1405	0.0000	0.0169

On the whale, the misclassification rates of Kernel method are smaller than other machine learning algorithms. In large data sets and small attributes, the SVM algorithm is better than others. From the above results, we find the CART is worse than other methods. Next, we present the cross-validation results among the competitive methods. To compute cross-validation, we

divide the data sample with n points into two subsamples which are a training sample with (n-m) points and a validation sample with m points. One sample is used to construct model. Another sample is used to assess the constructed model. We show the cross-validation results among statistical methods for classification.

Table 4. Cross-validations of statistical methods

data	LDA	QDA	NB
Bupa	0.3014	0.4058	0.4463
Diabetes	0.2253	0.2604	0.2473
Glass	0.3505	0.4206	0.6495
Ionosphere	0.1368	0.1197	0.1823
Iris	0.0200	0.0267	0.0467
Letter	0.2977	0.1135	0.3584
Satellite	0.1604	0.1428	0.2042
Segmentation	0.0835	0.1203	0.2035
Shuttle	0.0557	0.1048	0.1551
Sonar	0.2452	0.2404	0.3269
Spam	0.1130	0.1697	0.2849
Vehicle	0.2210	0.1442	0.5414
Vowel	0.4525	0.1980	0.4323
Wine	0.0112	0.0056	0.0281

The results of cross-validation of LDA and QDA are better than the NB result. Also, the performance of LDA is similar to the performance of QDA. According to the results of misclassification rates and cross-validations, we find that the performance of discriminant analysis is superior to Bayesian approach. In the following table, we present the results of cross-validation among the competitive algorithms of machine leaning for classification.

Table 5. Cross-validations of machine learning algorithms

data	Kernel	K-nn	SVM	CART
Bupa	0.3188	0.3594	0.3044	0.3507
Diabetes	0.2708	0.3060	0.2422	0.2643
Glass	0.2804	0.2944	0.2943	0.3878
Ionosphere	0.1852	0.1510	0.0570	0.0940
Iris	0.0333	0.0400	0.0334	0.0267
Letter	0.0407	0.0303	NA	0.8632
Satellite	0.1683	0.0831	NA	NA
Segmentation	0.0740	0.0372	0.0567	0.0658
Shuttle	0.0020	0.0020	NA	0.2171
Sonar	0.0000	0.1827	0.1539	0.2548
Spam	0.1852	0.1852	0.0659	0.0848
Vehicle	0.1690	0.2920	0.2222	0.2731
Vowel	0.0141	0.0283	0.0626	0.5252
Wine	0.0056	0.2416	0.0169	0.0674

In the above table, NA defines not answer. That is, it is extremely difficult to compute the cross-validation, because we use exhaustive cross validation. This divides the data set with n points into a sample with (n-1) points and a point. Then we construct a classification model using the sample with (n-1) points and perform cross-validation using a point. Similar to the results of misclassification rates of machine learning algorithms, the performances of cross validation show that the values of cross-validation of Kernel method are smaller than K-nn, SVM, and CART. Therefore we are able to find the result for optimal usage of popular classification algorithms as the following conclusions.

5. Conclusions

In this paper, we propose empirical comparisons among popular classification algorithms by misclassification rate and cross-validation. According to statistical assumptions, the classification methods are divided to statistical models and machine learning algorithms. From the results of our experiments, we found that disciminant analysis is better than Bayesian approach in statistical classification methods. Also, in machine learning algorithms, the Kernel algorithm has better performance than other machine learning algorithms. In the classification data sets from UCI machine learning repository, machine learning algorithms for classification have better performances than statistical classification methods. By all experimental results, we found the results of Kernel methods showed good performance in classification data. In future works, we will perform advanced empirical comparisons which have additional classification algorithms and mode detailed criteria in synthetic data using simulation study.

References

[1] L. Breiman, J. H. Friedman, R. A. Olshen, C. J. Stone, Classification and Regression Trees, Wadsworth Inc., 1984.

[2] V. Cherkassky, F. Mulier, Learning From Data Concepts, Theory, and Methods, John Wiley & Sons, 1998.

[3] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning: Data mining, Inference, and Prediction, Springer, 2001.

[4] R. A. Johnson, D. W. Wichern, Applied Multivariate Statistical Analysis, Prentice Hall, 1992.

[5] P. Giudici, Applied Data Mining, Wiley, 2003.

[6] J. Han, M. Kamber, Data Mining Concepts and Techniques, Morgan Kaufmann, 2001.

[7] S. Haykin, Neural Networks, Prentice Hall, 1999.

[8] K. B. Korb, A. E. Nicholson, Bayesian Artificial Intelligence, Chapman & Hall, 2004.

[9] T. M. Mitchell, Machine Learning, McGraw-Hill, 1997.

[10] A. S. Pandya, R. B. Macy, Pattern Recognition with Neural Networks in C++, IEEE Press, 1995.

[11] UCI Machine Learning Repository, http://www1.ics.uci.edu/~mlearn.

[12] V. N. Vapnik, The Nature of Statistical Learning Theory, Springer, 1995.

[13] V. Vapnik, Statistical Learning Theory, John Wiley & Sons, Inc. 1998.

[14] V. N. Vapnik, "An Overview of Statistical Learning Theory", IEEE Transactions on Neural Networks, vol. 10, no. 5, pp. 988-999, 1999.

Seung-Joo Lee

He received the BS degree in department of applied statistics from Cheongju University, Korea in 1985. Also, he received MS, and PhD degrees in department of Statistics, Dongkuk University, Korea, in 1987 and 1995. He is currently Professor in department of Bioinformatics & Statistics, Cheongju University, Korea. He has researched Bayesian statistics and multi-variate analysis.

Phone : +82-43-229-8204

Fax : +82-43-229-8432

E-mail : access@cju.ac.kr

Sung-Hae Jun

He received the BS, MS, and PhD degrees in department of Statistics, Inha University, Korea, in 1993, 1996, and 2001. Also, He received PhD degree in department of Computer Science, Sogang University in 2007. He is currently Assistant Professor in department of Bioinformatics & Statistics, Cheongju University, Korea. He has researched statistical learning theory and evolutionary algorithms.

Phone : +82-43-229-8205

Fax : +82-43-229-8432

E-mail : shjun@cju.ac.kr