Designing of Privacy Protection Platform Based on Data Mining

ZhouBing[1,a], Zeng Zhihua[1]

[1]City college Wuhan university of science and technology, Hubei Wuhan, 430000

[a]zhoubingzb1101@163.com

Keywords: Privacy Protection; Mechanical Analysis; Optimal Design Data Mining

Abstract. How to protect sensitive data privacy is one of the research focus in the field of data mining, especially in the case of the data distributed storage, and the meaning of data privacy protection is particularly important. Secure multi-party computation security multiparty computation password primitives in the privacy of distributed data mining related applications, the literature [1] of privacy protection data mining model based on security multiparty computation were analyzed, and the argument of this model is based on discrete logarithm public-key encryption protocol is not fully homomorphic characteristics, with a simple example. Thus, it is concluded that the privacy data mining model is not feasible.

Introduction

Multivariate statistical analysis and clustering method as the commonly used data mining techniques, in commercial, medical and other fields of knowledge discovery has important application. For example, multiple linear regression analysis through the establishment of forecast data driven and response variables, the regression model between the so as to realize the prediction model of variable [1]. Obtain accurate knowledge mining, however, at the same time, how to keep the original data of sensitive (e.g., customer information, patient information) are not leaked has become an important issue in the field of data mining. Privacy of data mining, privacy preserving data mining, and PPDM) is designed to ensure the reliability of the data mining process, at the same time avoid because of the information to participate in the body of the mining damage. In recent years, the research of PPDM technology mainly through two ways: data disorder and secure multi-party computation based on cryptography (security multiparty computation, SMC). Due to data disturbing method is easy to receive various types of attacks, and to meet the privacy and adding disturbing data can make the digging, according to the correctness of the result affected by certain so the method development is relatively slow. And secure multi-party computation method based on cryptography can be sensitive to in the process of digging, according to data encryption protection, and different kinds of privacy protection data digging according to the problem can be converted to a specific multilateral security task. So, according to the requirements of the specific dig according to design the corresponding SMC protocol become the mainstream research PPDM technology direction.

Designing of Privacy Protection

Anonymous wrong technical data mining result, also not the original data to disguise, but released with all data privacy, but others get privacy data but cannot deduce the owner's identity. Specific methods can be divided into the following two categories: Attribute set (1) to protect privacy Data released a single node, the node identifier portion is not encrypted, privacy properties section to separate the encryption. System after the collected data of each node can't see the privacy of each node data, and can only see the node id attribute data. System will classify the collected all the identity of the property, when the statistics of a node id attribute repetitions over iK in the whole system, system can decrypt according to iK the privacy of the ith node attributes. (2) the hidden identity property collection nodes to participate in the system of data mining, through a system algorithms require each node identification rules are presented. (3)With the constant promotion of the data mining technology, data mining if misused, so our social life might have been seriously affected. Huge amounts of data in the online community can provide many users greatly help, it is likely they will abuse of data mining technology, mining the information they want, but I don't pay attention to protecting the privacy of others. Therefore, for data mining in online community behavior must carry on the strict regulation, the privacy of data mining technology also needs to be vigorously promoted. Both use privacy protection technology in data mining and the regulation of data mining is the need for additional costs, how to reduce the cost efficiency is also a big challenge. Different methods of privacy protection at different levels of privacy protection, which does not have a clear need to how much privacy protection rules, should set up an evaluation system of data mining for protection of privacy and quantitative criteria.

In the given framework, the selection of the h initial nodes is divided into two stages: the enlightening stage and the greedy stage. In the enlightening stage, select the node with the maximum value of PI, and in the greedy stage, select the node with the "most influential" node. The inspiring factor $c(c \in 0,1)$ is introduced in the framework, $[ck]$ shows the step number of the greedy stage, and $k-[ck]$ shows the step number of the enlightening stage. Obviously, when there is $c=1$, the method in the framework is the HH algorithm.

In the linear threshold model, u_{uv} represents the existed impact of the activated node u on its adjacent node v, which is estimated by $1/d(v)$ shows the out-degree of node v) and means that the influences of all the neighbor nodes on node v are the same. Obviously, this assumption ignores the differences between nodes, and does not conform to the reality. Here, according to the different characteristics of the social network, different u_{uv} estimation formulas are given.

u_{uv} Estimation Formula on the Un-weighted Graph

The size of u_{uv} is a feeling of the node v itself, this feeling is a reflection of the authority of node v to the node that is pointed to node v, and has nothing to do with other nodes. Thus, there is only need to consider the structure relationship of the neighbor nodes of node v. The Neighbor Graph (NG) is made up of the node v, the neighbor nodes that are pointed to node v, and the relationship between the nodes. The degree of the nodes in the u_{uv} estimation formula is obtained according to the degree of the nodes in the neighbor graph. Its definition is given in formula (1).

$$NG(v) = G'(V',E'), V' = \{v\} \cup N(v),$$
$$E' = \{(x,y) \mid x,y \in V', (x,y) \in E\},$$
$$u_{uv} = \frac{outDeg(u)}{\sum_{w \in N(v)} outDeg(w)} \qquad (1)$$

Among which, $outDeg(u)$ represents the out-degree of node u in the neighbor graph, and $N(v)$ shows the set of the into-side neighbor nodes of node v in the neighbor graph. The influence of node u on node v is mainly decided by the neighbor structure of node v.

Figure 1 shows a simple example of the neighbor graph, in which dark node v, node u_1, u_2 and u_3 form a neighbor sub-graph. The out-degree of u_1, u_2 and u_3 in NG (v) is respectively 1, 1 and 2. Thus, the value of u_{u1v}, u_{u2v} and u_{u3v} is respectively 0.28, 0.28 and 0.7.

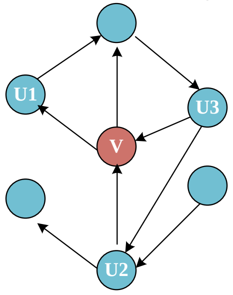

Figure1.The Neighbor Graph of Node v

Under the condition of considering the weighted edge, the influence of node u on node v is mainly determined by the weight of the edge. The definition of u_{uv} is specified as follows:

$$u_{uv} = \frac{s(u,v)}{\sum_{w \in N(v)} s(s,v)} \qquad (2)$$

Among which, $s(u,v)$ represents the weight of the edge (u,v), and $N(v)$ shows the set of the into-side neighbor nodes of node v in the neighbor graph.

Figure 2 shows a simple example, in which the weight of edge (u_1,v), (u_2,v) and (u_3,v) is respectively 2, 5 and 3. According to formula (2), the calculated value of u_{u3v}, u_{u2v} and u_{u1v} is respectively 0.3, 0.6 and 0.4.

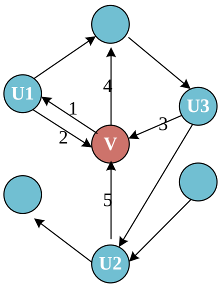

Figure 2. The Weighted Neighbor Graph of Node v

Test results

First of all, inspect the joint effect of the inspiring factor c and the size of the target set k, namely the influence of different values of c on the range of influence with the same value of k. The results of the experiment on the data set 1 are shown in figure 3. It is can be known from the figure that for different values of k, most of the ranges of influence with other values of c are bigger than the range of influence when the value of b is 1, except for the condition of b=0. When the value of b is 0.6 and the value of k is 60, the range of influence of the algorithm under the framework is about 10% higher than that of the greedy algorithm. When the value of c is 0, all the initial nodes are statically chosen from the nodes with the largest value of PI, which fails to consider the propagation process of influence, thus, its range of influence is the worst. In the following experiments, the condition of b = 0 is ignored.

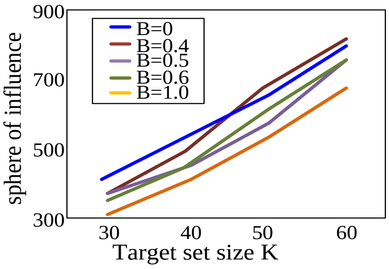

Figure 3.The Influence Curves with Different Values of k and c on Data Set 1

Conclusion

Privacy is one of the important topics in the area of data mining, and in order to achieve accurate knowledge discovery does not leak sensitive raw data at the same time as the goal. [1-2] trying to design based on the homomorphic multiparty secure cryptograph computing, privacy protection and build data mining model, to ensure that all participants of the original data without trusted third party in the process of data mining is not leaked. Such with the efficient design of SMC protocol to complete data privacy protection method is one of the developing direction of PPDM technology, is a meaningful attempt. But [1-2] are based on discrete logarithm problem of encryption protocol (CBSDLP) design of cipher algorithm is not fully homomorphic characteristics. Therefore, on the basis of building the privacy of multiple linear regression model and the clustering of privacy protection model is feasible.

References

[1]Y. Geng, J. Chen, K. Pahlavan, Motion detection using RF signals for the first responder in emergency operations: A PHASER project, 2013 IEEE 24nd International Symposium on Personal Indoor and Mobile Radio Communications (PIMRC), London,Britain Sep. 2013

[2]Y. Geng, J. He, K. Pahlavan, Modeling the Effect of Human Body on TOA Based Indoor Human Tracking[J], International Journal of Wireless Information Networks 20(4), 306-317

[3] Du W L, Han Y S. Privacy preserving multivariate statistical analysis: linear regression and classification[C]//Proc of the 4th SIAM International Conference on Data Mining. Florida: IEEE, 2004:222-233.

[4] VAIDYA J，CLIFTON C. Privacy preserving K-means over vertically partitioned data [C]//Proc of the SIGKDD International Conference on Knowledge and Data Mining. New York: ACM, 2003:490-510.

[5] KENNETH H. ROSEN. Elementary number theory and its applications（fifth edition）[M]. New York：Person Education，2005:245.