ISSN 1746-7659, England, UK
Journal of Information and Computing Science
Vol. 4, No. 4, 2009, *pp*. 283-289

Cooperative Classification under the Protection of PrivateInformation *

Lu Fang + [1, 2], Zhong Weijun[1] and Zhang Yulin [1]

[1] School of Economic & Management , Southeast University, Nanjing 211189, China)

[2] Institute of Management Science and Engineering, Hunan University of Technology, Zhuzhou 412008, China)

(*Received December 18, 2008, accepted May 24, 2009,*)

Abstract. The private information of enterprises is a bottleneck to enterprises' cooperation. Enterprises often analyze cooperatively their consumers' information, but laws and their image require enterprises to protect consumers' information. To resolve the conflict between information-sharing and information-protecting, a privacy preserving classification method with distributed private information is proposed. We uses the Warner model to hide the true private enumerative data of enterprises' consumers information. Then we introduces how to get the exact classifying result on the disturbed data and analyze the method's accuracy and privacy in theory. In the end, the method's feasibility and validity is proved by experiments.

Keywords: Private Information, Enumerative Data, the Warner Model.

1. Introduction

Nowadays, companies are achieving two changes: from "Product-centric" to "Customer-centric", from the traditional marketing model 4P（Product, Price, Place, Promotion）to the new business strategy 4C(Consumer, Cost, Convenience, Communication). Therefore, customers become a very important asset, but for companies not every customer is a great value. The Pareto theorem proves 80% of enterprises' profits come from 20% of the customers. Therefore, enterprises need to find the characteristics of the most value-creating clients in order to facilitate targeted marketing.To compete with large enterprises, small and medium enterprises need analyze cooperatively customers' information to achieve the analysis results of the customers of large enterprises.

But, from their own point of view, enterprises need to protect customers' information. Firstly, if enterprises abuse customers' information, customers will provide false personal information, so that these enterprises make wrong decisions based on the false customers' information; Secondly, if customers' information is well protected, enterprises will gain more customers' trust, more profit and more customers' information like DELL and AMAZONE companies. What's more, protecting consumers' information is laws' regulation. In Oct 1, 2008, "the Consumer Protection Bill " in Shangdong province has put into practice. The bill prohibits the use of improper disclosure or access to customers' information and violators will be received legal punishment. China's first "Personal Information Protection Act" also started the legislative process, which legally guarantees the protection of customers' information with enterprises' responsibility.

Enterprises need analyze cooperatively customers' information, but enterprises' image and laws require them to protect customers information, which leads to the conflict of sharing customer information and protecting private customers' information. To solve this problem, Agrawal proposed privacy-preserving data mining in 2000[1]. He first proposed how to get the extract accurate models and rules in the non-precise data. Then on this basis, a number of researchers develop a great deal of privacy- preserving data mining algorithms [3~5][9]. However, these algorithms mainly focused on the central databases [6], very few about distributed databases [7]. Kargupta[7] proposed a distributed privacy-preserving data mining by secure multiparty computing. But it is only suitable for the numerical data and the method's computing complexity

* Supported by National Natural Science Foundation of China (70771026)
+ Corresponding author. Tel:+86-25-52084633. *E-mail address*: lufang31@126.com

is very high. Du[8] also proposed distributed privacy-preserving data mining based on the Warner model, which is only suitable for the binary data. In real life, there are many types of enumerative data, for example, color, education and gender, etc. So we study the customers classification method applicable for multi-attribute value of the distributed data.

2. Cooperative Classification under the Protection of Private Information

2.1. Introduction of the Method

Assuming there are two cooperative enterprises, customers data items were n_1, n_2 . Each enterprise has 4 customer attributes: age (value: Old, Medium, Young), income (value: High, Medium, Low), marital status (value: Yes, No), gender (value: Male, Female). The enterprises divide customers into two groups: A (high-value customers), B (low-value customers). The "age" and "income" attributes are customers' private information, and enterprises need protect these consumers' private attributes. Therefore, enterprises need to hide "age" and "income" attributes during classifing consumers cooperatively.

Based on the Warner model [10], we propose the modified method using the randomized response answering. After two enterprises agree upon the disturbed coefficient θ, every enterprise does the following:

（1）The numerical data is generalized according to a certain rule. For example, the numerical data "age" and "income" are converted into "Old, Middle, Young" and "High, Middle, Low".

（2）Generate n_i random numbers $r_i\ (i=1,\cdots,n_i)$. On the ith data item, if $r_i \le \theta$, the ith data item won't be changed. Or, the ith data item will be changed to one attribute value randomly from the attribute values. For example, if $r_i \le \theta$ and age=O*, income= H, we choose randomly (YH) in (OH, OM, OL, MH, MM, ML, YH, YM, YL) and the probability of choosing (YH) is $1/(3\times 3)$ (the first 3 is the number of attribute age values , the second 3 is the number of attribute income values) ;

（3）Submit the disturbed data to the third party of operating data mining.

2.2. Information Gain Computation

The most important part of creating a decision tree is computing information gain. We set $p^*(\,)$ as the probability of the disturbed data, $p(\,)$ as the probability of the original data.

For example, $p^*(\mathrm{OH})$ represents the probability of age=Old and income=High in the disturbed data; $p(\mathrm{OH})$ represents the probability of age=Old and income=High in the original data.

If we disturb only private attributes: "age" and "income", attributes have the following several situations when we compute the information gain of attributes:

(1)All private attributes

$$\begin{aligned} p^*(OH) &= p(OH)\times\theta \\ &+\big(p(OH)+p(MH)+p(YH)+p(OM)+p(MM)+p(YM)+p(OL)+p(ML)+p(YL)\big)\times(1-\theta)\times\frac{1}{9} \\ &= p(\mathrm{OH})\times\theta+(1-\theta)\times\frac{1}{9} \quad \Rightarrow \\ p(\mathrm{OH}) &= \frac{p^*(\mathrm{OH})-(1-\theta)/9}{\theta} \end{aligned} \tag{1}$$

(2)Part of private attributes

$$\begin{aligned} p(O) &= p(OH)+p(OM)+p(OL) \\ &= \frac{p^*(OH)+p^*(OM)+p^*(OL)-3\times(1-\theta)/9}{\theta} \\ &= \frac{p^*(O)-(1-\theta)/3}{\theta} \end{aligned} \tag{2}$$

(3) All private attributes + non-private attributes

$$p^*(OHA) = p(OHA)\times\theta$$

$$+\left(p(OHA)+p(MHA)+p(YHA)+p(OMA)+p(MMA)+p(YMA)+p(OLA)+p(MLA)+p(YLA)\right)\times(1-\theta)\times\frac{1}{9}$$

$$= p(OHA)\times\theta + p(A)\times(1-\theta)\times\frac{1}{9}$$

$$p(OHA) = \frac{p^*(OHA) - p(A)\times(1-\theta)/9}{\theta} \tag{3}$$

(4) Part of private attributes + non-private attributes

$$p(OA) = p(OHA) + p(OMA) + p(OLA)$$

$$= \frac{p^*(OHA) + p^*(OMA) + p^*(OLA) - 3\times p(A)\times(1-\theta)/9}{\theta}$$

$$= \frac{p^*(OA) - p(A)\times(1-\theta)/3}{\theta} \tag{4}$$

（5）All non-private attribute

$$p^*(\,) = p(\,)$$

We use the following common formula (5) to describe situation (1)~(4)：

$$y = \frac{x - p\times(1-\theta)/|s_v|}{\theta} \tag{5}$$

x is the probability of the converted data; y is the probability of the original data; $|s_v|$ is the product of hidden attribute value's numbers ; p is the probability of non-disturbed attributes.

2.3. Algorithm

Algorithm 1: Generate_decision_tree（S,T）;

input：customer data samples，the set of candidate attritute: attribute_list.

output：a decision tree.

Create node N;

If samples are in the same class C, then return N as a leaf-node.

If attribute_list is null, then return N as a leaf-node.

Choose the attribute with the highest information gain as the test_attribute，the choosing function is select(S,T);

Tag the node N as test_attribute;

For every attribute values si in test_attribute, generate corresponding branches.

Add si to S;

If si is null, return a leaf-node

Else (T=attribute_list-test_attribute;

And add the subtree of generate_decision_tree (S,T))

Algorithm 2: select(S, T)

Input: the attribute values list S, attribute list T.

Output: the attribute of the highest information gain;

（1）Do (2)~(5) for every attribute in attribute list T.

（2）compute the information expectations of customer classification according S:

According to (5), we can get $p(SA)=\frac{p^*(SA)-p\times(1-\theta 0/|s_v|}{\theta}$. In the same way we get $p(SB)$. So it's easy to compute $I(S_{1,}S_2)=-p(SA)\log_2 p(SA)-p(SB)\log_2 p(SB)$.

(3) To every attribute values si in T， set $a_i=s_i\cup S$.

According to formula (5) ， compute the information expectation of every ai : $I(a_iA,a_iB)=-p(a_iA)\log_2 p(a_iA)-p(a_iB)\log_2 p(a_iB)$;

（4） $E(A_i)=p(a_i)I(a_iA,a_iB)$;

（5） $Gain(A_i)=I(S_1,S_2)-E(A_i)$;

（6）return test_attribute=argmaxGain(Ai);

3. Performance Analysis

3.1. Accuracy Analysis

From formula (5), we can get

$$E(\hat{y})=E(\frac{x-p\times(1-\theta)/|s_v|}{\theta})=y \tag{6}$$

$$\mathrm{var}(\hat{y})=\frac{\mathrm{var}(x)}{\theta^2}=\frac{x\times(1-x)}{n\theta^2}$$

$$\lim_{n\to\infty}\mathrm{var}(\hat{y})=\lim_{n\to\infty}\frac{x\times(1-x)}{n\theta^2}=0 \tag{7}$$

Formula (6) and (7) show formula (5) satisfy unbiased and consistency requirements, so we can use formula (5) to educe the estimates of original samples.

In formula (5), x is the probability of disturbed data and it must have a value range, that is $0\le x\le 1$. We set $p/|s_v|=t$. Because $0\le p\le 1$ and $|s_v|>0$, from $p/|s_v|=t$ we can get $0\le t<1$. And from $0\le\hat{y}\le 1$， we can get $0\le\frac{x-(1-\theta)\times t}{\theta}\le 1 \Rightarrow (1-\theta)\times t\le x\le\theta+(1-\theta)\times t$. We use the following figure to indicate the *x* range.

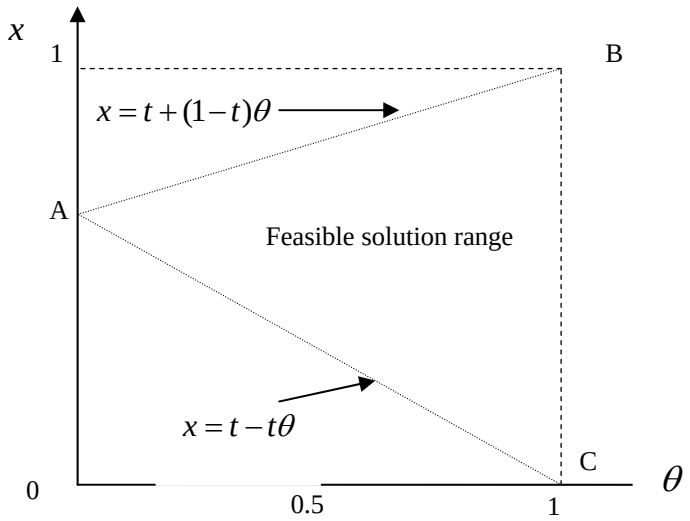


Figure 1 solution range

From figure 1, we can see when θ is closer to 1 the feasible solution range of *x* is greater and when $\theta=0$， the feasible solution range of *x* is 0. So from the method's feasibility,we hope θ is closer to 1.

DOI https://doi.org/2024-JICS-22737 | Generated on 2025-04-15 10:20:23

What's more, from $x = y \times \theta + p \times (1-\theta) / |s_v|$, when θ is closer to 1, x is closer to y, the accuracy is higher.

3.2. Privacy Analysis

Because $|s_v|$ is the product of hidden attribute value's numbers, formula (5) shows the more attribute is disturbed, the greater proportion of $|s_v|$ instead of real p, which leads to less accuracy. Formula (5) can be converted into the following formula:

$$x = y\theta + p(1-\theta) / |s_v| \quad (8)$$

In (8), when θ =1, y=x; and when θ=0, y is independent of x. That is, when θ is close to 0, the greater disturbance is, and the privacy is greater.

4. Experiment

Assuming two enterprises cooperate, and the numbers of customers' data items are 50 and 70. Every enterprise has 4 customer attributes: age(value:old, middle, young), income(value:high, middle, low), marital status(value:Yes, No), gender(value:Male, Female)). The enterprises divide customers into two groups:A (high-value customers), B (low-value customers). Attributes "age" and "income " are customers private information. As we only need to pay attention to several important customer characters, we set the height of the decision tree is 3.

The original decision tree is :

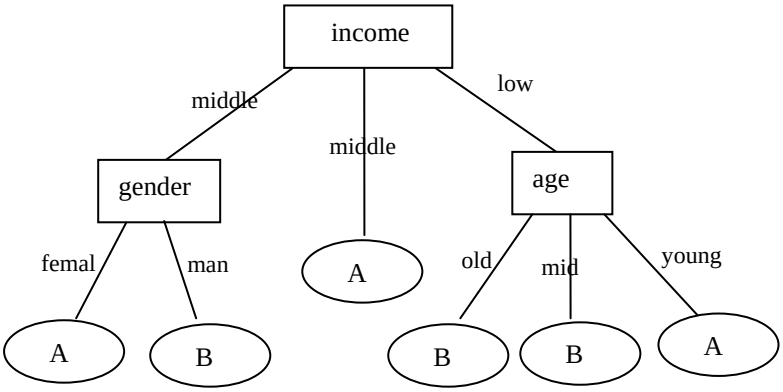
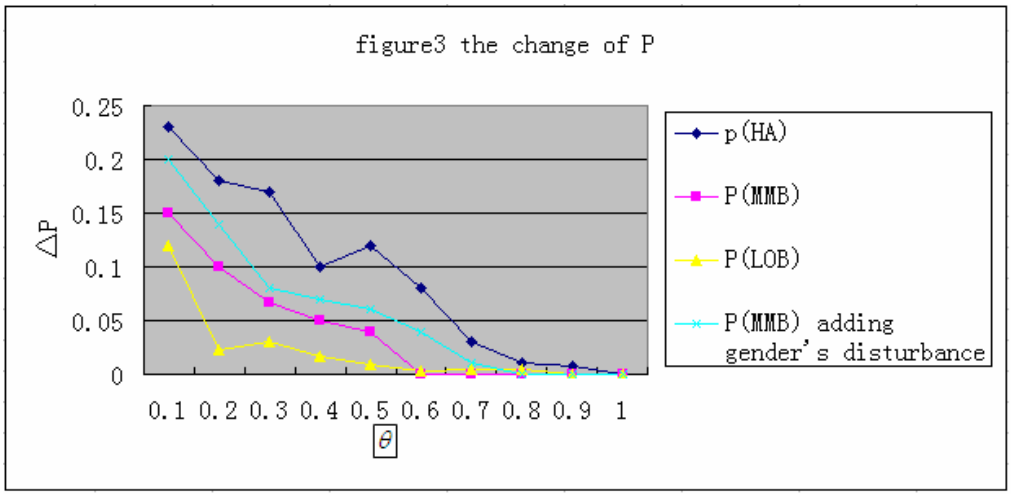


Figure 2 decision tree



Experiment 1: We set $\theta = 0.7$,perform algorithm 1 and algorithm 2, and get the decision tree. The decision tree is the same with figure 2. So the proposed method is feasible.

Experiment 2: We only hide attribute "age" and "income", θ changes from 0 to 1. The result is figure 3.

We can see θ is closer to 1, the change of attributes' probability is less and the accuracy is higher. And when θ in [0.65,1], p is close to 0. So enterprises can set θ in [0.65,1].

Experiment 3：We hide attribute "age", "income" and "gender", and compare the results. From figure 3, hiding "gender" doesn't affect p(HA) and p(LOB), but affect p(MMB). We can also see hiding "gender" increases the $\Delta p(MMB)$,that's decreases the accuracy.So, if possible, we should hide less attributes to get higher accuracy.

5. Example Analysis

Assuming a shop sells a product with a price 1000yuan. The cost including storage and rent is 500yuan every day and every product. The promotional leaflet is 2yuan for every customer; the net income is 500yuan if a customer buys a product. The number of people in the shop is 1000 and the response bottle line of customers is 1%. The shop can adopt the following projects.

If the shop promote the product to all shopping customer:

The promote cost : 1000*2=2000yuan;

The income of every response :1000*1%*500=5000yuan;

The whole income a week : 7*（5000-2000）=7*2100=14700yuan;

If the shop mines their customers information with another company a week, the cost of mining is 1000yuan a week and a company. The shop can promote the customer satisfying the mining rules. The customer satisfying the mining rules is the 30% of the number of whole shop customers. And the response increases to 3%.

The promote cost: 1000*30%*2=600yuan;

The income every response：1000*30%*3%*500=4500yuan;

The income a week :7*（4500-600）-1000=26300yuan;

So if we take project 2, we can make a higher profit: 26300-14700=11600yuan.

6. Concluding

The private information often hinders the enterprise information-sharing cooperation. To resolve the conflict between information-sharing and information-protecting, the paper uses the Warner model to hide the customers' private information with multi-value attributes. The method is simple and the accuracy is high. What' more, the cooperative enterprises can adjust the result's accuracy.

But, this mehod is only applied to enumeratied data type. if customers' information data is numerical , then the numerical data need to be converted. So, a common method without the restrictions of data type is a directions for furture research.

7. References

[1] R. Agrawal, and R. Srikant., *Privacy-preserving data mining*. Proc. Of the ACM SIGMOD, pp. 439-450, 2000.

[2] J. Michael, A. Berry, and S. Gordon, *Data mining techniques: for marketing, sales, and customer relationship management*, China machine press, 2006.

[3] W. Du , and Z.Zhan, *Using randomized response techniques for privacy-preserving data mining*. Proc. Of the ACM SIGMOD, pp. 24-27,2003.

[4] E. Alexandre, R.Srikant, and R. Agrawal, etc, *Privacy Preserving Mining of Association Rules*, Proc. Of the ACM SIGKDD,PP.343-364,2004.

[5] F. Lu, W. Zhong, and Y. Zhang, etc. *Privacy-preserving association rules mining using the grouping unrelated-question model*. Proc. of Wireless Communications, Networking and Mobile Computing, pp. 5589-5592,2007.

[6] K. Muralidhar, and R. Sarathy, *A theoretical basis for perturbation methods*. Statistics and Computing 13(2003),PP.329–335.

[7] H. Kargupta, K. Liu, and J. Ryan, *Privacy sensitive distributed data mining from multi-party data*, Proc of the first NSF/NIJ Symposium on Intelligence and Security Informatics, PP. 336-342,2003.

W. Du, and Z. Zhan, *Building Decision Tree Classifier on Private Data*, Proc. of IEEE Int'l Conf. Privacy, Security,

and Data Mining, 2002: 1-8.

[8] S. Oliveira, and O.Zaiane, *Privacy preservation when sharing data for clustering.* Proc of the International Workshop on Secure Data Management in a Connected World, 2004: 67–82.

[9] S. Warner. *Randomized response: a survey technique for eliminating evasive answer bias*, the American Statistical Association, 60(1965): 63-69.