International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 09 | Sep 2020 www.irjet.net p-ISSN: 2395-0072

DECISION TREE LEARNING TECHNIQUE FOR MULTI-RELATIONAL CLASSIFICATION IN INFORMATION LEAKAGE PREVENTION SYSTEM

Alese Boniface Kayode[1], Adewale Olumide Sunday[2], Alowolodu Olufunso Dayo[3],

Adekunle Adewale Uthman[4], Makinde Akindeji Ibrahim[5]

[1]Professor, Dept. of Cyber Security, Federal University of Technology, Akure, Ondo State, Nigeria
[2]Professor, Dept. of Computer Science, Federal University of Technology, Akure, Ondo State, Nigeria
[3]Lecturer, Dept. of Cyber Security, Federal University of Technology, Akure, Ondo State, Nigeria
[4]Student, Dept. of Computer Science, Federal University of Technology, Akure, Ondo State, Nigeria
[5]Student, Dept. of Computer Science, Federal University of Technology, Akure, Ondo State, Nigeria

---

Abstract - *In a complex database schema, it is always difficult to identify all the relationship among various attributes. A multi-relational classification technique helps to identify all the multiple interlink relationship in a relational database in which a model is generated based on training dataset and the model is used for predicting a set of unknown data. This paper uses a multi-relational classification for information leakage prevention. The result of the generated ranked list of subschemas maintains the predictive performance and predictive accuracy of confidential attributes.*

Key Words: multi-relational classificaiton, relational database, information leakage prevention, decision tree, database schema.

1. INTRODUCTION

The widespread use of information technology and cloud computing has increased the threats of data leakage, manipulation and distribution. While the world is increasingly relying on the cloud computing, the security and privacy concerns of the data stored and shared on the distributed networked are also increasing (Subashini, 2011). As reported by BBC in a report of 2014, many enterprises had been prey to the theft, loss, leakage, manipulation of the sensitive business data (Arora, 2017). Such kind of sensitive direct and indirect data loss of an individual or a corporate poses a great threat on the business reputation and trust on ad individual (Norberg, 2007). Out of many possible and obvious reasons of data leakage and loss, the most commonly reported reason is the casual and negligent behaviour of employees while interacting with the shared data or files of a corporate which are shared through a covert channel (Williams-Banta, 2019).

Large percentage of today's real-world data is structured i.e. such data has no natural representation in a single tabular table and instances in these data are represented by structured terms than static feature vectors (Alphonse. 2004). Information encoded in structure of the data needed to be taken into account when learning from structured data since such structure represents how different objects in the data relate and demonstrate some useful patterns in mining tasks (cooley, 2000). Multi-relational algorithms search for patterns across multiple related tables in a relational data.

To classify objects in one relation, other relations provide important information. To classify data from relational data there is a need of multi-relational classification which is used to analyze relational data and used to predict behavior and unknown pattern automatically. Classification in data mining is a two-stage process (Macskassy, 2007). Stage one learn classification model from training dataset. While stage two classify testing set using the classification model. In multi-relational classification, given a collection of relational data, each table contain a set of attribute, only target table contains class attribute. The goal of classification is to predict a class to unknown test set accurately.

2. RELATED WORK

Agrawal and Srikant (2000) present privacy-preserving data mining that address the problem of privacy preserving data mining. The work is motivated by the need to both defend confidential information and enable its use for research or other purposes. The problem stated above was solved using known generic protocols. Despite the solution, data mining algorithms are complex and, the input usually having large data sets.

Agarwal et. Al. (2012) designed a Robust Data leakage and Email Filtering System which arise from the common applications such as email and other Internet channels. The Electronic Mail filtering was implemented based on fingerprints of the message bodies, the black and white lists of the email addresses and the words exact to spam. The research work also emphasized that distributor need to estimate odds that disclosed records corresponding data leakage from one or more agents. The research used the data allocation methods or injecting "realistic but fake" data records for increasing detection of the leakage.

Shu et. Al. (2016) authored a Fast Detection of Transformed Data Leaks which focused on inadvertent leak

detection. Identifying the exposure of sensitive data was difficult due to data transformation in data content. In the model designed, two types of sequences where analyzed i.e. sensitive data sequence that requires to be protected from an unauthorized parties and content sequence which is to be examined for leaks. The content may be data extracted from the file systems on workstations, distributed system or personal computers from supervised network channels. The sensitive data sequences are known to the analysis system. The sensitive data sequences utilized sequence alignment approaches for detecting the complex data-leak patterns which are known to the analysis system.

Zang et. al. (2005) presented a privacy preserving naive bayes classification. In many scenario, data is divided between multiple organization. The organizations may want to utilize all of the data to create more accurate predictive models. The paper make use of Naive Bayes Classifier which is a simple but efficient baseline classifier.

Evfimievski et al. (2004) presented a framework titled Privacy preserving mining of association rules for mining association rules from transactions consisting of categorical items where the data has been randomized to preserve privacy of individual transactions. Even though it is reasonable to recover association rules and preserve privacy using a straight forward randomization, the rules that are found leads to privacy leaks.

3. DATA LEAKAGE PREVENTION IN MULTI-RELATIONAL CLASSIFICATION

The aim of this approach is to prevent the prediction of confidential attributes and at the same maintaining the predictive performance of the target variable.

A relational database can be described by a set of tables where each tables constitutes a set of tuples. In each table, a row is used to describe a single record while column is used to show values of some attribute in the table. Multi-relational pattern involves a relational database and can be seen as pieces of subschema which we want to meet in the scheme of the objects that we are considering. The patterns with a large support that are above some predefined threshold is known as frequent.

Adapting the construction approach presented Guo et. al. (2010), the method consists of the following four stages.

1. Firstly, correlated confidential are identified. The term correlation to refer to the associations and interrelationships between attributes in the database.

2. Secondly, the degrees of sensitivity for different subschemas of the database are calculated based on the correlation computed from the first stage.

$$P = \frac{\sum_{j=1}^{n} TC}{TS} \qquad (1)$$

Where P is the degree of sensitivity, TC is the number of tuples covered by each rule, and TS is the total number of tuples

3. Thirdly, construction of subschemas consisting of different tables of the database

$$I = \frac{k\overline{V_{cf}}}{\sqrt{k+k(k-1)\overline{V_{ff}}}} \qquad (2)$$

Where I is the subschema informativeness, k is correlation between the sub-graphs, $(V\_cf)^{-}$ is the average correlation between the variable and the target variables, and $(V\_ff)^{-}$ is the average dependence between the variables themselves

4. Fourthly, computation of individual subschema performance when predicting the target variable and privacy sensitivity level thereby leading to a ranked list of subschemas.

$$Pr = I * (1 - P) = \frac{k\overline{V_{cf}}\,(1-P_k)}{\sqrt{k+k(k-1)\overline{V_{ff}}}} \qquad (3)$$

Where Pr is the value of a subschema, I is the target variable and P is the confidential variable.

Table -1: The System Multi-Relational Classification Algorithm

The System Multi-Relational Classification Algorithm
Input: a relational database (target variable , confidential attribute)
Output: a ranked list of subschemas of relational database.
construct a set of high quality rules using
derive the subschema privacy sensitivity (S) from the set of rules learned
convert schema into undirected graph
construct a set of subgraphs
for each subset **do**
compute the privacy-informativeness (PI) of the subcheme
end for
rank the subschema based on their privacy-informativeness values
return the ranked subschema

4. EXPERIMENT AND RESULTS

In this section, we evaluate the performance of the information leakage prevention in multi-relational classification on the dataset generated from the system database as shown in Fig. 1 The one the university cooperative society database consists of nine tables, Members, Staff, Account, Payslip, Demographic, Card Details, Pension, Savings, and Loan as shown in Figure 1. The target attribute is the Loan Flag in the Loan table. The loan flags are represented in the table with One (1) for Not Approved, Two (2) for See committee, Three (3) for complete previous loan, and Four (4) for Approved

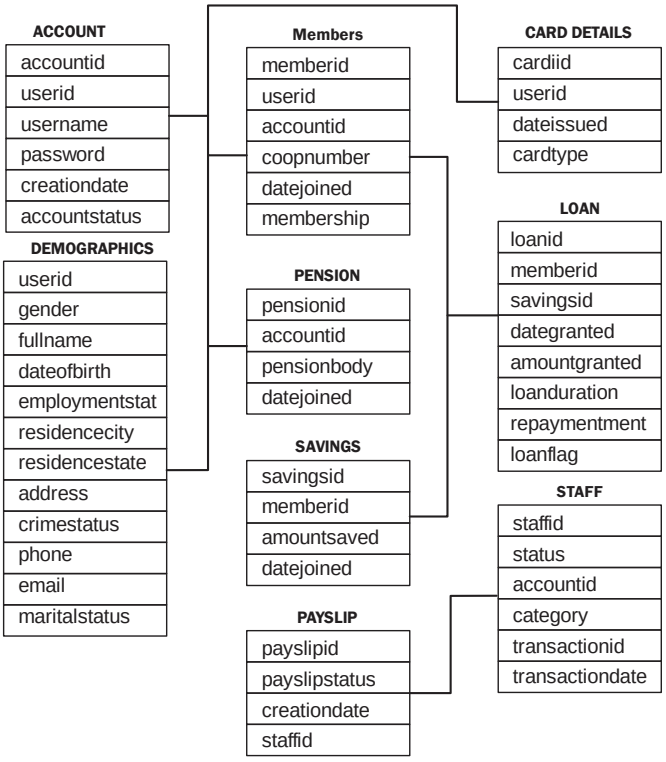

Fig -1: Cooperative Database

In this experiment, transactionid, transactiondate and repayment are very confidential records. This are association with loan repayment. There are about 2,230 transaction records in the database and much of it are on loan repayment. Therefore, there is need to protect confidential information of the members. Loan transaction records are being shifted from the loan table to the payslip table, and other traces kept in the staff table.

The degree of sensitivity for each subschema as shown in Table 2 is computed using Equation 1, and the result is presented in Table 3. Each subschema sensitivity is determine by the number of tuples. The subschema which consists of tables {Account, Members, Loan, Demographic, Card detail, payslip} has the highest degree of sensitivity.

Table -2: Tuples for some selected rules

Subschemas	**Tuples**
{Account, Members, Loan}	540
{Account, Members, Loan, Demographic}	1203
{Account, Members, Loan, Demographic, Card details}	1530
{Account, Members, Loan, Demographic, Card detail, payslip}	2,230
{Account, Members, Savings}	2,100
{Account, Members, Savings, Staff}	30
{Account, Members, Loan, Staff}	18
{Account, Loan, Payslip}	650
{Account, Members, Loan, Pension}	320
{Account, Members, Pension}	180

Table -3: Privacy Sensitivity of subschemas

Subschemas	Tuples
{Account, Members, Loan}	0.240
{Account, Members, Loan, Demographic}	0.540
{Account, Members, Loan, Demographic, Card details}	0.690
{Account, Members, Loan, Demographic, Card detail, payslip}	1.000
{Account, Members, Savings}	0.940
{Account, Members, Savings, Staff}	0.013
{Account, Members, Loan, Staff}	0.008
{Account, Loan, Payslip}	0.291
{Account, Members, Loan, Pension}	0.143
{Account, Members, Pension}	0.081

Table 4 represents the subschemas generated from the cooperative database and it shows the tested results against the target label. It also shows the confidential attribute. The multi-relational classification created a list of subschemas with various predictive capability against the target variable and the confidential attribute. Table 4: The top 10 ranked subschemas and their accuracies obtained against the target and sensitive attributes, for the cooperative database

Table -4: The top 10 ranked subschemas and their accuracies obtained against the target and sensitive attributes, for the cooperative database

Subschemas selected for release	**Target**	**Sensitivity**
{Loan, payslip, Member}	83.7	61.2
{Loan, payslip, Member, Demographic}	83.2	58.9
{Loan, Payslip, Members, Demographic, Savings}	80.9	61.8
{Loan, Payslip, Members, Demographic, Savings, Account details}	80.9	59.8
{Loan, Payslip, Members, Demographic, Account details}	81.5	62.4
{Loan, Payslip, Members, Account details}	83.0	50.1
{Loan, Demographic, Account details}	81.1	60.5
{Loan, Demographic, Savings, Account details}	80.2	62.1
{Loan, Savings, Account details}	78.9	51.4
{Loan, Savings}	76.9	51.1
All table in the database	**83.7**	**62.7**

With Table 4, one is able to identify dangerous subschema that pose a high data leakage risk. The confidential attribute and the accuracy attribute for the subschema containing

table{Loan, Payslip, Members, Account details} drops from 62.7% to 50.1% while the accuracy against the target is higher than that against the full database schema. In this case, the database administrator will determine if the high leakage is acceptable or not. Subschema containing tables {Loan, Payslip, Members, Demographic, Account details} and the subschema containing {Loan, Demographic, Savings, Account details} can be used to predict the sensitive attribute with an accuracy of over 62%. In summary, subschemas can be used to attack the confidential attribute with accuracy only slightly lower than 83.7% against the original, full database.

5. CONCLUSION

In this section this research work included literature review of works on multi-relational classification based on several methods. The proposed method generated ranked list of subschemas which maintain the prediction performance and predictive accuracy of confidential attributes leading to gives more accurate results.

REFERENCES

[1] S. Subashini, and V. Kavitha (2011). A survey on security issues in service delivery models of cloud computing. Journal of network and computer applications, 34(1), 1-11.

[2] A. Arora, and A. Mendhekar (2017). Threats to Security and privacy of Information due to growing use of social media in India. Asian Journal of Managerial Science, 6(2), 42-49.

[3] P. A. Norberg, D. R. Horne, and D. A. Horne (2007). The privacy paradox: Personal information disclosure intentions versus behaviors. Journal of consumer affairs, 41(1), 100-126.

[4] P. E. Williams-Banta (2019). Security Technology and Awareness Training; Do They Affect Behaviors and Thus Reduce Breaches? (Doctoral dissertation, Northcentral University).

[5] É. Alphonse, and S. Matwin (2004). Filtering multi-instance problems to reduce dimensionality in relational learning. Journal of Intelligent Information Systems, 22(1), 23-40.

[6] R. W. Cooley, and J. Srivastava (2000). Web usage mining: discovery and application of interesting patterns from web data. Minneapolis, MN: University of Minnesota.

[7] S. A. Macskassy, and F. Provost (2007). Classification in networked data: A toolkit and a univariate case study. Journal of machine learning research, 8(May), 935-983.

[8] R. Agrawal, and R. Srikant (2000, May). Privacy-preserving data mining. In Proceedings of the 2000 ACM SIGMOD international conference on Management of data (pp. 439-450).

[9] M. Agarwal, K. Gaikwad, and V. Inamdar, "Robust Data leakage and Email Filtering System," in International Conference on Computing, Electronics and Electrical Technologies (ICCEET), IEEE, 2012, pp. 1032-1035.

[10] P. Zhang, Y. Tong, S. Tang, and D. Yang (2005, July). Privacy preserving naive bayes classification. In International Conference on Advanced Data Mining and Applications (pp. 744-752). Springer, Berlin, Heidelberg.

[11] A. Evfimievski, R. Srikant, R. Agrawal, and J. Gehrke (2004). Privacy preserving mining of association rules. Information Systems, 29(4), 343-364.

[12] X. Shu, J. Zhang, D. Yao, and W. C. Feng, "Fast Detection of Transformed Data Leaks," IEEE Transactions on Information Forensics and Security, vol. 11, no. 3, pp. 528-542, March 2016.

[13] H. Guo, H. L. Viktor, and E. Paquet (2010, December). Identifying and preventing data leakage in multi-relational classification. In 2010 IEEE International Conference on Data Mining Workshops (pp. 458-465). IEEE.