Evaluation of Classification System Using Naïve Bayes Classifier and Feature Selection Algorithms

Aung Nway Oo
University of Information Technology
aungnwayoo78@gmail.com

Abstract

In machine learning, feature selection, also known as variable selection, attribute selection or variable subset selection, is the process of selecting a subset of relevant for use in model construction. The feature selection approach gives enhanced prediction and reduces the computation time. This paper presents the comparative analysis of Naïve Bayes (NB) classifier with using two feature selection approaches namely Principal Component Analysis (PCA) and Correlation-based Feature Subset Selection (CFS). The experimental results prove that feature selection based Naïve Bayes classifier achieve higher accuracy rate.

Keywords: feature selection, NB, PCA, CFS

1. Introduction

The data mining algorithms should be computationally feasible for data analysis but takes low human intervention. As mentioned, data mining can be performed by using several techniques [1]. Among those techniques, classification [2] is very popular and this technique is being intensively used in many real business applications now-a-days [3]. In this paper, we used the Naïve Bayes algorithms for classification. The Naïve Bayes classifier (NB) is one of the most popular data mining techniques for classifying the large dataset. It has been successfully applied to the different problem domains of classification

Feature selection is an effective and an essential step in successful high dimensionality data mining applications [8]. Feature selection involves identifying a subset of the most useful features that produces compatible results as the original entire set of features. Proper feature selection facilitates the model building by reducing noise and the dimensionality of the problem. The reduced features are further classified by Naive Bayes classifier to produce better accuracy results. Two feature selection algorithms namely Principal Component Analysis (PCA) and Correlation-based Feature Subset Selection (CFS), are used to classify the different datasets.

The remainders of the paper are organized as follows. We describe the feature selection algorithm PCA and CFS in section 2 and 3. In Section 4 overview of the Naïve Bayes Classifier is discussed. The experimental results are described in section 5. Finally, conclusion is presented in section 6.

2. Principal Component Analysis (PCA)

Principal Component Analysis is an unsupervised Feature Reduction method for projecting high dimensional data into a new lower dimensional representation. PCA is a standard statistical technique that can be used to reduce the dimensionality of a data set. It is known as Karhunen-Loeve transform [4]. PCA combines the essence of attributes by creating an alternative, smaller set of variables. The initial data can then be projected onto this smaller set. PCA often reveals relationships that were not previously suspected and thereby allows interpretations that would not ordinarily result [7]. The basic procedure is as follows:

1. The input data are normalized, so that each attribute falls within the same range. This step helps ensure that attributes with large domains will not dominate attributes with smaller domains.

2. PCA computes k orthonormal vectors that provide a basis for the normalized input data. These are unit vectors that each point in a direction perpendicular to the others. These vectors are referred to as the principal components. The input data are a linear combination of the principal components.

3. The principal components are sorted in order of decreasing "significance" or strength. The principal components essentially serve as a new set of axes for the data, providing important information about variance.

4. Because the components are sorted in decreasing order of "significance," the data size can be reduced by eliminating the weaker components, that is, those with low variance. Using the strongest principal components, it should be possible to reconstruct a good approximation of the original data.

PCA can be applied to ordered and unordered attributes, and can handle sparse data and skewed data.

3. Correlation-based Feature Subset Selection (CFS)

CFS evaluates and ranks feature subsets rather than individual features. It prefers the set of attributes that are highly correlated with the class but with low intercorrelation [9]. With CFS various heuristic searching strategies such as hill climbing and best first are often applied to search the feature subsets space in reasonable time. CFS first calculates a matrix of feature-class and feature-feature correlations from the training data and then searches the feature subset space using a best first. Equation 1 (Ghiselli 1964) for CFS is

$$\text{Merit}_s = \frac{k\overline{rcf}}{\sqrt{k+(k-1)\overline{rff}}} \qquad (1)$$

Where Merit_s is the correlation between the summed feature subset S, k is the number of subset feature, $\overline{rcf}$ is the average of the correlation between the subsets feature and the class variable, and $\overline{rff}$ is the average inter-correlation between subset features.

4. Naïve Bayes Classifier

The Bayesian Classification represents a supervised learning method as well as a statistical method for classification. Assumes an underlying probabilistic model and it allows us to capture uncertainty about the model in a principled way by determining probabilities of the outcomes. This Classification is named after Thomas Bayes (1702-1761), who proposed the Bayes Theorem. Naive Bayes algorithm is based on Bayesian Theorem. Given training data X, posterior probability of a hypothesis H, P(H|X), follows the Bayes theorem

$$P(H|X)=P(X|H)P(H)/P(X) \quad (2)$$

Naive Bayes classifiers are highly scalable, requiring a number of parameters linear in the number of variables (features/predictors) in a learning problem.

5. Experimental Results

The WEKA data mining tool is used for evaluation and testing of algorithm. These dataset are collected from UCI Repository in the website www.ucirepository.com. Three datasets from UCI, mushroom dataset, diabetes dataset and Ionosphere dataset are used for comparing results. The results of different datasets are described in the following tables. The following figure visualizes the accuracy results of different datasets with different algorithms.

Table 1. Results of Mushroom Dataset

Algorithms	Correctly classified instance	In Correctly classified instance	Accuracy
NaiveBayes	2625	137	95.0398 %
NaiveBayes+CFS	2724	38	98.6242 %
NaiveBayes+PCA	2641	121	95.6191 %

Table 2. Results of Diabetes Dataset

Algorithms	Correctly classified instance	In Correctly classified instance	Accuracy
NaiveBayes	201	60	77.0115 %
NaiveBayes+CFS	213	48	81.6092 %
NaiveBayes+PCA	204	57	78.1609 %

Table 3. Results of Ionosphere Dataset

Algorithms	Correctly classified instance	In Correctly classified instance	Accuracy
NaiveBayes	98	21	82.3529 %
NaiveBayes+CFS	103	16	86.5546 %
NaiveBayes+PCA	103	16	86.5546 %

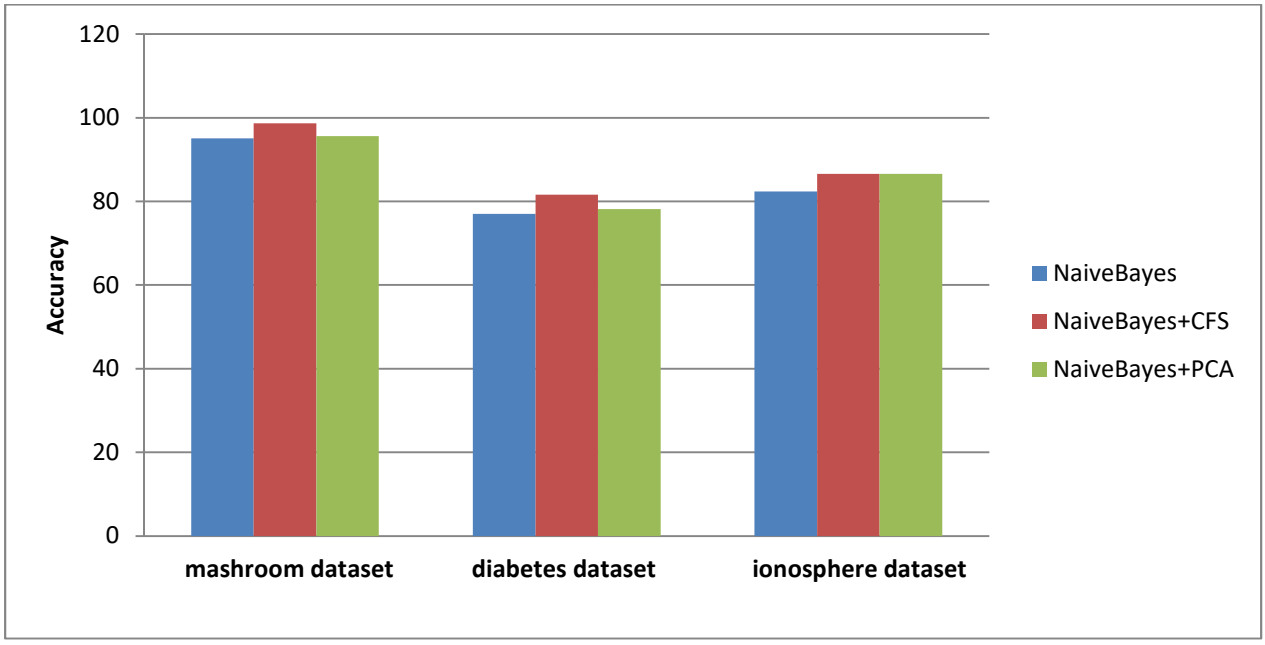

Figure 1: Accuracy results of different datasets

6. Conclusion

The experiment results showed, the accuracy results of Naïve Bayes classifier using feature selection approach is more than Naïve Bayes only. The result of Naive Bayes and CFS is better than other approaches. The evaluation results showed feature selection important for classification task of data mining.

REFERENCES

[1] Maindonald, J. H. 'New approaches to using scientific data statistics, data mining and related technologies in research and research training' Occasional Paper 98/2, The Graduate School, Australian National University, 1999.

[2 Quinlan, J. ,"Induction of Decision Trees," Machine Learning, vol. 1, pp.81-106, 1996.

[3] Berson, A., Smith, S. J. and Thearling, K. Building Data Mining Applications for CRM McGraw-Hill, 1999.

[4] L.Breiman, J.H. Friedman, R.H. Olshen, Stone C.J., Classification and Regression Trees, Wadsworth and Brooks, Monterey, CA, 1984.

[5] Kotu V and Deshpande, Predictive Analytics and Data Mining(Waltham: Morgan Kaufmann), 2015

[6] Kavitha R and Kannan, An efficient framework for heart disease classification using feature extraction and feature selection technique in data mining IEEE Int. Conf. on Emerging Trends in EngineeringTechnology and Science(ICETETS)pp 1-5, 2016

[7] Data Mining: Concepts and Techniques, Third Edition (The Morgan Kaufmann Series in Data Management Systems) 3rd Edition, 2012

[8] Liu H ,Setiono R, Motoda H, Zhao Z Feature Selection: An Ever Evolving Frontier in Data Mining, JMLR: Workshop and Conference Proceedings 10: 4-13 The Fourth Workshop on Feature Selection in Data Mining.

[9] I.H. Witten, E. Frank, M.A. Hall " Data Mining Practical Machine Learning Tools & Techniques" Third edition, Pub. – Morgan kouffman.