Application of Bayesian rules based on Improved K-means classification on Credit Card

Xie Meiping
School of Information Management and Engineering, Shanghai University of Finance & Economics, Shanghai 200433
E-mail:xiemp@shufe.edu.cn

Abstract

K-means clustering *algorithm is a method of cluster analysis which aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean. It is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. It is similar to the expectation-maximization algorithm for mixtures of Gaussians in that they both attempt to find the centers of natural clusters in the data. Bayesian rule is a theorem in probability theory named for Thomas Bayesian. It is used for updating probabilities by finding conditional probabilities given new data. In this paper, K-mean clustering algorithm and Bayesian classification are combined to analysis the credit card. The analysis result can be used to improve the accuracy.*

KeyWords*: K-Means clustering, Bayesian Rule, Credit card*

I. Introduction

[1]Since the data mining is the synthetic product of multi-subjects, it has drawn the advantages from all of these subjects including database technique, artificial neural network nets, statistical methods, mode identification, information searches, database visualization and so on. A commonly acknowledged definition is a complex process of picking the cryptic, unknown, potentially great useful and valuable models, rules or some practical knowledge from the database. It is actually a kind of deep-layer data analysis method. In order to improve the conciseness of basic Bayesian classification, many literatures have made advancements in broadening the independence of the conditions. In fact the elements that affect the conciseness is not only the relativity among the attributes but also the completeness. In response to this reason, the introduction of the clustering arithmetic for the K average into the Bayesian classification method is aimed to improve the conciseness. K-means clustering has been used in many fields[1-2], Bayesian rule is a theorem in probability theory named for Thomas Bayesian. It is used for updating probabilities by finding conditional probabilities given new data. It has been used in many fields [3-4]. Credit card is the fastest-growing banking industry of the financial business. In order to prevent risks, many methods are developed to mine or analysis the customers [5-7]. In this paper, K-mean clustering algorithm and Bayesian classification are combined to analysis the credit card. The analysis result can be used to improve the accuracy and shows that the method is feasible.

II. K-Means Rules

The key point in the K-means clustering rule is to divide the data into different clusters through iterative method. The ultimate aim is to get the target function minimized, the cluster produced will be as close and independent as possible.

Input: expected number of clusters: k, the database of n objects.

Output: k clusters which make the square error criteria function to be the minimal one.

Steps:

(1) Selecting k as the of the original cluster centroid;

(2) Calculating the distances between the objects and every cluster centroid, then divide the objects to the closest cluster;

(3) Re-calculating the average of the every new cluster;

(4) Keep doing this until the centroid tends to be unchangeable

Features:

(1) A pre-fixed k;

(2) Creating a initial division ,then using the position relocation technique of the interactive

(3) The distances and matrixes can be unsure

(4) Less calculating than the hierarchical clustering method and it is suitable for dealing with huge sample database.

(5) It is suitable for the discovery of the ball-like ones.

The K-Means algorithm has the advantages of fast clustering and easy realization. But there is a pre-fixed number k of the clusters. This condition has affected and

[1]This paper is supported by Leading Academic Discipline Program , 211 Project for Shanghai University of Finance and Economics (the 3rd phase)

978-0-7695-3817-4/09 $26.00 © 2009 IEEE
DOI 10.1109/WISM.2009.11

constrained the rationalization of its utilities. What's more, as for this algorithm, the choosing of the center of the origin cluster is stochastic which may bring instability to the result. Hence it is of high value to improve the quality and stability in the cluster analysis.

III. Bayesian Rule

Bayesian rule is a method belonging to the statistics. They can forecast the rate of whether some target data belongs to some certain category.

Bayesian rule makes a hypnosis that the of all attributes are independent from each other. This hypnosis is also be named as: independent, it helps to effectively reduce the calculation work when the Bayesian classification rule is found.

The basic Bayesian classifier is described as the following figure 1. Suppose a variable quantity collection $U=\{A_1,A_2,\cdots,A_n, C\}$. Among them, $A_1,A_2,\cdots,A_n$ is the variable attribute quantity in the practices.

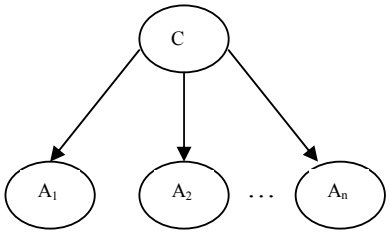

Fig.1 Bayesian classifier

The results of the study in comparing the classification methods show that the Bayesian classifier has the same function as neural net work. And in analyzing the large database, the Bayesian classifier has shown high conciseness and calculating function.

The Bayesian classifier has the features as following:

(1) It doesn't assign an object to a certain category unconditionally. Instead, it works out the rate through calculation. The category which has the largest rate is the one that object belongs to.

(2) Commonly, all the attributes function invisibility. It means it not that several attributes that determine the classification but all the attributes.

(3) The attributes of the objects can be discrete, consecutive and mixed also.

Compared to other classification methods, the Bayesian classifier has the lowest mistake rates.

A. Bayesian Theories

Suppose x to be a sample database whose belonging is unknown, Suppose H as a hypothesis, for example, sample database X belongs to a specific category C. As for the classification, our goal is to fix P(H|X)---fixing a observed sample database X and the rate when H is supposed to be right.

P(H|X) is posterior probability which is the rate of the rightness of H under condition X. For example, Suppose the sample database is fruit, the attributes described are colors and shapes. Suppose X signifies red color and round shape, H is the hypothesis that X are apples. So P(H|X) presents the rate of the fact that X are apples when fruit X are known as red and round.

To the contrary, P(H) is the priori probability, in the examples above, P(H) signifies the rates of the fact that the sample is apple no matter what color it is and what shape it is. P(H|X) is based on more information. While P(H) has no relation with X.

Similarly, P(X|H) is the after-rate of the foundation of X under conditions H. Which means, if it is already known that H is apple, the rate of X being red and round can be signified as P(X|H). P(X) is the priori probability of X, which is also the rate of picking up a sample which is red and round from the collection.

Bayesian rule describes how to work out the P(H|X) according to the P(X),P(H) and P(X|H). Among which, the rate of P(X),P(H) and P(X|H) can be get from the data collection.

B. The basic Bayesian classification procedure

(1) Every sample database uses n-dimension vectors to signify the specific number of its n attributes.

(2) Suppose there are m different categories, $C_1, C_2, \ldots C_m$ .An unknown data sample X is given. The classifier, when X is known, predict the category which X most likely belongs to. Which is, when the basic Bayesian classifies the unknown sample X into category C_i, only when (1) is true.

$$P(C_i \mid X) > P(C_j \mid X) \qquad 1 \le j \le m, j \ne i \quad (1)$$

Which is also that $P(C_i \mid X)$ is largest. The category C_i is called the supposition of the largest after-rate.

$$P(C_i \mid X) = P(X \mid C_i)P(C_i) / P(X) \qquad (2)$$

Suppose there are m different categories, $C_1, C_2, \ldots C_m$. A sample database X whose category is still unknown is given.

(3) Since P(X) is the same to all the categories, it will be ok with the largest $P(X \mid C_i)P(C_i)$ .And because that the pre-rate of each category is unknown, the occurrence rates of each categories are supposed to be the same, that is P(C_1)=P(C_2)=…=P(C_m). This way, formula (2) chose the maximum then it turns out to seek the largest P(X|C_i),otherwise the largest P(X|C_i) and P(C_i) must be largest. While the pre-rate of the pre-rate can be estimated through using formula P(C_i)=s_i/s, the s_i is the

number of the C_i category in the sample collection. S is the size of the training sample collection.

(4) If according to the offered database which includes a few attributes, there will be quite a large amount of computation to work $P(X|C_i)$ out directly. In order to estimate $P(X|C_i)$ effectively, Bayesian classifier usually suppose that each category is independent from each other. Which means, the attribute values are independent. For a certain category, its attributes are independent of each other. There are :

$$P(x \mid C_i) = \prod_{j=1}^{n} P(x_j \mid C_i) \tag{3}$$

Values of $P(x_1|C_i)$, $P(x_2|C_i)$, …, $P(x_n|C_i)$ can be estimated according to the training samples. Detailed explanation of the method is as followed:

If A_k is a symbol quantity, $P(x_k|C_i)=s_{ik}/s_i$, s_{ik} is the sample's number when the category is C_i and A_k's value is v_k. And it is also the sample's number which falls in the C_i category.

If A_k is a consecutive quantity, and suppose the attributes are in line with Gaussian distribution property, hence there will be:

$$P(x_k \mid C_i) = g(x_k, \mu_{C_i}, \sigma_{C_i})$$

$$= \frac{1}{\sqrt{2\pi\sigma_{C_i}}} \exp(-\frac{(x_k - \mu_{C_i})^2}{2\sigma_{C_i}^2} \tag{4}$$

In order to predict the category of an unknown sample X, we can estimate the corresponding value of $P(X|C_i)P(C_i)$. Sample X will belong to category C_i, only when

$$P(C_i \mid X) > P(C_j \mid X) \qquad 1 \le j \le m, j \ne i \tag{5}$$

IV. Simulation

Computation method of K average and the integration of Bayesian classification integration

As we all know, banks as the pillar of a nation's finance industry, a large amount of the interest comes from the banks' loan business. And the credit risk of the loaners make many money can't be repaid, we call this kind of money "bad money". Thus great lost will be caused to the bank. In order to prevent this lost from happening and lower banks' potential risk of debt collection, we can use the data digging to make analysis of the old debt collection cases thus help banks do the credit rating. As a result, we can prohibit the come-into-being of the bad money ahead.

The paper is conducted against the background of banks' credit problems. Some practical examples will be cited to show how to use the new computation method which has combined the K average computation with Bayesian classification to deal with those data. The picture beyond represent the tabular form of dealing with the data.

Table 1.

Age	*Income*	*Assets*	*Debts*	*Want*	*credit*	*Risk*	*On-time*
28	*216844*	*161566*	*248903*	*3000*	*red*	*high*	*1*

In order to improve the conciseness of Bayesian computation method, the data will be classified through K-means computation method. The conditions for the classification is shown as following

Data classification:

Table 2.

Age	*values*	*Debts*	*values*
[20,30)	*0*	*[0.25000)*	*0*
[30,40)	*1*	*[25000,50000)*	*1*
[40,50)	*2*	*[50000,100000)*	*2*
[50,60)	*3*	*[100000,200000)*	*3*
[60,60+)	*4*	*[200000,200000+)*	*4*
Income	*values*	*Want*	*values*
[0,10000)	*0*	*[0,3200)*	*0*
[10000,50000)	*1*	*[3200,6400)*	*1*
[50000,90000)	*2*	*[6400,9600)*	*2*
[90000,130000)	*3*	*[9600,12800)*	*3*
[130000,130000+)	*4*	*[12800,12800+)*	*4*
Assets	*values*	*Credit*	*values*
[0,30000)	*0*	*red*	*0*
[30000,80000)	*1*	*green*	*1*
[80000,140000)	*2*	*amber*	*2*
[140000,270000)	*3*	*Risk*	*values*
[270000,270000+)	*4*	*high*	*0*
		medium	*1*
		low	*2*

Using the 500 data of more than 600 lines of data given as the modal, and the rest will only be used in authentication and modification. Using the former two lines of data as the central point, we will do the K average analysis one by one. The data will be classified into two categories after several times of repetition.

For the sake of convenience, the ability conception is referred in this paper which equals the sum of income and assets, than minus debts. We can use the ability to pay as the criteria. We can conclude the table 3..

Table 3.

Age	*Want*	*credit*	*Risk*	*On-time*	*Ability*
24	*1500*	*red*	*high*	*1*	*34406*

Finally, Bayesian computation will be conducted to every group of data, as the following icon shows:

P(On-time=1)=0.95
P(On-time=0)=0.04
P(Age=4|On-time=1)=0.05
P(Age=3|On-time=1)=0.09
P(Age=2|On-time=1)=0.25
P(Age=1|On-time=1)=0.59
P(Age=0|On-time=1)=0.01
P(Age=0|On-time=0)=0.08
P(Age=1|On-time=0)=0.67
P(Age=2|On-time=0)=0.08
P(Age=3|On-time=0)=0.08
P(Age=4|On-time=0)=0.08
P(Want=0|On-time=0)=0.83
P(Want=1|On-time=0)=0
P(Want=2|On-time=0)=0.17
P(Want=3|On-time=0)=0
P(Want=4|On-time=0)=0
P(Want=4|On-time=1)=0.004
P(Want=3|On-time=1)=0.03
P(Want=2|On-time=1)=0.07
P(Want=1|On-time=1)=0.17
P(Want=0|On-time=1)=0.72
P(credit=0|On-time=0)=0.25
P(credit=1|On-time=0)=0.25
P(credit=2|On-time=0)=0.5
P(credit=0|On-time=1)=0.1
P(credit=1|On-time=1)=0.61
P(credit=2|On-time=1)=0.28
P(Risk=0|On-time=0)=0.58
P(Risk=1|On-time=0)=0.08
P(Risk=2|On-time=0)=0.42
P(Risk=0|On-time=1)=0.14
P(Risk=1|On-time=1)=0.026
P(Risk=2|On-time=1)=0.83
P(Ability2Pay=0|On-time=0)=0.75
P(Ability2Pay=1|On-time=0)=0.25
P(Ability2Pay=1|On-time=1)= 0.86

The simulation of the established model is extremely necessary. This paper picked one data from the left data and used this model to confirm the correctness of this model. The results are shown as following.

P(On-time=1)=0.95
P(On-time=0)=0.04
P(Age=1|On-time=1)=0.59
P(Want=0|On-time=1)=0.72
P(Age=1|On-time=0)=0.67
P(Want=0|On-time=0)=0.83
P(Ability2Pay=0|On-time=1)=0.13
P(Ability2Pay=0|On-time=0)=0.75
P(Risk=2|On-time=1)=0.83
P(Risk=2|On-time=0)=0.42
P(credit=2|On-time=1) =0.28
P(credit=2|On-time=0)=0.5
P(X|On-time=1)=0.01
P(X|On-time=1)*P(On-time=1)=0.01
P(X|On-time=0)=0.09
P(X|On-time=0)*P(On-time=0)=0.004

We can conclude that the result of On-time is 1.

The upper right corner is the original value of the On-time. While the final result is concluded using models. After computation, the conclusion is living up to our model. Which means that the model is feasible.

V. Conclusion

On the bases of basic Bayesian classification, this paper raise the Bayesian classification model set up on the K average computation method. Its function and conciseness are better than the traditional basic Bayesian classification. While further consideration should be given to the matter whether it is better than other classification method. The bank credit rating system though comprehensive in functions and conveniences in operation, still has some problems in the details of its practical use. This is needed to be further solved.

References

[1] M.Y.Kiang, A.Kumar, “A comparative analysis of an extended SOM network and K-means analysis”, International Journal of Knowledge-based and Intelligent Engineering Systems, IOS Pres, 2004, pp.9-15.

[2] Sergio M. Savaresi, Daniel L. Boley, “A comparative analysis on the bisecting K-means and the PDDP clustering algorithms”, Intelligent Data Analysis 8 (2004), pp.345-362.

[3] Hiroyuki Kashima, “An application of a minimax Bayesian rule and shrinkage estimators to the portofolio selection problem under the Bayesian approach”, Statistical Papers 46, 2005, pp. 523-540.

[4] Lili Liu, Andrew K. C. Wong, Yang Wang, “A global optimal algorithm for class-dependent discretization of continuous data”, Intelligent Data Analysis 8,2004, pp. 151-170.

[5] Tian Xiaoguang, Kong Dejing, “The Application of Data Mining on Credit Card Issuing”, Science & Technology Information, No.5. 2008, pp.64-66.

[6]Ge Jike, Zhao Yongjin, Wang Zhenhua, Yu Jianqiao, “Application of Data Mining Technique to Personal Credit Evaluating Model”, Computer Technology and Development, Vol. 16, No.12, 2006, pp.172-177.

[7] Hua Bei, “Research on credit card approval models based on data mining technology”, Computer Engineering and Design, Vol.29, No.11, 2008,pp.2989-2991.