Classification of Paddy by using NAIVE BAYESIAN Classifier

Thida San, Thwe
University of Computer Studies, Mandalay
sunandaroo09@gmail.com

Abstract

Classification is a form of data analysis that can be used to extract models describing important data classes or to predict future data trends. Data classification is a two step process. In this system, a model is built on the training datasets by using the Naive Bayesian classification algorithm. And then, a model is used to test the unknown datasets. The performance of classifier is estimated by using the hold-out method. The Naive Bayesian (NB) classifiers have been one of the most popular techniques as basis of many classification applications both theoretically and practically. This system presents a Naive Bayesian classification learning in order to evolve useful subset of paddy features for classification task. This system is determined the kind of paddy by using Naive Bayesian classification.

1. Introduction

This system classifies the type of paddy by using Naive Bayesian Classification. Based on high, spike-count, spike long one spike-fruit, seed weight and season etc, paddy name are classified.

Classification is the process of finding the common properties among different entities and classifying them into classes. This system is to search the input feature used in the Naive Bayesian classification. These feature subsets are evaluated using a limited amount of training data.

Analysis can be helped provide us with a better understanding of the data at large. Whereas classification predicts categorical labels, prediction models continuous valued functions.

Many classification and prediction methods have been proposed by researchers in machine learning, pattern recognition and statistics. Most algorithms are memory resident typically assuming a small data size. To be learn about extensions to these techniques for their application to classification and prediction in large data base [5].

The remainder of the paper is organized as follow. Section 2 presents related work. Section 3 represents proposed system. Section 4 discusses the Naïve Bayesian Classification. Section 5 is the experimental result of the system, section 6 presents system implementation and section 7 is conclusion.

2. Related work

Many classification techniques have been used for document classification. Nowadays, the most popular text categorization algorithms are Decision Tree Induction, Naive Bayesian classification, Neural Network, K-Nearest Neighbor and so on.

Arul Prakash Asirvatham, Kranthi Kumar Ravi presented web page classification based on document structure in which web pages are classified into three categories: research page, information page and personal homepage. And textual information and image information are considered as features [1].

Web mining is discussed and how web page classification can be achieved is described. In this classification, web page binary classification system was implemented in three stages: Representation, Learning and Classification algorithm. Naïve Bayesian classification method is used for classification [3].

H.Kim et.al [2] described a nontrivial extension of document classification methodology from a fixed set of classes C = c_1, c_2, ..., c_n to a knowledge hierarchy. To develop the text mining classifier, called Associative Naive Bayesian Classifier.

3. Proposed System

In this system, paddy database can be partitioned into training dataset (2/3) and testing dataset (1/3) by using hold-out method. The hold-out method is a simple technique that uses a training dataset of class-labeled samples.

These samples are randomly selected and are independent of the testing samples. And then, can be made the input data and calculate the result by using Naive Bayesian classification. After calculation produce the paddy name and calculate the accuracy. The accuracy of a model on a given training dataset is the percentage of test set samples that are correctly classified by the model.

This system is used from Naive Bayesian classification equation

(1) $X = (x_1, x_2, ..., x_n)$

Where, x_1 = high, x_2 = spike – count,..... x_n = season

(2) $P(C_i) = \frac{S_i}{S}$ where, C_i = paddy name,

S_i = the number of paddy samples of class C_i

S = the total number of paddy samples

(3) $P(x_k / C_i) = \frac{S_{ik}}{S_i}$

where, x_k = the input features of paddy

S_{ik} = the number of paddy samples of class C_i having the value of x_k

S_i = the number of paddy samples belong to C_i

3.1. Data Flow Diagram of the System

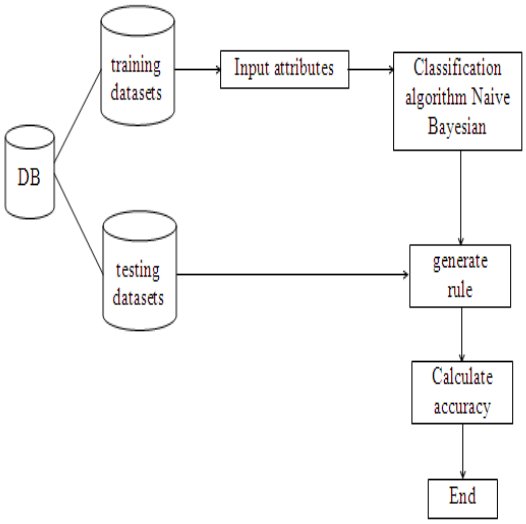

Figure 1. classification of paddy by using the Naive Bayesian classifier.

The data flow diagram of the system is presented in figure 1. In this system, database is divided into training and testing by using holds-out method. When user inputs attributes, the system generates rules by using Naïve Bayesian classification. Finally, the system calculates the accuracy.

3.2. Classification

Data classification is a two-step process. In the first step, a model is build describing a predetermined set of data classes or concepts.

The model is constructed by analyzing database tuples described by attributes.

Each tuple is assumed to belong to a predefined class, as determined by one of the attributes called the class label attribute. The data tuples analyzed to build the model collectively form the training data set. The class label of each training sample is provided, this step is also known as supervised learning.

Typically, the learned model is represented in the form of classification rules, decision tress, or mathematical formulae. The rules can be used to categorize data samples, as well as provide a better understanding of the database contents.

In the second step the model is used for classification [5].

3.2.1. Bayesian Classification

Bayesian classifiers are statistical classifiers. They can predict class membership probabilities. Bayesian classification is based on Bayes theore. Studies comparing classification algorithms have found a simple Bayesian classifier known as the naive Bayesian classifier to be comparable in performance with decision tree and neural network classifiers. Bayesian classifiers have also exhibited high accuracy and speed when applied to large database.

Naive Bayesian classifiers assume that the effect of an attribute value on a given class is independent of the values of the other attributes. This assumption is called class conditional independence. Bayesian belief networks are graphical models, which unlike naive Bayesian classifiers, allow the representation of dependencies among subsets of attributes [4].

4. Naive Bayesian Classification

The naive Bayesian classifier, or simple Bayesian classifier, works as follows:

1. Each data sample is represented by an n-dimensional feature vector, X = (x_1, x_2,....x_n), depicting n measurements made on the sample from n attributes, respectively, A_1, A_2, ..., A_n.
2. Suppose that there are m classes, C_1, C_2, ..., C_m. Given an unknown data sample, X (i.e., having no class label), the classifier will predict that X belongs to the class having the highest posterior probability, conditioned on X. That is, sample X to the class C_i if and only if the naive Bayesian classifier assigns an unknown

$$P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)}$$

3. As P(X) is constant for all classes, only $P(X|C_i)P(C_i)$ need be maximized. If the class prior probabilities are not known, then it is commonly assumed that the classes are equally likely, that is, $P(C_1) = P(C_2) = ...= P(C_m)$, and we would therefore maximize $P(X|C_i)$. Otherwise, we maximize $P(X|C_i)P(C_i)$. The class prior probabilities may be estimated by $P(C_i) =$

$\frac{s_i}{s}$, where S_i is the number of training samples of class C_i, and S is the total number of training samples.

4. Given data sets with many attributes, it would be extremely computationally expensive to compute $P(X|C_i)$. In order to reduce computation in evaluating $P(X|C_i)$, the naive assumption of class conditional independence is made. This presumes that the values of the attributes are conditionally independent of one another, given the class label of the sample, that is, there are no dependence relationships among the attributes. Thus,

$$P(X|C_i) = \prod_{k=1}^{n} P(x_k \mid C_i).$$

The probabilities $P(x_1|C_i)$, $P(x_2|C_i)$,....., $P(x_n|C_i)$ can be estimated from the training samples, where

If A_k is categorical, then $P(x_k|C_i) = \frac{s_{ik}}{s_i}$, where s_{ik} is the number of training samples of class C_i having the value x_k for A_k, and s_i is the number of training samples belonging to C.

In order to classify an unknown sample X, $P(X|C_i)P(C_i)$ is evaluated for each class C_i. Sample X is then assigned to the class C_i if and only if$P(X|C_i)P(C_i) > P(X|C_j)P(C_j)$ for $1 \leq j \leq m, j \neq i$

In other words, it is assigned to the class C_i for which $P(X|C_i)P(C_i)$ is the maximum.

4.1. Classifier Accuracy

Estimating classifier accuracy is important in that it allows one to evaluate how accurately a given classifier will label future data, that it is, data on which the classifier has not been trained. Accuracy estimates also help in the comparison of different classifiers.

$$\text{Accuracy} = \frac{\text{No of correct Paddy}}{\text{No of total Paddy}} \text{ x } 100$$

The accuracy for each interval is the percentage of documents correctly classified out of the number of documents in that interval.

5. Experimental Result

The main purpose of this system is to classify the paddy name. The naive Bayesian classifier could produce a satisfied classification. In this system paddy name are classified into fifty classes such as Manawthuka, Thihtetyin, Shwethweyin and so on. The attributes of paddy are high, spike count, spike length, seed per spike, seed weight, color of pollen, color of leaf, structure of plant, structure of leaf, flower life and season. The training data set are shown in table 1.

In this system, 600 paddy are used as training data: 12 paddy each for class. So, $P(C_i)$ is obtained 12/600 for each class with 11 attributes. Eg. $P(C_{Manawthuka}) = 0.02$

Since the attribute value in this system are counts to estimate $P(X_k/ C_i)$. So,

$P(X_n/C_i) = P(X_1/ C_i) \text{ x } P(X_2/C_i) \text{ ....x } P(X_{11}/ C_i)$

By this way, P(X/Manawthuka),P(X/ Tuithiri), P(X/ Thihtetyin),P(X/ Shwethweyin) and so on will be obtained. Paddy that has the highest probability of $P(X / C_i)$ will belong to the class C_i.

In this system, data are partitioned into two sets, a training set and testing set. Two thirds of the data are allocated to the training set and the remaining one thirds is allocated to the test set. When calculate the accuracy, the system calculate the probability by using Naive Bayesian classification and number of correct paddy is divided by number of total paddy. This system provides 73.3 % classification accuracy for 300 paddy, 78.5 % for 400 paddy and so on as shown in Table 2.

Table1. Training Data Set Table

ID	Attribute				Class
	High (x_1)	Spike Count (x_2)	---	Season (x_{11})	
1	3.0-3.5	10	---	Rain	Manawthuka
2	3.0-3.5	10	---	Summer	Manawthuka
3	2.5-3.0	8	---	Rain	Thihtetyin
4	2.5-3.0	10	---	Rain	Shwethweyin
5	4.5-5.0	10	---	Rain	Ayarmin
6	3.0-3.5	10	---	Rain	Mahawbe(1)
7	2.5-3.0	10	---	Summer	Thihtetyin
-	---	---	---	---	---

Table2. Classifier Accuracy Table

Training data	Testing data	Accuracy
108	54	66.6%
200	100	73.3%
266	134	78.5%
332	168	86.5%
600	300	91.6%

6. System Implementation

In this system implementation, the system calculates the probability of paddy after user fill input shown in figure 2. Figure3 is classified the type of paddy and figure 3 is define the accuracy.

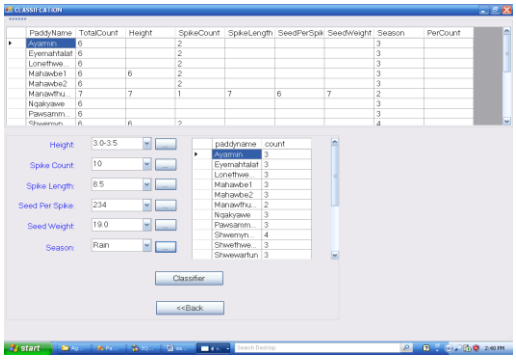

Figure 2.Classification of Input Form

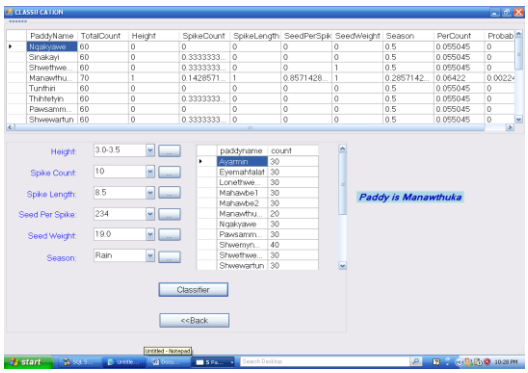

Figure3.Classification of Define Paddy Name Form

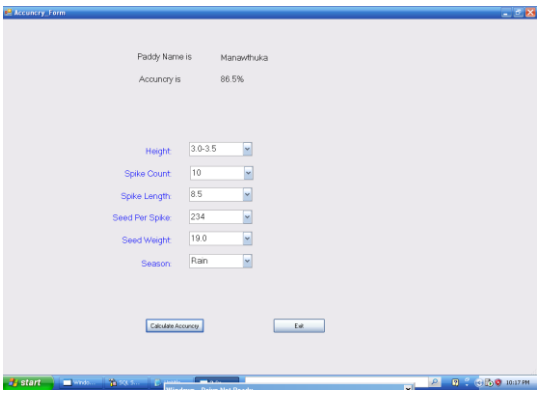

Figure4.Classification of Accuracy Form

7. Conclusion

Classification is the process of finding a set of models that describe and distinguish data classes or concepts. Bayesian classification is based on Bayes theorem. The Naive Bayesian classification use is focused on the Bayesian formula to calculate the probability of each class given the values of all attributes.

This paper can be applied to predict the class label of unknown sample given the sample data. The relative performance of the Naive Bayesian classification can serve as an estimate of the conditional independence of attributes.

This system involves classification of paddy name, based on features such as high, spike count; spike long, one spike-fruit, seed weight and season by using Naïve Bayesian classification.

This paper has presented generating of classification from large data sets. This approach demonstrates efficiency and effectiveness in dealing with paddy data for classification which considers the classification problem of paddy by using Naive Bayesian classification. This paper is to support for Agricultural Research Department. This paper provide high accuracy when large amount of data.

References

[1]A.P. Asirvatham, K. K. Ravi, "Web page classification based on Document Structure"

[2]H. Kim, S Chen Associative Naive Bayes classifier: Automated linking of gene ontology to medline documents. Pattern Recognition (30 January 2009)

[3]H. KOU, "Text Mining" IFT6255. Information Retrieval "Text classification".

[4]Jiawei Han and Micheline Kamber, Data Mining concepts and techniques, First edition, page:296

[5]Jiawei Han and Micheline Kamber, Data Mining concepts and techniques, second edition, page : 285-286.