International Journal of Engineering &Technology, 7 (4) (2018) 6862-6865

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET
doi: 10.14419/ijet.v7i4.27434
Research paper

Privacy preservation of class association rules and its optimization by utilizing genetic algorithm

Darshana H. Patel [1] *, Dr. Saurabh Shah [2], Dr. Avani Vasant [3]

[1] *Head and Assistant Professor, Department of Information Technology, V. V. P. Engineering College, Gujarat, India*
[2] *Director and Professor, Department of Computer Engineering, C.U. Shah University, Wadhwan, Gujarat, India*
[3] *Professor and Head, Department of Computer Science and Engineering, Babaria Institute of Technology, Gujarat, India*
**Corresponding author E-mail: darshana.h.patel@gmail.com*

Abstract

With the advent of various technologies and digitization, popularity of the data mining has been increased for analysis and growth purpose in several fields. However, such pattern discovery by data mining also discloses personal information of an individual or organization. In today's world, people are very much concerned about their sensitive information which they don't want to share. Thus, it is very much required to protect the private data. This paper focuses on preserving the sensitive information as well as maintaining the efficiency which gets affected due to privacy preservation. Privacy is preserved by anonymization and efficiency is improved by optimization techniques as now days several advanced optimization techniques are used to solve the various problems of different areas. Furthermore, privacy preserving association classification has been implemented utilizing various datasets considering the accuracy parameter and it has been concluded that as privacy increases, accuracy gets degraded due to data transformation. Hence, optimization techniques are applied to improve the accuracy.

Keywords: *Data Mining; Associative Classification; Privacy Preserving Data Mining; Optimization; Genetic Algorithm.*

1. Introduction

With the technological revolution, a huge amount of data is being collected and as a consequence data mining technologies are used for the extraction of useful information from huge compilation of digital data. However, this immense quantity of data, if publicly available, can be employed for growth and development as well as in several applications. An enormous amount of data is processed to obtain certain useful information in the area of data mining. Major techniques of data mining are classification, clustering and association [1].

Classification is an important technique which is used in our day to day life. Classification process is a supervised learning process that classifies by keeping the target class into consideration. Various classification methods such as Decision Tree, Naive Bayesian classifier, Support Vector Machine, Neural Network, Associative Classification etc. exists [1].Amongst Classification, associative classification technique is preferred due to its simplicity and interpretability characteristics[3]. After classification task, such data is used for analysis and prediction. But at the same time that data contains private or sensitive information due to which sometimes there might be inappropriate data available that leads to low mining result. Hence, to solve this problem, there subsists a field of privacy-preservation of data [6].

Privacy-Preserving technique preserves the private or sensitive information of an individual and at the same time utility of that data is also maintained [2]. Hence, focus is made on the construction of class association rules generated by associative classification and by applying privacy-preserving techniques on these rules to prevent its disclosure to uncertified population or nation [11]. Thus, privacy preserved class association rules are formed which protects the sensitive patterns from being disclosed.

Privacy-preserving associative classification (PPAC) covers a wide range of applications such as recommended system used in online shopping credit card fraud detection, phishing detection video surveillance, , Stock trading data that helps in finding signals to sell and buy, e-mail authorship, etc.[2] [5] [6].

Optimization technique is used to find the better solution either maximizing or minimizing depending on the problem [10]. Here, optimization methods are used for maximizing the outcome or accuracy as it is degraded while preserving the privacy.

The structure of this paper is: This section contains the basic thought of optimizing privacy preserved class association rule. Section 2 includes generation of class association rules. Section 3 provides generation of privacy preserved class association rules. Section 4 constitutes optimization Section 5 includes implementation and discussions. Section 6 contains conclusion and future scope.

2. Generation of class association rules

Class association rules (CARs) are type of rules which provides association between frequent items in such a way that frequent items are on the antecedent part and class is on the descendent part of the rule [15]. CARs can be produced by various associative classification algorithms. Several algorithms exist for associative classification namely Classification based on Multiple Association Rules (CMAR), Classification based on Association Rules Generated in a Bidirectional Approach (CARGBA)Classification based on Predictive Association Rules (CPAR), Classification based on Association (CBA) [4] [12]. Amongst these, CBA algorithm has

been utilized for the generation of CARs as it is simple algorithm and provides better accuracy as compared to C4.5 [15].
Classification based on association (CBA) comprises of three steps. The first step includes continuous attributes upon which discretization has been performed. In second step, using CBA-RG algorithm, class association rules (CAR) are generated, and in the third step CBA-CB is used to build a classifier which makes usage of newly formed CARs [3].
The core practice of CBA-RG algorithm is to determine each rule items that consist of support greater than minimum support [3]. The CBA-CB algorithm yields the supreme classifier from the total package of set of laws which comprises a calculation of every practicable parts of it from the training data and the part had been chosen whose rule sequence is correct, furthermore with the target of generating minimum number of errors [3]. There exits 2m such parts, where m stands for quantity of rules [3]. This algorithm follows heuristic approach and the classifier built by CBA performs better compared to C4.5
For building CAR, CBA-RG called rule generator algorithm is used. The main task of CBA-RG is to discover all the rules whose support is greater than specified minimum support [3]. A rule item is of the form:
f_items -> c
Where, f_items is frequent items and c is class label.
The support of a rule can be calculated as
(rulesupCount / |D|) *100% [3]
Where |D| is the dataset.
The confidence of a rule can be considered as,
(rulesupCount / f_itemssupCount)*100% [3]
The advantage of CBA is that it is simple algorithm which provides useful class association rules. Also it has some limitation like training the dataset often generates large number of rules leading to redundancy and it has the issue of over fitting [5]. This issue of over fitting can be solved by pruning strategies such as support, confidence, coverage etc.

3. Generation of privacy preserved class association rules

Privacy preserved association rules (PPCARs) are types of CARs in which sensitive rules are protected through privacy preserving techniques [6]. Many privacy preserving techniques namely Anonymization, Randomization, Cryptography etc. [5] [11] exists. Anonymization has been employed as it is more flexible compared to other existing techniques [11]. It is a process of removing/modifying the identifying variables or identifiers contained in the micro-data dataset.
Identifiers can be of the form explicit, quasi, sensitive and non-sensitive [8]. Explicit are those identifiers which provide direct specific information. (Example: name, phone number etc.) Quasi-identifiers can be single attribute or combination of attributes which leads to the specific identification. (Example: age, occupation etc.) Sensitive attributes are such attributes which reveals the private information of a specific individual. Non-sensitive attributes if disclosed in any condition may not lead to release the privacy of any individual.
The process of anonymization can be described visibly as given in below figure 1&2:

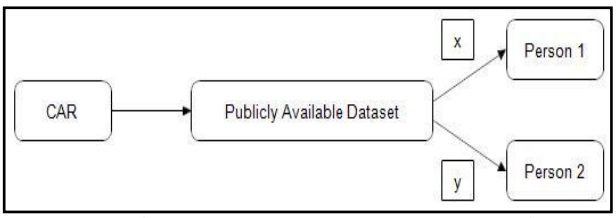

Fig. 1: Before Privacy-Individual Person Can Be Identified.

The CARs produced can be integrated with the publically available information for tracing the sensitive information or personal identification. But, after applying k-anonymization i.e. k=2; the scenario changes as depicted in below figure: 2, wherein the disclosure of the person cannot be done or sensitive information remains hidden. Similarly, according to the requirement, the level of k can be increased for preserving the privacy by anonymization.

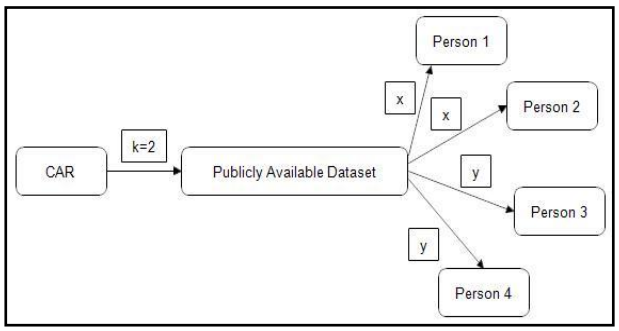

Fig. 2: After Privacy-Individual Person Cannot Be Identified.

4. Optimization

The area of data mining greatly augments and adapts several methods from optimization [13]. The choice of suitable optimization method depends on the type of optimization problem.
The two foremost optimization approaches used are Combinatorial Optimization and Mathematical Programming [13] [15]. The main difference between these two approaches are combinatorial optimization is used for discretized datasets whereas mathematical programming can accept continuous values also.
An example of combinatorial optimization includes genetic algorithm, particle swarm optimization, ant colony optimization etc. Amongst these genetic algorithms has been selected for maximizing the outcome or output due to its ability to find optimal solution in less amount of time to NP-hard problems [20]. Genetic algorithms are search methods based on the principle of natural selection and concept of survival of the fittest [13].
Examples of mathematical programming comprises of neural network, integer programming, linear programming, quadratic programming etc. Amongst these, neural network has been considered due to its characteristics such as high tolerance to noisy and incomplete data. Neural network is a set of inter connected units [10]. Each connection has a weight associated with it and this weight will be updated as per the error found until the terminating condition gets satisfied.
The complete view of the work done in this paper can thus be expressed as given in below figure:

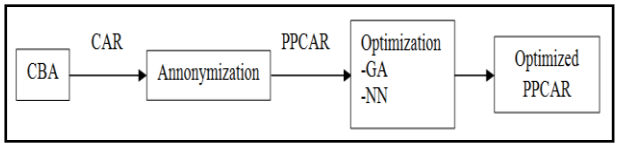

Fig. 3: Optimization of PPCAR.

The genetic algorithm has been proposed to improve the accuracy of PPCARs. Hence, various parameters required to implement the genetic algorithm are fitness function, selection process that comprises of selection of parents and genetic operators namely crossover and mutation, encoding as well as decoding of chromosomes.
Here, the fitness function has been proposed to check the fitness. The Fitness Function is divided into three parts. First part is about "Attribute Coverage", second part denotes "confidence" of each CAR and third part denotes "support" of each CAR.
Suppose S = {C1, C2…Cn}
Each CAR (class association rule) is defined as a combination of antecedents and class label. CAR [QI], QI = Quasi Identifier, Aj=items in CAR

So the fitness function will be, FV (k) = Count (A) in PPCAR + conf (ck) in S+ Sup (ck) in S

Let's take an example to understand the calculation of fitness function. Suppose following are the CARs obtained.
[6 15 23] → 1
[6 15] → 1
[6 23] → 1
[5 16] → 0
[5 16 23] → 0→
Then, to find the fitness value according to the proposed fitness function, following two steps are required to be carried out:

Table 1: Binary Encoding Into Chromosomes (Step1)

Original CARs	Frequent Items 5	6	15	16	23	Class Label	Chromosomes
[6 15 23] →1	0	1	1	0	1	1	011011
[6 15] →1	0	1	1	0	0	1	011001
[6 23] →1	0	1	0	0	1	1	010011
[5 16] →0	1	0	0	1	0	0	100100
[5 16 23] →0	1	0	0	1	1	0	100110

In step 1, original CARs are converted into chromosomes by binary encoding. Here, frequent items from all the CARs are taken into consideration and if particular item is present then it is marked as '1' else '0' as shown in above table 1.

Table 2: Fitness Value Calculations (Step2)

Chromosomes	Count of Antecedents	Confidence of Rule	Support of Rule	Fitness Value
011011	3	0.925	0.27057	4.19557
011001	2	0.91379	0.29068	3.20447
010011	2	0.91371	0.32709	3.24278
100100	2	0.91209	0.30347	3.21556
100110	3	0.90184	0.26874	4.17058

Then, calculation of fitness value is done considering the parameters as specified and the fittest will be taken as parents for the production of generation and the it continues until the termination criteria is satisfied.

5. Implementation

The four real datasets obtainable from UCI data repository are taken into contemplation. The MATLAB tool and Windows operating system are utilized to carry out the implementation. Several preprocessing required such as removal of noisy and missing values, discretization and sampling were carried out on these datasets. The dataset composition [21] is as given in the below table 3:

Table 3: Dataset Information

Datasets	Number of Attributes	Number of Instances	Number of Class
Census Income	11	30719	2
Mammography	6	961	2
Bank Marketing	17	45211	2
Contraceptive Method Choice (CMC)	9	1473	3

Initially, associative classification algorithm namely class based association (CBA) is considered for producing class association rules (CARs). Then, calculation of accuracy and training time by setting the appropriate level of confidence threshold as 0.6 and the support threshold at 0.2 and is done for various dataset
The below given table 4 shows the results of accuracy achieved and execution time required to run each dataset. Here, accuracy has been considered as an evaluation parameter and it is calculated as:
Accuracy (%) = Number of correctly classified instances/Total number of rows * 100

Table 4: Accuracy and Execution Time

Datasets	No. of CARs	Accuracy (%)	Execution Time (sec)
Mammography	8	82.042	156.920
CMC	12	70.842	933.612
Census Income	20	72.081	40449.029
Bank Marketing	34	75.173	75432.983

Then, privacy preserving technique specifically k-anonymization is applied to preserve the privacy of these CARs. Now, after applying k-anonymity on rules, following results have been obtained using MATLAB tool.

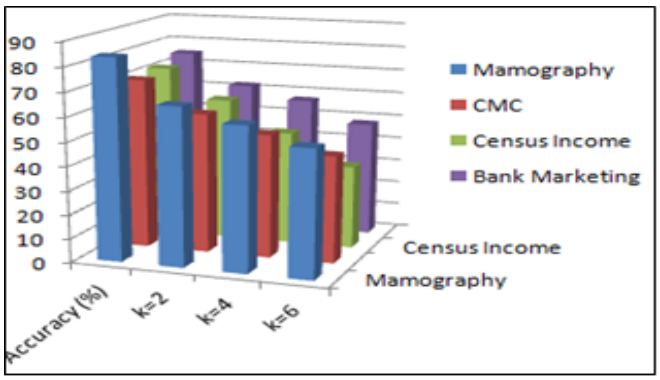

Fig. 4: Accuracy and Privacy.

It can be seen from the above figure 4, that at various privacy levels, accuracy has been inspected considering different database. However, it can be concluded that accuracy decreases with increase in privacy level. i.e. accuracy achieved at k=2 for a particular dataset is more than accuracy achieved at k=6. Because as data is more transformed, privacy is secured better but at the same time due to transformation of data, accuracy is degraded.
Privacy preserved class association rules (PPCARs) are generated using class based association (CBA) technique and anonymization respectively and such rules were optimized by genetic algorithm (GA) considering the accuracy parameter. The result generated regarding their comparisons are as follows:

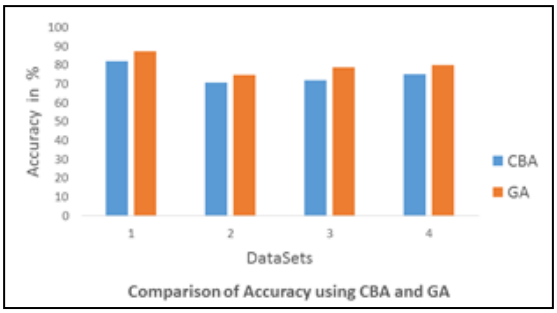

Fig. 5: PPCAR Accuracy (%) by Utilizing CBA and GA.

As shown in above figure 5, accuracy calculated for PPCARs considering all the datasets is found more in GA which performs optimization as compared to CBA method. In above figure Datasets 1: Mammography 2: CMC 3: Census Income 4: Bank marketing is to be taken into consideration.
Further PPCARs were optimized by neural network (NN) also considering the accuracy parameter. The result generated regarding their comparisons are as follows:

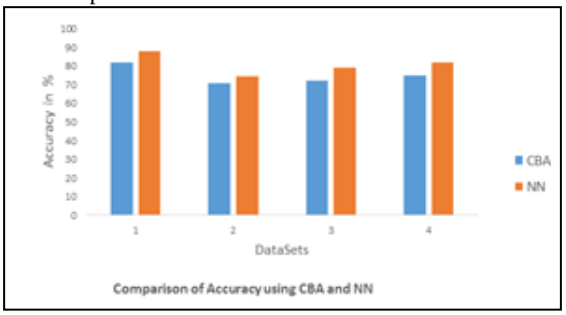

Fig. 6: PPCAR Accuracy (%) by utilizing CBA and NN.

As shown in above figure 6, accuracy calculated for PPCARs considering all the datasets is found more in NN which performs optimization as compared to CBA method. In above figure 6, on the x-axis datasets represents; Dataset 1: Mammography 2: CMC 3: Census Income 4: Bank marketing is to be taken into consideration.
GA and NN techniques were compared considering accuracy parameter and it has been found that both the techniques almost provide the same accuracy which can be clearly concluded from below given table 5:

Table 5: Comparison of Genetic Algorithm and Neural Network

Data set	No. of PPCARs	Accuracy (in %)	
		NN	GA
Mammography	8	75.12	74.56
Contraceptive Method Choice (CMC)	12	64.68	65.56
Census Income	20	66.98	68.79
Bank Marketing	34	71.89	71.23

However, GA and NN techniques differs while considering the time factor into consideration and the graph for training the different datasets are as given below.

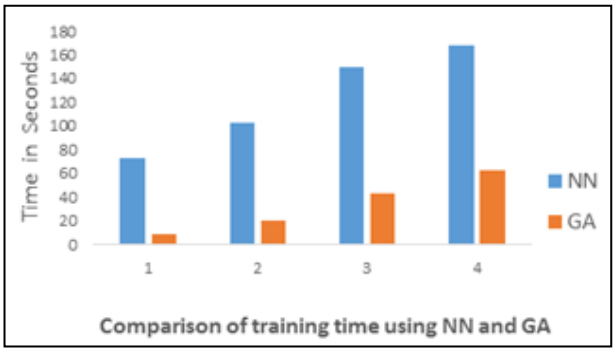

Fig. 7: Time Taken for NN and GA.

It can be concluded that neural network consumes more time to train the network as compared to genetic algorithm. In above figure Datasets 1: Mammography 2: CMC 3: Census Income 4: Bank marketing is to be taken into consideration.

6. Conclusion

Class association rules (CARs) produced by associative classification technique namely class based on association (CBA) method discovers the interesting patterns which can be helpful for making decisions as well as classification purpose. But, an attacker could misuse this information and hence privacy preserved CARs were formed by employing anonymization wherein privacy gets preserved. Certainly accuracy gets degraded due to the data transformation and hence as privacy increases, accuracy decreases. Thus, accuracy is improved by utilizing optimization techniques specifically genetic algorithm (GA) and neural network (NN).
However, GA and NN were compared with CBA wherein it was found that they outperform over CBA considering accuracy parameter. Further, it can also be concluded that GA is better than NN considering time factor whereas the accuracy achieved is almost similar by both the methods.
The work can further be extended by considering the distributed scenario as future scope. Also, other evaluation parameters can be considered for efficiency as well as the side effect of hiding PPCARs can also be explored.

References

[1] H Jiawei, K Micheline, Data Mining: Concepts and Techniques, Morgan Kaufmann Publishers, Elsevier, 2011.
[2] R Liu, H Wang, Privacy-preserving data publishing, Proc. – Int. Conf. Data. Eng., (2010) 305–308.
[3] B Liu, W Hsu, Y Ma, Integrating Classification and Association Rule Mining, KDD98-012pdf 1998.
[4] N Harnsamut, J Natwichai, X Sun, X Li, Privacy preservation for associative classification, Comput. Intell. (2014) 752–770 https://doi.org/10.1111/coin.12028.
[5] S Gokila, P Venkateswari, A Survey on Privacy Preserving Data Publishing, Int. J Cybern. Informatics (2014) 1–8. https://doi.org/10.5121/ijci.2014.3101.
[6] F Thabtah, P Cowling, Y Peng, Multiple labels associative classification, Knowl. Inf. Syst. (2006) 109–129. https://doi.org/10.1007/s10115-005-0213-x.
[7] B Seisungsittisunti, J Natwichai, Incremental privacy preservation for associative classification, Proceeding ACM first Int. Work. Priv. anonymity very large databases - PAVLAD (2009) https://doi.org/10.1145/1651449.1651458.
[8] G Nayak,S Devi, A Survey on Privacy Preserving Data Mining Approaches and Techniques, Int. J Eng. Sci. (2011) 2127–2133.
[9] N Safaei, S Sadjadi, M Babakhani, an efficient genetic algorithm for determining the optimal price discrimination, Appl. Math. Comput. (2006) 1693–1702 https://doi.org/10.1016/j.amc.2006.03.022.
[10] K Park, J Lee, J Choi, Deep Neural Networks for News Recommendations, Proc. 2017 ACM Conf. Inf. Knowl. Manag. - CIKM '17 (2017) 2255–2258 https://doi.org/10.1145/3132847.3133154.
[11] D Patel, R Kotecha, Privacy Preserving Data Mining: A Parametric Analysis, Proceedings of the 5th International Conference on Frontiers in Intelligent Computing: Theory and Applications (2016) 139–149 https://doi.org/10.1007/978-981-10-3156-4_14.
[12] Y Jiang, J Shang, y Liu, Maximizing customer satisfaction through an online recommendation system, A novel associative classification model Decision Support Syst. (2010) 470–479 https://doi.org/10.1016/j.dss.2009.06.006.
[13] D. Martín, J. Alcalá-Fdez, A. Rosete, F. Herrera, "A Niching Genetic Algorithm to mine a diverse set of interesting quantitative association rules", Elsevier, Information Sciences, Volumes 355–356, pp. 208-228, (2016). https://doi.org/10.1016/j.ins.2016.03.039.
[14] J Natwichai, Privacy preservation for associative classification: an approximation algorithm, ECTI-CON2010: The 2010 ECTI International Confernce on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (2011) 127-131.
[15] D Patel, R Kotecha, Associative Classification: A Comprehensive Analysis and Empirical Evaluation, Nirma University International Conference on Engineering (NUiCONE) (2017). https://doi.org/10.1109/NUICONE.2017.8325616.
[16] M Ouda, S Salem, I Ali, E Saad, Privacy-Preserving Data Mining (PPDM) Method for Horizontally Partitioned Data, International Journal of Computer Science (2012) 339–347.
[17] Y Zhu, Y Tang, G Chen, A privacy preserving algorithm for mining distributed association rules, Int. Conf. Comput. Manag. CAMAN (2011) https://doi.org/10.1109/CAMAN.2011.5778775.
[18] S Wedyan, Review and Comparison of Associative Classification Data Mining Approaches, Int. J. Comput. Information Syst. Control Eng. (2014) 34–45.
[19] A Haris, M Abdullah, A Othman, F Rahman, Optimization and data mining for decision Making, World Congr. Comput. Appl. Inf. Syst. WCCAIS (2014) https://doi.org/10.1109/WCCAIS.2014.6916587.
[20] M.Hassoon and M. S.Kouhi and M.Zomorodi-Moghadam and M.Abdar, "Rule Optimization of Boosted C5.0 Classification Using Genetic Algorithm for Liver disease Prediction", International Conference on Computer and Applications (ICCA), pp. 299-305, (2017). https://doi.org/10.1109/COMAPP.2017.8079783.
[21] M Lichman, UCI Machine Learning Repository Irvine CA, University of California School of Information and Computer Science, https://archiveicsuciedu/ml/indexphp (2018).
[22] D. Martín, J. Alcalá-Fdez, A. Rosete, F. Herrera, "A Niching Genetic Algorithm to mine a diverse set of interesting quantitative association rules", Elsevier, Information Sciences, Volumes 355–356, pp. 208-228, (2016). https://doi.org/10.1016/j.ins.2016.03.039.
[23] A.Alexander, C.Stefan, "Respecting Data Privacy in Educational Data Mining: An Approach to the Transparent Handling of Student Data and Dealing with the Resulting Missing Value Problem," IEEE 27th International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE), Paris, pp160-164. (2018).