An Implementation of Naive Bayesian based Bagging Method for Advertisements Prediction

Moh Cherry Maung, Dr.Myat Thuzar Tun
University of Computer Studies, Yangon
sakurasan22@gmail.com, mtzucsy@gmail.com

Abstract

Nowadays, the field of advertising is more spread. Broadcasting media receives advertisements from advertisement companies. When programme are shown, these advertisements used to be broadcasted. Most of the advertisement companies install to broadcast its advertisements when advertisement time. As Myanmar rule, it must be broadcasted one third of the programme. So, media cannot receive all advertisements. By using this system, it is easy to know which advertisement companies should be imperative if new programme broadcast. In this system, Bagging and Naive Bayesian Classification methods are used. Bagging method is one of the well-known ensemble techniques that build bags of data of the same size of the original data sets by applying random sampling with replacement. Naive Bayesian Classification method is widely used for probabilities estimations. Using the mixture of these algorithms, we get the probabilities from multiple models .Then, averaging probability values are calculated with bagging method. Finally, the system extracts the company names with highest probabilities. Companies can be classified into "classes". Another way, users can view products information and reports at different sites. No need to look large amount of historical data.

1. Introduction

Competitive advantage requires abilities. Abilities are built through knowledge. It comes from data. The process of extracting knowledge from data is called Data Mining. Data mining involves analysis tools to discover previously unknown, valid patterns and relationships in large data set. These tools can include statistical models, mathematical algorithms, and machine learning methods. It provides tools for automated learning from historical data and developing models to predict outcomes of future situations. Consequently, data mining consists of more than collecting and managing data, it also includes analysis and prediction. It is becoming an increasingly important tool to transform this data into information. It is commonly used in a wide range of profiling practices, such as marketing, fraud detection and scientific discovery [2].

Data mining is the task of discovering interesting patterns from large amounts of data where the data can be stored in database, data warehouse or other information repositories [3]. Data mining has been rapidly evolving into a widely used technique for dealing with large amount of data for enterprises to analyze advertisements prediction. Data mining technology is also the rapidly developing key technology, whose development and perfection accomplish the aims of Naive Bayesian based Bagging method.

Data mining has many classification algorithms such as decision tree, Naive Bayesian, rule-based classification, backpropagation, neural network, clustering and so on. This paper is organized as follows- Section 1, discuss introduction of this system. In Section 2, we explain classification method and in Section 3, presents related work and then in Section 4**,** describe Naive Bayesian Classification method. Section 5, discusses about bag classifiers and how it works. Section 6 is the System Overview of the paper and proposed system of the paper. We finally present some conclusion in Section 7.

2. Classification

Data tuples can be referred to as sample. The class label of each training tuple is provided; this is known as supervised learning. Classification and prediction are supervised learning. They also have numerous applications such as target marketing, performance prediction, manufacturing and medical diagnosis. Naive Bayesian method has been widely used for clustering and classification. Comparing with other method, Naive Bayesian classification method is simple and effectiveness. The class label of each training tuple is not known and the number of set of classes to be learned may not be known in advance is called unsupervised learning (clustering) [1].

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.

3. Related Work

This paper attempts to improve the Naive Bayesian algorithm by bagging classifier. The related system intends to implement a bagged classifier based on naive Bayesian classifications to predict the class label of an unknown sample. The implemented classifier can be used as a supporting tool for decision making problems in [5]. Bagging Classifier has been implemented in classification and prediction of protein domain structural class in [4].

4. Naive Bayesian Classification

Bayesian classifiers are statistical classifiers. They can predict class membership probabilities. Bayesian classification is based on Bayesian Theorem. A simple classifier is known as the Naive Bayesian Classifier to be comparable in performance with decision tree and neural network classifiers. Bayesian classifier has exhibited high accuracy and speed when applied to large database. Bayesian theorem is below [2].

$$P(H|X) = \frac{P(X|H)\,P(H)}{P(X)} \qquad (1)$$

X =data sample whose class label is unknown.
H =some hypothesis such as X belongs to specified class.
P(H|X)=posterior probabilities that hypothesis H holds given the observed data sample X.
P(X|H) = posterior probabilities of X conditioned on H.
P(H) = prior probability of H and
P(X) = prior probability of X

Suppose that there are m classes, C1, C2… C_m. Given an unknown data sample, X, the classifier will predict that X belongs to the class having the highest posterior. Probability conditioned on X. That is, the naive Bayesian classifier assigns an unknown sample X to the class C_i if and only if

$$P(X \mid Ci) = \prod_{k=1}^{n} P(xk \mid Ci) \qquad (2)$$

$$P(x1 \mid Ci) * P(x2 \mid Ci) * .... * P(xn \mid Ci) \qquad (3)$$

5. Bagging Method

Bagging method is the acronym of 'bootstrap aggregating'. Bootstrap aggregating, or "bagging", this paper attempts to improve the Naive Bayesian algorithm by bagging classifier. It has been shown to be very successful in improving the accuracy of certain classifiers for artificial and real-world datasets. Suppose we have a model fit to a set of training data. The training set is Z= (z_1, z_2...z_n) where Z_i= (x_i ,y_i)

The basic idea is to randomly draw datasets with replacement from the training data. Each sample set is the same size as the original training set. This is done B times and we will have B bootstrap datasets [6].

Bootstrap sampling with replacement from the original data. These are the effects of noisy data. The increased accuracy occurs because the composite model reduces the variance of the individual classifiers. For prediction, it was theoretically proven that a bagged predictor will always have improved accuracy over a single predictor derived from D.

5.1 Advantages of Bagging Method

If the base classifier is stable, the Bagging may adversely deteriorate the classification accuracy because each classifier receives lesser of the training data. The Bagging algorithm reduces the variance of the classification. It can improve the classification accuracy significantly if the base classifier is properly selected. It is also not very sensitive to noise in the data [8].

5.2 Bagging Averaging

We form a new predictor based on the average of the bootstrap predictor and can be used to assess the accuracy of a parameter estimate or prediction. In bagging, we use it to improve the estimation or prediction itself. For each bootstrap sample set, b=1,2,…,B , giving prediction the bagging estimate is

$$\hat{f}_{bag}(X) = \frac{1}{B}\sum_{b=1}^{B} \hat{f}^{*b}(X) \qquad (4)$$

5.3 Bagging Maximum

Let be a classifier for a k-class response. Consider an underlying indicator vector function

$$\hat{f}(x) = (0,...,0,1,0,...,0) \qquad (5)$$

The entry in the i^{th} place is 1 if the prediction for the ith class, such that

$$\hat{G}(x) = \arg\max\ \hat{f}(x) \qquad (6)$$

Then the ^{k}bagged estimate where the proportion is of base classifiers predicting class at where [7]. Finally,

$$\hat{G}_{bag}(x) = \arg\max_{k}\ \hat{f}_{bag}(x) \qquad (7)$$

6. System Overview

This paper is to predict company names which are possible to advertise in broadcasting new programme. The detail processes of proposed system are as shown in Figure 1. Firstly, this system starts with login process to determine whether the user is administrator or not. If the user is administrator, user can do two processes –
(1) Analysis process and
(2) Enquire process

Analysis process consists of four main steps and uses Naive Bayesian Classification method and Bagging method.

Step 1: the system specifies number of classifiers and attributes values. The accepted attribute values are:

(1) number of classifiers (1,2,3,…10),
(2) series (Korea series, Chinese series , melody, Myanmar video…),
(3) weather(summer , rainy, weather),
(4) day (Monday, Tuesday,., Sunday),
(5) duration (5s,10s,….),
(6) time (11:00AM, 7:00 to 8:00 PM,…) .

Step 2: Each classifier works on randomly drawn datasets and accepted attribute values. This dataset is drawn with replacement from the training set. Each classifier calculates probabilities for each advertisement company by using naive Bayesian classification method.

Step 3: Average probability for individual advertisement company is calculated by means of bagging Averaging method. Then, based on these averaging probabilities, maximum probabilities are determined by Bagging Maximum.

Step 4: Finally, Names of companies, which probabilities are selected as maximum probabilities, are extracted.

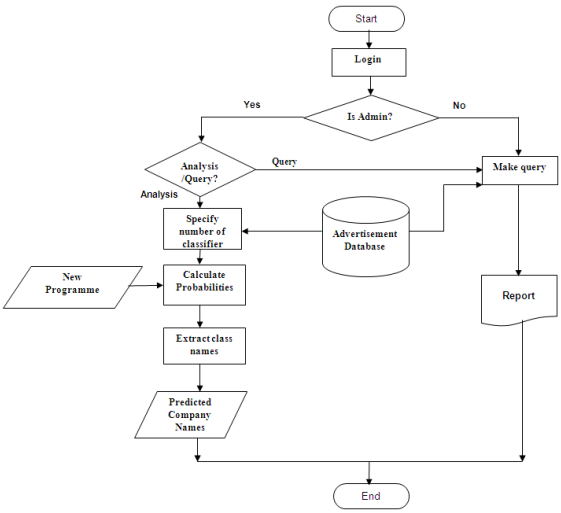

Figure1. System Overview

Another process is enquiring process. If the user is administrator, which can view and retrieve reports based on following criteria by counting number of advertisements or total duration.

1. Advertisement category
2. Programme
3. Time and
4. Weather

If the user is guest, user can enquire advertisement's product information only. Guest user can view advertisement companies for desired product. Consequently, he can view advertisement information for selected company which is resulted from previous query.

6.1 Analysis Process

Using the sample data in table 1, the sample calculation process for analysis process is as follow:

Table1.Sample training data set

Weather	Day	Time	Duration	Programme	Company Name
Summer	Saturday	7:00 to 8:00 PM	30	Korea Series	SES
Summer	Sunday	7:00 to 8:00 PM	30	Korea Series	SES
Summer	Friday	7:00 to 8:00 PM	30	Korea Series	SES
Summer	Friday	9:00 PM	30	Living Songs	Zomia
Summer	Wednesday	7:00 to 8:00 PM	10	Korea Series	Ki Ki
Summer	Monday	7:00 to 8:00 PM	10	Korea Series	Majestic
Summer	Saturday	7:00 to 8:00 PM	60	Korea Series	SES
Summer	Saturday	7:00 to 8:00 PM	60	Korea Series	SES
Summer	Tuesday	7:00 to 8:00 PM	30	Korea Series	Ki Ki
Summer	Saturday	7:00 to 8:00 PM	30	Korea Series	M&G
Summer	Wednesday	7:00 to 8:00 PM	30	Korea Series	Zomia
Summer	Sunday	9:00 PM	40	Puzzle Palace	Moe Sann Pann
Summer	Wednesday	7:00 to 8:00 PM	40	Korea Series	Zomia
Summer	Saturday	7:00 to 8:00 PM	20	Korea Series	Kaung Hein
Summer	Wednesday	7:00 to 8:00 PM	10	Korea Series	Kaung Hein
Summer	Saturday	7:00 to 8:00 PM	15	Korea Series	M&G
Summer	Tuesday	7:00 to 8:00 PM	15	Korea Series	SES
Summer	Sunday	7:00 to 8:00 PM	20	Korea Series	Zomia
Summer	Sunday	7:00 to 8:00 PM	20	Korea Series	Zomia
Summer	Saturday	7:00 to 8:00 PM	20	Korea Series	Zomia
Summer	Wednesday	7:00 to 8:00 PM	20	Korea Series	Zomia
Summer	Wednesday	7:00 to 8:00 PM	30	Korea Series	Zomia
Summer	Thursday	7:00 to 8:00 PM	30	Korea Series	Zomia
Summer	Tuesday	7:00 to 8:00 PM	30	Korea Series	Zomia
Summer	Monday	6:00 to 7:00 PM	40	Thwet Lat Hlote Shar Ka Sar Soe Lar	Kaung Hein
Summer	Friday	7:00 to 8:00 PM	40	Korea Series	Multi
Summer	Saturday	9:00 PM	20	Lu Shwin Taw Shwe Pha lar	KI Ki
Summer	Tuesday	7:00 to 8:00 PM	15	Korea Series	Multi
Summer	Thursday	7:00 to 8:00 PM	15	Korea Series	Ki Ki
Summer	Wednesday	7:00 to 8:00 PM	15	Korea Series	MTZ

Firstly, if the user is administrator, user puts all inputs for analysis and then specifies number of classifiers .Secondly, by depending number of classifiers, this system divides training data sample .Each data set can randomly draw from the training data and then calculated Naive Bayesian classification equation. Then, this system gets probabilities for each class label. The sample calculation is follow;

Input-

Number of classifiers =3
Series =Korea Series
Weather =Summer
Day =Sunday
Duration =at least 10s
Time =7:00 to 8:00 PM

Let C1= SES, C_2= Zomia, C_3= Ki Ki, C_4 =Majestic, C_5=M & G, C_6=MoeSannPann, C_7 =Kaung Hein , C_8= Multi, C_9 = MTZ, C_{10}= Thu Ta Yate Mon, C_{11} = Love, C_{12} = Creation

Classifier 1-
$P(X|C_1)$ =6/6*6/6*6/6*1/6*6/6 =0.166
$P(X|C_2)$ =9/10*10/10*9/10*2/10*10/10=0.162
Classifier 2-
$P(X|C_1)$= 12/12*12/12*12/12*2/12*12/12 =0.166
$P(X|C_2)$= 2/3*3/3*2/3*1/3*3/3 =0.148
Classifier 3-
$P(X|C_1)$=4/4*4/4*4/4*1/4*4/4 =0.25
$P(X|C_2)$=5/6*5/6*6/6*2/6*6/6 =0.231
$P(X|C_8)$=3/3*3/3*2/3*1/3*3/3 =0.222
$P(X|C_{11})$=4/7*7/7*4/7*2/7*5/7 =0..666
$P(X|C_{12})$=4/4*4/4*4/4*1/4*4/4 =0.25

Then, these probabilities are find averaging with Bagging Averaging method. The following averaging probabilities are obtained.
$P(X|C_1)$ =(0.166+0.166+0.25)/3 =0.194
$P(X|C_2)$ =(0.162+0.148+0.231)/3 =0.180
$P(X|C_8)$ =0.222/3 =0.074
$P(X|C_{11})$=0.666/3 =0.222
$P(X|C_{12})$=0.25/3 =0.083

Finally, if the admin can choose three highest companies name's probabilities, this system find with bagging maximum method. The following maximum results are obtained.
$P(X|C_{11})$ =Love =0.222
$P(X|C_1)$=SES =0.194
$P(X|C_2)$ =Zomia =0.180
$P(X|C_{12})$ =Creation=0.083
$P(X|C_8)$ = Multi =0.074

Among them, the highest three maximum company names are obtained.
Company names - Love
SES
Zomia

6.2 Enquire Process

Moreover, administrator can make by two ways. They are –
(1)View summarize by company
(2) View summarize by category

(1) Views summarize by company-This view has four categories. They are -
(i)advertisement category (stationary, English medicine, Myanmar medicine, men wear, etc.),
(ii)programme (Korea series, Chinese series, melody world, etc),
(iii)time (7:00 to 8:00 PM, 9:00 PM, etc.),
(iv)weather (summer, rainy, winter).
Admin can view each or all companies reports at different sites by calculating number of advertisement or total duration.
Example enquire- Admin enquires number of advertisements for KiKi Advertisement Company's advertisement rating .Then, the results in table2 are obtained.

Table2. KiKi advertisement Company's advertisement rating

Time	No of advertisement
11:00 AM	212
6:00 to 7:00 PM	227
7:00 to 8:00 PM	1921

(2)View Summarize by Category-This view has three categories. They are –
(i)company (KiKi, Zomia, Love, SES, etc),
(ii)programme (Korea series, Chinese series, etc),
(iii)time (5:00 to 6:00 PM, 2:00 PM, 11:00 AM, etc).
Admin can look each or all advertisement categories at different sites by calculating number of advertisements or total duration.
Example enquire-admin enquires number of advertisements for food category and corresponding companies. In table 3, display results.

Table 3. Food category is advertised Company names

Company	No of food advertisement
Creation	262
Emperor	189
ITM	205
Kaung Hein	845

Another user is guest who can look the advertisement information. The guest may advertise their own new product. So, user can enquire which kinds of products are advertised and which companies broadcasted this product.
Example enquire-If user choose food category, companies which advertised food category are displayed .And then ,user choose one company ,the food advertisements of that company are displayed .

7. Conclusion

This paper tends to predict the class label for a new programme. In this system, Naive Bayesian and Bagging approach are used together for effective prediction. The aim of using bagging is that this paper tends to predict the class from multiple models rather than one for the accurate prediction. Over thousand of training advertisement data are used in this system. In classification and prediction process, the system will give the successful and popular advertisement company names which are suitable to select for broadcasting of new programme by using mixture of Bayesian and Bagging Method.

References

[1] Bayesian network,
http://en.wikipedia.org/wiki/Bayesian_network

[2] Jiawei Han and Micheline Kamber
Data Mining: Concepts and Techniques
Second Edition.

[3] J. Ross Quinlan. Morgan Kaufmann, San Mateo, Calif., C4.5 Programs for Machine Learning.

[4] Liuhuan Dong, Yuan Yuan, Yudong Cai
Using Bagging Classifier to Predict Protein Domain Structural Class Dept of Combinatory and Geometry CAS-MPG Partner Institute for Computational Biology Shanghai Institutes for
Biological Sciences Chinese Academy of Sciences China 200031 Department of Statistics Jiaxing College, Zhejiang, China 314000
http://www.jbsdonline.com

[5] Myat Htun Oo, Dr. Myat Thuzar Htun
University of Computer Studies, Yangon
Implementation of Bagged Classifier Based on Naive Bayesian Classification.

[6] Thomas G. Dietterich
An experimental comparison of three methods for costructing ensembles of decision trees: Bagging, boosting, and randomization. Machine Learning, (40):139–157, 2000.

[7] Xiaogang Su
Bagging and Random Forests
Department of Statistics and Actuarial Science University of Central Florida

[8] Y. C. Tzeng
Using Modified Bagging and Boosting Algorithms in Multiple Classifiers System for Remote Sensing Image Classification.