Prediction of Bank Loan Type Using Naive Bayesian Classification

Thandar Aung, Thin Zar Win
University of Computer Studies, Kyainge Tong
ninichay@gmail.com; thinzarwin07@gmail.com

Abstract

Data Mining is the process of storing through large amount of data and picking out relevant information. Classification can be used as in the form of data analysis that can be used to extract models describing the important data classes. Classification is the task to identify the class labels for instances based on a set of features (attributes). As Myanmar is the rice-mart of the world, agricultural remains the vital sector of the economy and measures has been taken to increase productivity of paddy and other crops. It needs to predict the possibility of default of a potential counterparty before they extend a loan to the growers. In this paper, we proposed a bank loan-type prediction system for Myanmar agriculture development Bank by using Bayesian classification. Our system's performance results are also discussed in this paper.

1. Introduction

Data mining is a powerful technology that converts raw data into an understandable and actionable form, which can then be used to predict future trends or provide meaning to historical events It is becoming increasingly common in both the private and public sectors. It can be performed on data represented in quantitative, textual, or multimedia forms. Data mining applications can be used as a variety of parameters to examine the data. It includes association, sequence or path analysis, classification, clustering and forecasting.

The Bayesian classification method is a generative statistical classifier. Studies comparing classification algorithms have found that the simple or naive Bayesian classifier provides relatively good performance compared with other more complex algorithms. Accuracy of classification is a very important property of a classifier, measure of which can be separated in two parts: measure of accuracy in case of trained samples and measure of accuracy in case of untrained samples. Naive Bayesian classification is generally very accurate since all testing samples are trained before and has no outliers.

The Naïve Bayesian classifiers have been one of the most popular techniques as a basic of much classification application both theoretically and practically. It provides a flexible way for dealing with any number of attributes of classes and is based on probability theory. It is fastest learning algorithm that examines all its training input. Classification of a collection consists of dividing the items that make up the collection into categories or classes.

Bank officers need to learn which loan applicants are “safe” and which are “risky” for the bank by analyzing their data. Our proposed system intends to predict loan type of Myanmar agriculture development bank by using Bayesian classifier based on the bank attributes, such as Year, State/Division, Saving-account Installment, Housing, season, No of field, Type of loans.

This paper is organized as follow .In section 2, a description of related work is given. Section 3 describes classification and prediction and Bayesian classification. Section 4 describes proposed system. Section 5 describes Data set description .Section 6 describes Experimental Result and Section 7 describes conclusion.

2. Related Work

Supervised Naive Bayesian classification algorithm allows the use of fuzzy attributes in [1]. They argue that this algorithm is suitable in applications that provide only small training sets and require a compact representation of the internal model. Naïve Bayes Classifier method was presented in [2] for phoneme classification in the reconstructed phase space. This method is a novel approach substantially different from existing techniques. A Naïve Bayes classifier uses the probability mass estimates for classification.

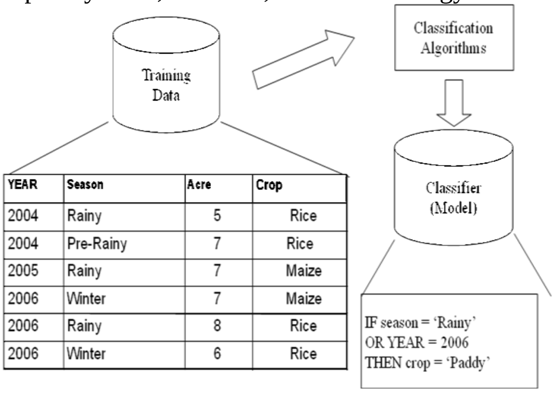
A robust model has been developed in [3] for the forecasting of long term interest rates in India. They employed multivariate time-series techniques to develop the model. Based on this, their estimated model is used to generate out-of-sample forecasts. Subsequently the forecasts are compared with naïve models such as random walk. This approach is verified using isolated fricative, vowel, and nasal phonemes, a software package for flexible data analysis in predictive data mining tasks. It is based on a simple Bayesian network, the Naive Bayes classifier. It also provides a feature selection scheme which can be used for analyzing the problem domain, and for improving the prediction accuracy of the models constructed by BAYDA.

In paper [5], Limsombunchai and et.al proposed a lending decision model (credit scoring) for the agricultural sector in Thailand. The logistic regression and artificial neural networks (ANN) were used to identify critical factors in lending decision process in the agricultural sector and to predict the borrower's creditworthiness.

This paper describes for the bank loan application by using naïve Bayesian classification. Naive Bayesian algorithm builds a model to classify new examples based on observed probabilities and supporting evidence from the training data. An accurate model can assist the banks about the future expected loan type during each of the time periods.

3. Classification and Prediction

Classification and prediction are two forms of data analysis that can be used to extract models describing important data classes or to predict future data trends. Whereas classification predicts categorical labels, prediction models continuous-valued functions. For example, a classification model may be built to categorize bank loan applications as either safe or risky, while a prediction model may be built to predict the expenditures of potential customers on computer equipment given their incomes and occupations many classification and prediction methods have been proposed by researchers in machine learning, expert systems, statistics, and neurobiology.



YEAR	Season	Acre	Crop
2004	Rainy	5	Rice
2004	Pre-Rainy	7	Rice
2005	Rainy	7	Maize
2006	Winter	7	Maize
2006	Rainy	8	Rice
2006	Winter	6	Rice

IF season = 'Rainy'
OR YEAR = 2006
THEN crop = 'Paddy'

Figure 1. Overview of Classification

3.1. Bayes Theorem

Let X is a data sample whose class label is unknown. Let H be a hypothesis that X belongs to class C. For classification problems, determine

P(H|X)-the probability that the hypothesis holds given the observed data sample X.
P(H)-prior probability of hypothesis H (i.e. the initial probability before we observe any data, reflects the background knowledge)
P(X)-probability that sample data is observed
P(X|H)-probability of observing the sample X, given that the hypothesis holds.

Given training data X, posteriori probability of a hypothesis H, P(H|X) follows the Bayes theorem

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

3.2. Bayesian Classification

Bayesian classifiers are statistical classifiers. They can predict class membership probabilities, such as the probability that a given sample belongs to a particular class. It is based on Bayes theorem and has also exhibited high accuracy and speed when applied to large database. A naive Bayes classifier is a term in Bayesian statistics dealing with a simple probabilistic classifier based on applying Bayes' theorem with strong (naive) independence assumptions. A more descriptive term for the underlying probability model would be "independent feature model".

In simple terms, a naive Bayes classifier assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature. For example, a fruit may be considered to be an apple if it is red, round, and about 4" in diameter. Even though these features depend on the existence of the other features, a naive Bayes classifier considers all of these properties to independently contribute to the probability that this fruit is an apple.

Depending on the precise nature of the probability model, naive Bayes classifiers can be trained very efficiently in a supervised learning setting. In many practical applications, parameter estimation for naive Bayes models uses the method of maximum likelihood; in other words, one can work with the naive Bayes model without believing in Bayesian probability or using any Bayesian methods. Suppose that there are m classes, C_1, C_2… Cm. Given an unknown data sample, X, the classifier will predict that X belongs to the class having the highest posterior probability,

$P(C_i \setminus X) > P(C_j \setminus X)$ for $1 \le j \le m$ and $j \ne i$, Where

$$P(C_i \setminus X) = P(X \setminus C_i)P(C_i)/P(X)$$

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

For Instance, X= (year=2001, State=rakhine, season=rainy, SC=less then 1000DM, Installment=Bank, Housing=Rent, No of field=2, Loan type= paddy)

The system need to maximize P(X/C$_i$) P (C$_i$) the prior probability of each class, can be computed based on the training sample :
P (C$_i$) =P (loan type=paddy) = total no of "paddy "loan type/total no of loan type
P(X/C$_i$) = P (year=2001/loan type=paddy)

The probability of other attributes is also calculated. An advantage of the naïve Bayes classifier is that it requires a small amount of training data to estimate the parameters (means and variances of the variables) necessary for classification.

4. Proposed System

Our proposed system has two basic steps. The first one is modeling of classifier and the second one is testing the model. At the first step, the system constructs the model for classifier with Naive Bayesian classification structure. The system uses information from training dataset to construct model and calculate the prior probability of each label. Then it calculates the posterior probability

(P [X\Ci]) based on the attribute values. Then it calculates the right part of the equation by multiplying the obtained two probabilities. Finally the system maximizes the products to define the predicted label.

At the second step, the system fire the test data sample to the model and count the hit and miss of each label to calculate the accuracy of the model. In addition, the class label of unknown class sample is accepted; the system predicts the type of loan based on information. Our proposed system's architecture is shown in figure (2).

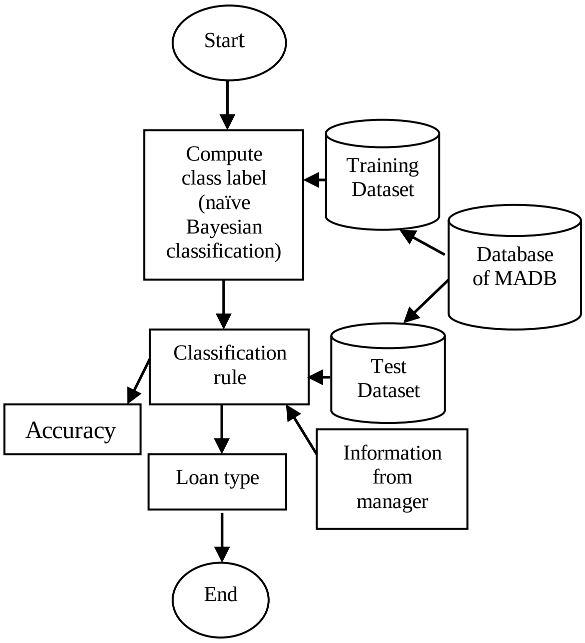


Figure 2. Proposed system Architecture

In our system, customer data collected from Myanmar Agricultural Bank which is used as training data. When users provide loan information like year, division and town, number of acres, season and Installment, the system will predict the loan type of the crops with most loan amount. There are basically two kinds of loan types that used in Myanmar Agricultural Development bank. These are short term loan and long term Loan based on the growing time of the crops. Our system is focus on short term loan.

5. Dataset and Description

First the data all of which have known class are partitioned into a training set and testing set. Typically, two thirds of data are allocated to the training set, and the remaining one third is allocated to testing set. There are 8 attributes in our system and the sample training data set can be seen as in Table 1.

1. Year {Real}
2. State {Yangon, Rakhine,...}
3. Season {rainy, winter, pre-rainy}
4. Saving- Account (SC) {less100DM,less 500DM, less
   1000DM, over10000DM}
5. Installment {bank, store, none}
6. Housing {Rent, own, free}
7. No of field {Real}
8. Type of loan {paddy, pea, groghum, sesame}

Table 1: Sample Training Dataset

year	state	season	SC	In-stall-ment	hous ing	no of field	type of loan
2001	Shan	rainy	Less 500 DM	bank	Rent	2	pad dy
2001	Mon	winter	Less 100 DM	store	free	1	pea
2001	Shan	rainy	Less 100 DM	bank	Rent	2	pea
2003	Mon	winter	Less 500 DM	store	own	1	Ses ame

6. Experimental Results

In Classification Problem, it is commonly assume that all samples are uniquely classifiable, that is, that each training sample can belong to only one class. This system evaluate how accurately a given classifier will label future data, that is, data on which the classification has not been trained.

The Holdout Method are used to evaluate the accuracy in this system. In holdout method, the given data are randomly partitioned into two independent sets, a training set and test set.

Two-thirds of the data sets are allocated to the training set and the remaining One-third of the data sets are used to the test set.

Suppose that, we have trained a classifier to classify loan type. In loan type, “paddy” or “not-paddy” are classified. An accuracy rate of, say, 90% may make the classifier seem quite accurate, but if only, say, 3-4% of the training samples are actually “paddy”. Clearly, an accuracy 90% may not be acceptable- the classifier could be correctly labeling only the “not-paddy” sample, for insteance.

Instead, we would to be able to access how well the classifier can recognize “paddy” samples (positive samples) and how well it can recognize “not-paddy” samples (negative samples). The sensitivity and specificity measures can be used, respectively. In addition, we may use precision to access the percentage of samples labeled as “paddy” that actually are “paddy” samples. These measures are defined as

$$sensitivity = \frac{t\_pos}{pos}$$

$$specificity = \frac{t\_neg}{neg}$$

$$precision = \frac{t\_pos}{(t\_pos + f\_pos)}$$

$$accuracy = sensitivity\frac{pos}{(pos+neg)} + specificity\frac{neg}{(pos+neg)}$$

Where, t_pos is the number of true positives(“paddy” samples that were correctly classified as such), pos is the number of positive (“paddy”) samples, t_neg is the number of true negatives(“ not-paddy” samples that were correctly classified as such), neg is the number of negative (“not-paddy”) samples, and f_pos is the number of false positives (“not_paddy” sampless that were incorrectly labeled as “paddy”).

In Table 2, the classification of the system based on training dataset has 448 correct data and 252 incorrect data. Testing dataset has 198 correct data and 102 incorrect data. Therefore, the accuracy of our system is 64% on training and 65% on testing.

Table 2. Training and Testing of Bank loan Using Naïve Bayesian classification

Dataset	Correct	Incorrect	Prediction
Training	448	252	64%
Testing	198	102	65%

The accuracy of classification Algorithm can be estimated by calculating the percentage of patterns in a testing dataset. In figure 3 show accuracy of testing in loan type of bank. In training data 700 and testing data 300, accuracy we got is 65.45%.In training data 600 and testing 300, accuracy we got is 63.57%.In training data 500 and testing 300, accuracy we got is 61.23%.

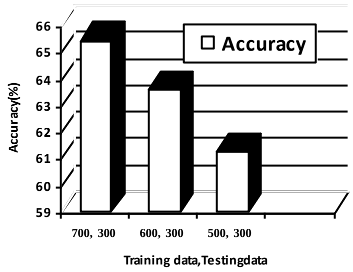


Figure 3. Classification accuracy of crop’s loan type

7. Conclusion

The paper intends to improve the prediction accuracy of classifier from Myanmar Agricultural Development Banks point of view. It will provide the bank option and quick and easy finds of information with respect to bank needs with minimum effort within relevant time interval. In this paper, we presented the benefit of naïve Bayesian classification that is used in bank Loan application. Although our system can support to predict for short term loan, in future we intend to undertake training with larger data sets by using different customers’ dataset and various types of loan. So, it can be exploited as a useful tool in Myanmar Agricultural Development Bank although it needs further modifications.

8. References

[1] [1].H.P.Storr,”Acompact fuzzy extension of Naïve Bayesian classification Algorithm”, AIInstitute Science Technique University at Dresden

[2].J.Ye, R.J.Povinelli, M.T.Johnson “Phoneme Classification Using Naïve Bayes Classificaton In Reconstructed Phase Space” Department of Electrical And Computer Engineering, MarquetteUniversity,ilwaukee,WI

[3] Kapil S and Rudra.S “Forecasting Long Term Interest Rate: An Econometric Exercise for India” Indian Institute of Management, Prabandh Nagar, Off Sitapur Road Lucknow-226013, India, Feb 2006

[4] Kontkanen, Petri and Aki, Petri Myllym and Sil Tomi and Tirri, Henry “BAYDA: Software for Bayesian Classification and feature selection ”Proceedings of the Fourth International Conference on knowledge Discovery and Data Mining (KDD-98)

[5] Limsombunchai, V. 2C. Gan and 3M. Lee Commerce Divisions, Lincoln University, New Zealand 3Department of Economics, American university of Sharjah, UAE

[6] Martin, Andrew D. "Bayesian Analysis", Washington University, National Science Foundation Methodology, Measurement, and Statistics Section, December 29, 2005