Classification of Medical Diagnosis by using Naive Bayesian Classifier

Tin Moe Htwe, Win Mar Oo
Computer University, Taunggyi
moehtwe13.tgi@gmail.com

Abstract

Classification is a form of data analysis that can be used to extract models describing import data class or to predict future data trends. This paper contains two important aspects of pattern recognition that classification problem and evaluate the performance are studied on real-world datasets. This system is to study the Naive Bayesian classifier and to classify the class label of two datasets. In this system, classifier is built on the training dataset and test the unknown dataset on the testing dataset. And then, calculate the accuracy of classifier by using the hold-out method. Before the classifier is built, missing value is filled by using mean value and feature value is normalized by using min-max normalization such as preprocessing step. The experiment is performed on two medical datasets from University of California, Irvine (UCI) machine learning database and General Hospital in Mandalay.

1. Introduction

Applying machine learning techniques to medical diagnosis tasks has the advantages of saving time and reducing cost. Machine learning provides tools by which large quantities of data can be automatically analyzed. Classification is a form of data analysis that can be used to extract models (rules) describing import data class or to predict future data trends. This paper is, to construct a rule (classifier) based on two medical datasets, such as Breast cancer and Hepatitis. The decision rule, called the classifier, is trained by using a number of observations with known class labels. This approach is also known as "supervised learning." The classifier is used to classify new instances with unknown classification. In two medical datasets, only two classes, i.e., the "Absent" class and the "Present" class are considered.

In the machine learning literature, many different techniques can be applied to medical diagnosis [7], including the naive Bayesian (NB) method [8], [6]. In this paper, Naive Bayesian (NB) method is used to build a classifier based on two datasets. Naive Bayesian Classification (NBC) assumes that the attributes are mutually independent. Although in practice this assumption is not quite true, experience shows that the NBC approach in medical application is effective and gives relatively good classification accuracy in comparison with other, more elaborate learning methods. The accuracy of Naive Bayesian algorithm is measured by using hold-out method of a given datasets into disjoint train and test sets. To use these tools effectively, an important part is preprocessing in which data is processed before it is presented to any learning, discovering, or visualizing algorithm [9]. The product of data preprocessing is the final training datasets.

1.1 Related Method

Decision trees ad their induction is one of the most important and thoroughly investigate methods of machine learning [1]. There are many existing algorithms proposed for induction of a decision tree from a collection of records described by attribute vectors. A decision tree forms a model which is then used to classify new records. In general, a decision tree is constructed in a top-down fashion, from the root node to leaves. In each node an attribute is chosen under certain criteria and this attribute is used to split the collection of records covered by the node. The nodes are split until the of records have the same value of the decision attributes covered by the node. The critical point of this general approach in thus the selection of the attribute upon which the records are spilt. The selection of the splitting attribute is the major concern of the research in the area of decision trees.

The classical methods of attribute selection, implement in well-known algorithm ID3 and C4.5 [2, 3], are based on minimizing the entropy of information gain, i.e. the amount of information represented by the clusters of records covered by nodes created by nodes created upon the selection of the attribute.

2. Data preprocessing

Data preprocessing is an important issue for both data warehousing and data mining, as real-world data

trend to be incomplete, noisy, and inconsistent. A common occurrence for many learning problems is when some examples are missing values for some of the features. This can be a complicating factor during the learning phase, when assessing the important of features in forming some learning results, and in classification, when making a decision when values of some of the features are unavailable. So need to fill the missing value. In this paper, missing values are filled by using the attribute mean value. Then attribute values are transformed by using the min-max normalization. Suppose that $\min_A$ and $\max_A$ are the minimum and maximum values of an attributes A. Min- max normalization maps a value v of A to v' in the range [new-$\min_A$, new-$\max_A$] by computing.

$$v' = \frac{v - \min_A}{\max_A - \min_A}(new\_\max_A - new\_\min_A) + \text{new\_min}_A$$

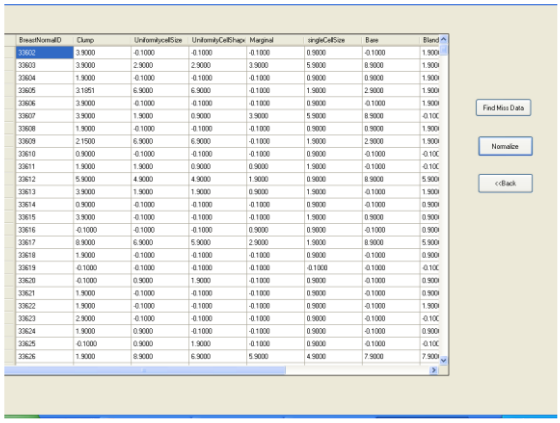

Figure 1: Preprocessing for normalization based on breast cancer

2.1 Classification

The authors proposed in [4], [5], Classification is an important data mining technique which predicts the class of a given data sample. Classification is a two step process. In the first step, a model is built describing a set of predetermined classes. The set of instances used for model construction is training set. This model is represented as classification rule, decision tree, or mathematical formulae.

In the second step, this model is used for classifying future of unknown objects and to estimate accuracy of the model. If the accuracy is acceptable, use the model to classify instances that class labels are not known. One type of classification patterns used in this paper is the simple Bayesian classifier (also called the Naive Bayesian Classifier). Bayesian classification method is used to classify a given data set.

2.1.1 Naive Bayesian classifier

The Naive Bayes algorithm employed a simplified a version of Bayes formula to decide which class a novel instance belongs to. The posterior probability of each class is calculated, given the feature values present in the instance; the instance is assigned the class with the highest probability. Equation (1) shows the naive Bayes formula, which makes the assumption that feature values are statistically independent within each class.

$$p(C_i \mid x_1, x_2, ..., x_n) = \frac{p(C_i)\prod_{j=1}^{n} p(x_j \mid C_i)}{p(x_1, x_2, ..., x_n)} \quad (1)$$

The left side of Equation (1) is the posterior probability of class C_i given the feature values, $< x_1, x_2, \ldots, x_n >$, observed in the instance to be classified. The denominator of the right side of the equation is often omitted because it is a constant which is easily computed if one requires that the posterior probabilities of the classes sum to one. Naive Bayes classifier is straightforward and involves simply estimating the probabilities in the right side from the training instances. The result is a probabilistic summary for each of the possible classes. If there are numeric features it is common practice to assume a normal distribution again the necessary parameters are estimated from the training data.

2.2 Hold-out method

The holdout method as shown in figure 2, sometimes called test sample estimation, randomly partitions the data into two mutually exclusive subsets called training set and the test set, holdout set. It is common to designate 2/3 of the data as the training set and the remaining 1/3 as the test set. The training set is used to train the classifier. The training set is also used for validation. The testing set is used to estimate the error rate of the trained classifier. It is important to take note of the fact that this method assumes there is lots of data available.

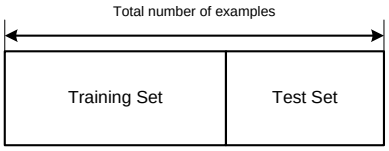

Figure 2: Holdout Method

3. Proposed system

Proposed system has four stages as shown in figure 3. In the first stage, the data are partitioned into training and testing using hold-out strategy over the whole data set. We believe the accuracy calculated in this way is more reliable in the classification method. Training data set is used to build the classifier. Testing data set is used to test the unknown data set. In the second stage, Training data are preprocessed to fill the missing value by using the mean value and normalize the features by using the min-max normalization. In the third stage, these training data are used as the input features to build a classifier by using Naive Bayesian algorithm and evaluate the accuracy of the classifier. And then system produces the accuracy of the classifier as the result. Finally, unknown data set is tested on the testing data set in the fourth stage.

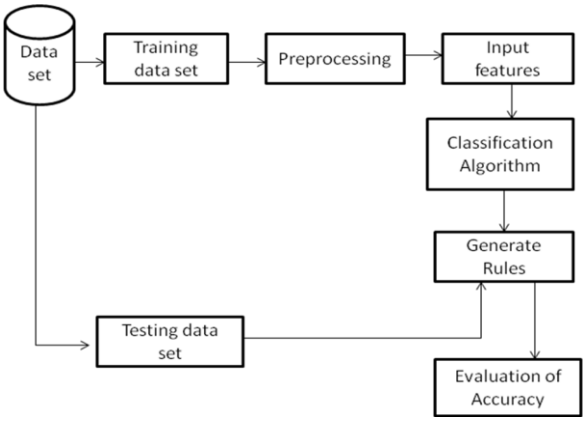

Figure 3: System flow for classification

4. Experimental results

The experiments described in this paper to compare Naive Bayesian Classifier using the accuracy the evolution criteria. The data used for these experiments obtain from UCI machine learning dataset repository and General Hospital in Mandalay. These data sets are summarized in Table1.

Accuracy of algorithm is measured using the hold-out method of a given datasets into disjoint train and test sets. An algorithm is trained on the training dataset and the induced theory is evaluated. Table 2 shows the train and test set size used with the real-world datasets.

Dataset	# of feature	Instances	Missing value	Class
Breast cancer	9	699	Yes	2
Hepatitis	16	150	Yes	2

Table 2. Training and testing set sizes of the real-world datasets.

Dataset	Training set	Testing set
Breast cancer	466	233
Hepatitis	100	50

Table 1. Datasets used in experiment.

In order to evaluate the result of the classifiers, some means of measuring or evaluating classifier performance is required. In some situations and for some tasks, it is important to evaluate the overall performance of the classifier. This overall performance is measured in terms of accuracy. There are four kinds of subsets after classification:

- True positive answers denoting correct classifications of positive cases (true positives-*T-Pos*).
- True negative answers denoting correct classification of negative cases (true negatives-*T-Neg*)
- Positive answers denoting the number of positive samples (*Pos*).
- Negative answers denoting the number of negative samples (*Neg*).

The classification accuracy measures the proportion of correctly classified cases. The result of experiment for accuracy measured by using Naive Bayesian classifier is presented in Table 3.

$$accuracy = \frac{T\_Pos + T\_Neg}{Pos + Neg}$$

Table 3: comparison of accuracy measured breast cancer and hepatitis

Dataset	Native Bayes
Breast cancer	80%
Hepatitis	86%

Table 3 shows the performance of Naive Bayesian algorithm. Naive Bayesian is more maintains or improves the accuracy on hepatitis than breast cancer.

5. Conclusion

In the paper, the important aspects of pattern recognition are studied on the real-world datasets are classification problem and performance evaluation. Classification can be used for prediction the class label of dataset and to extract model describing important data class. In this system, Naive Bayesian classifier is used as a basis for comparing the performance of breast cancer and hepatitis dataset. Naive Bayes classifier is a Bayesian learning method that has been found to be useful in many practical applications. It is called "Naive" because it incorporates the simplifying assumption that attribute values are conditionally independent, given the classification of the instance. The experiments also show that the accuracy performance of one classification method, Naive Bayesian. This classifier is evaluated by using the attributes values. The result shows that the Naive Bayes classification method gives efficient accuracy rate of hepatitis than breast cancer.

References

[1] H . Almuallim and T.G Dietterich, " Learning with many Irrelevant Features", AAAI-91 proceedings, 9th National Conference on Artificial Intelligence, 1991.

[2] P.S Bradley, U.M. Fayyad, and O.L, Mandasarian, "Mathematical Programming for data mining: formulations and challenges", INFOMS Journal on Computing, pp. 217-238,1999.

[3] L. Bremian, "Bagging predictors" , Machine learning pp.123-140,1996.

[4] Breiman, "Bias, variance, and arcing classifiers", Technical Report 460, University of Califormia, Berkeley, CA., 1996.

[5] M. L. Ginsberg, "Essentials of Artificial Intelligence", Morgan Kaufmann Publishers, Inc, 1993.

[6] M. I. Jordan, Why the Logistic Function? A Tutorial Discussion on Probabilities and Neural Networks Tech. Rep. 9503, 1995, MIT Computational Cognitive Science Report.

[7] I. Kononenko, "Machine learning for medical diagnosis: history, state of the art and perspective," *Artif. Intell. Med.*, vol. 23, pp. 89–109, 2001.

[8] P. Langley, W. Iba, and K. Thompson, "An analysis of Bayesian classifiers,"in *Proc. National Conf. Artificial Intelligence (AAAI-1992)*, 1992, pp. 223–228.

[9] H. Liu and H. Motoda, "Feature Selection for Knowledge Discovery Data Mining", Boston: Kluwer Academic Publishers, 1998.