Accuracy-Privacy Comparison for Enhanced Grouping using a Hybrid Data Mining Technique

Dilbahar Singh[1], Sumit Kumar Yadav[2]

[1]School of Computer Science and Engineering, Lovely Professional University,
Phagwara, Punjab, India

[2]School of Computer Science and Engineering, Lovely Professional University,
Phagwara, Punjab, India

Abstract: *Privacy preserving data mining techniques play very important role in protecting the sensitive and private information of users from distributed data sets. In this modern field of business world applications the capability of storage and record the personal data or information of customers or of any user increase day by day due to this users are worried about the misuse and about the protection of their sensitive information in many data mining techniques that why privacy preserving data mining plays very important role for security and also protect the private data to a greater privacy and proper utilization by making a balance point between privacy and accuracy. In this paper, we are proposing a technique for extended grouping of data sets which provide privacy preserving data mining using ID3 algorithm with k-mean clustering algorithm on randomization response technique. It concludes that the accuracy of the extended group can be increase if we can classify the bounding limit of the data set using Genetic algorithm supported by k- mean. The proposed work is to check out the accuracy level of the dataset if the clusters are divided into inner clusters so that their privacy can be increased.*

Keywords: Data Mining, Privacy Preserving, Accuracy, ID3 algorithm, k-mean clustering algorithm

1. Introduction

Data Mining is a main platform for search and researches of data. By mining of the data we mean to say fetching out a piece of data from a huge data block. The fundamental work in the data mining can further be classified in two following ways one is classification and another is clustering. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. Data mining is mainly used by companies for retail, financial, communication, and marketing organizations. It enables companies to establish relationships among various "internal" factors, i.e., price, product positioning, and "external" factors, i.e., economic indicators, competition, and customer demographic send to determine the its effect on sales, customer satisfaction, and corporate profits. Data Mining, also popularly known as Knowledge Discovery in Databases (KDD), refers to the extraction of hidden, previously unfamiliar and potentially useful information from databases.

Why we need privacy preserving, there are the various reasons for this question; in daily life we face this situation often when we have to submit our details in some outlet or somewhere which requires our information. This information is considered as customer information. The customer information may consist of a lot of personal information which a customer would not like to be published in general. But this protection of the published information leads to a less accuracy of the data. Now by means of less accuracy, we mean to say that if you will not be allowed to go through the data content, that particular data content won't be included into the searching. In this modern field of business world applications the capability of storage and record the personal data or information of customers or of any user increase day by day due to this users are worried about the misuse and about the protection of their sensitive information in many data mining techniques. Hence we can

say that for avoid the misuse and disclose of the personal and sensitive information in front of others that's why privacy preserving keep in mind.

And if two or more parties want to share their databases in business world problems and in secure multi party computation then they also have to maintain privacy because one party can have their some personal and sensitive information of customers and of their users that's why they do not want to disclose their information to other party hence, in this situation also need privacy preserving against data mining attackers. So, we can say that privacy preserving play very important role for protection of the personal information against the false detector or attacker to the data sets. In today real world business application areas data collection and transactional record is stored somewhere to make the databases with this huge collection information contain personal information and records for the analysis purpose and proper classification by maintain the balance between privacy and accuracy that's why privacy preserving data mining come into consideration. But here in our work, we are using ID3 (Iterative Dichotomiser3) and Genetic algorithm followed by k-mean clustering algorithm for building decision trees classifier and cluster classification of data sets and also a randomization method is using as privacy preserving data mining technique for protecting the personal and sensitive data with a better accuracy by extended data into groups. So here we are going to take a small look on the algorithms which we are using in our work. ID3 is a decision tree building algorithm which use for classification of data sets by testing the values of their attributes. It helps in building the decision tree from a set of objects or data and their attribute values in a top down fashion. At every node of the tree test the attribute value which helps in partition the data sets. The testing of attribute values helps in classification of data sets into sub sets which depend on maximum information gain and minimum entropy for making the decision nodes of the tree classifier, gain will

be calculated for every attribute and attribute with the higher information gain plays very important role to split the data set S into subsets t which help in making the decision tree node containing that attribute. This process will be applicable on every subset until all the elements or objects of that subset belong to the same class means this node will become the leaf node or terminal node. When all the subsets will classify properly then ID3 algorithm is terminated.

The role of Genetic algorithm play very important role in our work used for calculating the threshold value of every elementary data sets which is represented by theta value. Basically, it is used to calculate the useful solutions for any search problems and provides accurate results because in calculation of the solution, it generates optimization which satisfies the minimum criteria for the analysis purpose.

K-mean algorithm helps in classify or clustering (grouping) the data objects on the basis of their attribute or properties values into k number of clusters. It is a simple iterative method to partition a given dataset into a user specified number of clusters. We are using k-means in our work because it helps in calculating the accurate classification and data mining analysis from the data sets. Specifically, k-mean has been used here because a very difficult structure is drawn within a single data set. It means that there is a non hierarchal architecture within a specified cluster. Any other algorithm might have not suitably concluded the values of data sets for proper classification taken for consideration. And it computes the accuracy level of all the groups for each and every data set taken for analysis by the help of clustering technique for average computation at each node or group.

2. Literature Survey

Data mining is that technique which provides knowledge discovery information from large databases for accurate decision making purpose and extracts useful information according to their need. And also use for privacy preserving in various applications for protecting the sensitive data of customers to better and accurate utility of data. In this modern field of business world applications the capability of storage and record the personal data or information of customers or of any user increase day by day due to this users are worried about the misuse and about the protection of their sensitive information in many data mining techniques that's why privacy preserving data mining plays very important role for security and also protect the private data to a greater privacy by making a balance point between privacy and accuracy. Here, we are introducing some author's views about privacy preserving data mining and techniques related to privacy and accuracy of data analysis.

(Zhijun Zhan, 2003), they developed a randomized response technique by which they collect the data to form a database or datasets for data mining purpose which holds privacy to a great extent. Means they preserving the private data and here in data mining process, they developed a method by which no actual and private information of any user or customer found easily but on randomly selection of data from datasets provides better results in data mining process and also enhance the privacy level. In their paper, they are taking the

multi group scheme in which they are dividing the user's information's and customer feedbacks into multi groups and apply randomized technique by which they represent a decision tree classifier concept for clustering the data from the datasets. Here, they are taking the three group scheme in which the datasets are disguised into three different groups according to their attribute values independently and using ID3 algorithm for decision tree classifier and provides the information about the data groups. They also told that if the number of groups are increasing then the privacy level also increases and accuracy level decrease in mining the information form data groups and all the things in decision tree building depends on information gain and entropy form ID3 algorithm [1].

(Zhouxuan Teng, 2007), they mentioned a hybrid multi-group privacy preserving approach for building decision trees on vertically partitioned data. And this hybrid approach is the combination of the randomization and secures multi-party computation (SMC) techniques which map the efficiency and accuracy constraints. Here they divide the data attributes into the groups for group-based randomization by which data miner can extract useful and accurate information making the balance between privacy and accuracy. This multi-group scheme is implemented by them using ID3 decision tree algorithm and form their results conclude that this scheme provides better accuracy as compare to the randomization and take less computation time as compare to the SMC. But here they use the fixed window size and then suggest that performance can be increase by using the dynamic window size and it can be different for different types of tree nodes depending on the randomization results [2].

(Zhengli Huang, 2008), they introduce an optimizing randomized response schemes for privacy preserving data mining. In this technique, the sensitive or private information disguised. And told that previously, there is no any optimal randomized response schemes have been proposed. So, here they introduce a technique or method for better utility and privacy quantification as estimate problems obtain from estimate theories and to obtain optimal disguise matrices for randomized response techniques using multi-objective optimization method. And they show that their scheme have much better performance as compare to other randomized response techniques [3].

(Chen Jin, 2009), in their paper they describe that the decision tree is one of the best methods for classification and prediction in data mining based on ID3 algorithm. Here they discussed the new decision tree algorithm which is the combination of the ID3 and Association Function. In which they calculate the relation degree function values of the attributes of the data sets but doing this the computational complexities increased [4].

(Ning Zhang, 2011), used a differential privacy term for distributed data mining which provided the better security measure for the privacy protection of the information stored in different data places against the data miner attackers. They proposed an algorithm which combines techniques from secure multiparty computation and differential privacy concept named as secure group differential private query

(SDQ). They provided a system model which consists of n different number of databases and each database has an individual owner of its data. The data miner used a decision tree induction for issued the count queries. Here they used a problem of distributed knowledge extraction while preserving the privacy due to this select the query capabilities randomly for the different databases which significantly reduced the probability of collision attacks. Therefore, their results show that using the decision tree induction with SDQ can achieve better privacy [5].

(MohammadReza Keyvanpour, 2011), they also told about the privacy of the data or private and sensitive information of the users and in databases. Here they provides the Classification and Evaluation the Privacy Preserving Data Mining Techniques by Using a Data Modification–based Framework in which they are concerning about the protection privacy of the increasing data day by day which contains critical and sensitive information means private information. They provide a technique by divides into two approaches one is perturbation approach and another is Anonymization approach. Privacy preserving data mining techniques are used by data modification framework which presents the classification and evaluation of the data. These two approaches are represented for providing the protection against the different attacks against the privacy by calculating a critical value and also compare the privacy and accuracy balance respectively but not provide the better results in privacy and data mined accuracy that is not the better and suitable approach [6].

(Xingwen Liu, 2011), they provided a Novel Method for Inducing ID3 Decision Trees Based on Variable Precision Rough Set. In which classification play very important role in many search algorithm of data mining and classification by the decision tree building algorithm provides a better and a clear view of it for the classification of the data sets in decision taking tasks and ID3 is one of the algorithm which is used for decision tree building algorithm and provides the best way and path for the decision tree classifier. Decision tree building depends on the information gain calculated from the ID3 algorithm after applied to the data sets and attributes of the data because this value is used for testing the attributes in order to make the decision tree nodes basically in ID3 algorithm the information gain value must be maximum and entropy must be minimum for better classification of data sets. On the basis of the gain value we split the nodes in the decision tree that is root node and sub nodes. But an attribute can have more values which not able to provide the better results in node splitting hence not a best attribute for decision making. Here they provided an improved information gain for the selection of the optimal splitting attribute for the decision attribute which is based on the dependency degree of condition attributes which is better than the previous one. Here they used a rough set approach for ID3 algorithm for generating the best decision attribute having the optimal splitting ability. Hence, we can say that they provide a improved solution (VPRSID3) based on dependency of attributes in variable precision rough set theory (VPRS) to select the splitting attributes. They also provided the relationship between the condition and decision attributes using information gain by the help of dependency among the attributes for accurate classification [7].

(Imas Sukaesih Sitanggang, 2011), they suggested that the classification of the non spatial data is more easy as compare to the spatial data that why they described an extended ID3 decision tree algorithm for the classification of the spatial data. Because spatial data mining algorithm also take interest to the neighbors objects in order to extract useful and meaningful classification. Hence they described a new spatial decision tree algorithm which is based on ID3 algorithm for discrete features represented in points, lines and polygons. Here the selection of the attribute or best decision attribute is based on the spatial information gain for best splitting nodes of the tree which will be calculated from spatial measure formula. And they provide the result by applying the algorithm on two different spatial data objects and construct the spatial decision tree from the small spatial datasets containing the discrete features like points, lines and polygons [8].

(V. Narmada, 2011), introduced an enhanced security algorithm in privacy preserving for distributed data bases. In this modern field of business world applications the capability of storage and record the personal data or information of customers or of any user increase day by day due to this users are worried about the misuse and about the protection of their sensitive information in many data mining techniques that why privacy preserving data mining plays very important role for security and also protect the private data to a greater privacy by making a balance point between privacy and accuracy. In this paper, they define the various privacy preserving data mining techniques and also provide and suggest a novel algorithm that is CryptDB algorithm for security. Here they define randomization technique and k-anonymization technique and partition the distributed database into the horizontally and vertically data sets and provides the security by the different layers of encryption and this encrypted data is processed by the help of SQL key queries in which these keys provided to the client with minimum decryption capabilities. So, here privacy will increase but accuracy decrease [9].

(Nishant Mathur, 2012), they also introduced the ID3 algorithm for decision tree building but here they are using the concept of Havrda and Charvat Entropy in place of Shannon Entropy by which they provides a decision tree from the calculated information as the root and sub roots of the decision tree. Here decision tree are used for better analysis and for better decision about the data. And decision tree algorithm with ID3 algorithm provides the best path for better analysis about the data division. They suggest that the number of nodes is small in the decision tree and it is less complex on the basis of the lesser value of the alpha that is less than one [10].

(Hem Jyotsana Parashar, 2012), in their paper they are also told about the decision tree building by using ID3 algorithm for better and efficient classification of the data sets by removing the limitations of ID3 algorithm by doing some modification in it and proposed a new or improved algorithm. In which the attributes are divided into the groups and applied the improved algorithm for determining the information gain valve for the proper classification of the data sets or attributes. And if the information gain is not

good then, further divide the attributes valve for good information gain and better classification which classified the datasets more accurately and efficiently. The improved algorithm is a recursive algorithm means recursively performed for getting the high and 100% classification results and the provides the simpler form of decision tree with minimum depth for a particular data set [11].

(Majid Bashir Malik, 2012), suggested that in data mining techniques and process privacy preserving play very important role means not disclosed the sensitive and private information of any user means protect the privacy. Because users are well aware about their private data and not want to share their information that is they worried about it. That's why privacy preserving data mining is important here. There are the various factors which are responsible for the better privacy preserving data mining these are its performance, utilization of data and the resistance and uncertainty levels. They provides the privacy preserving data mining (PPDM) techniques and told about the approaches and dimensions on the basis of which (PPDM) can be classified. These are the methodologies for protecting the private and sensitive data and data mining process and the dimensions are data distribution, data modification, and data mining algorithms, data or rule hiding and privacy preservation. And on the basses of these dimensions the PPDM techniques can be classified into the following parts these are Anonymization based PPDM, Perturbation based PPDM, Randomized Response based PPDM, Condensation approach based PPDM and Cryptography based PPDM. At last they conclude that no anyone single privacy preserving data mining algorithm which fulfill all the requirements and criteria like performance, utility, cost, complexity and tolerance etc. but one algorithm may perform better than other in different- different criteria and conditions [12].

(Savita Lohiya, 2012), they provide a hybrid approach for privacy persevering in data mining. They told about the when two party or originations sharing their data in many applications and in business planning and in other forms. In this shared environment of data there may be the some private and sensitive data which must not disclosed to any other parties means there must be the some privacy to some data in this shared environment. In their paper they are taking scenario of Health Insurance Portability and Accountability Act (HIPAA), in which they are protecting the patient privacy and medical information or data that why they are providing a hybrid approach for privacy preserving in which protecting the private data for better accuracy by using randomization on original data followed by the generalization on randomized or modified data. They are describing the various techniques for privacy preserving like as method of k-anonymity, random perturbation, blocking based method, cryptographic techniques and condensation approach and define the hybrid approach algorithm by taking an example for privacy preserving. And they conclude that these various techniques have limitations like anonymity has limitations of homogeneity and background attacks but it provide privacy and usability of data, random perturbation technique not usefully provide usability of data, blocking method lacks in providing full information means information loss is there, cryptographic techniques having more computational time and not provide the usability of

data but provides security, not proper information found in condensation and randomized techniques but they preserve privacy. But in their hybrid algorithm they used k-anonymity and randomization techniques in which because of randomization, attacker not able to identify the data patterns and anonymity has limitations of homogeneity and background attacks but in combine effect attacker not able to identify the homogeneity and background attacks [13].

(Hiroaki Kikuchi, 2012), they told about a randomized perturbation scheme for privacy preserving of data where a server which allows the users to predict the rating value from the disguised data and in original rating value user can add a random noise for providing higher privacy. In their scheme they use a posterior probability distribution function which inherit form Bayes' estimation for the reconstruction of the original distribution to check the similarity between the computed data from the disguised data matrix. And some incensement is shown here in their work [14].

(Thanveer Jahan, 2012), according to this paper privacy preserving data mining plays very important role in many applications for security of private or sensitive information. Here they provide a data distortion method in which they used two techniques one is the singular value decomposition (SVD) and another is sparsified singular value decomposition (SSVD) with feature selection scheme for the reduction of feature space. So original data sets are converted into the distorted data sets by the help of (SSVD), and then feature selection apply on the distorted data sets which discard the low value perturbed data. They are using real world data sets and sparsified singular value decomposition (SSVD) with feature selection in their experiment for providing better accuracy and also used data mining classifiers like SVM, ID3 and C4.5 for extracting accurate information or for testing the mining utility [15].

3. Problem Formulation

In daily life we face this situation often when we have to submit our details in some outlet or somewhere which requires our information. This information is considered as customer information. The customer information may consist of a lot of personal information which a customer would not like to be published in general. But this protection of the published information leads to a less accuracy of the data. Now by means of less accuracy, we mean to say that if you will not be allowed to go through the data content, that particular data content won't be included into the searching. Now in the work till now, ID3 algorithm have been implemented to check out the accuracy and there was only a single predefined threshold value for all the elements of one data groups due to which the accuracy of the data decreases as the threshold value pick only those data values whose defining values are more than threshold values. According to the current implemented situation, we have found that the accuracy decreases if the numbers of groups are three and privacy increases at the same time. In the previous work, ID3 algorithm is implemented up to the three data groups for privacy preserving data mining using randomized response techniques in which their results show that when number of groups increase the privacy level also increase but accuracy level decrease means, not provide feasible solution and

accurate utilization of data. Hence due to this problem, decrease in accuracy we are proposing a hybrid data mining approach for enhanced data grouping.

Now our problem is to create a new group to check out whether the implemented algorithms can make any change in the accuracy if we increase the privacy level or not. For this purpose, we are proposing privacy preserving data mining technique using ID3 (to design the decision tree) and Genetic algorithm (to define threshold value of elementary data sets) followed by k-mean clustering algorithm for better result analysis (classification) of data sets into different data groups which may enhance the accuracy level in extended grouping of data sets.

4. Proposed Method

Here, we are adding a fourth group and applying ID3 and Genetic algorithm followed by k-mean clustering algorithm for checking whether the accuracy will increase or not. Now for this thing to be implemented, first of all we need random attributes. Now we need to perform the randomization procedure on every dataset which we have considered. The proposed work is to check out the accuracy level of the dataset if the clusters are divided into inner clusters so that their privacy can be increased. For this purpose we have taken the data into binary format which has been converted as per the ASCII values of the data.

In the proposed work, we are dividing the current data set taken for the analysis into one extended group. All the previous work done till now has taken the ID3 algorithm for the basic classification on the basis of which they define a particular threshold value which provides the input output accuracy vintage point. But the problem is that the threshold value is same for each and every data. When, we have done the classification on the basis of two additional algorithms known as Genetic and k-mean algorithm. At the first step, we are applying the ID3 algorithm to the basic data sets for the testing as per the previous algorithm and randomized the data sets elements into groups by the help of some parameters obtained from ID3 algorithm.

Now we need to define the threshold value according to each elementary data as per their specification. The inner data elements of a data set are again clustered according to their elementary properties so that the Genetic algorithm can draw a threshold for each subsequent group. Now, when the classification is done we assumed this threshold value as theta. As the theta value is different for every elementary group the classification for the accuracy becomes much simpler as the group is only classified according to its own threshold value not by the overall threshold value. While classifying, the mean value defines the average of the true positive picked up at different instances (values) of theta. In the similar manner, the difference between two mean values is defined as variance. According to our algorithmic assumption the variance should be small and accordingly the probability to finding true positive increasing.

And at the final stage, we will compute the different accuracy level formed by variance at different theta value with the help of k-mean clustering algorithm. The k-mean algorithm will compute the accuracy level of all the four groups for each and every data set taken for the analysis. By the end of the statement, it picks up a clustering technique for average computation at each node or group. The decision tree will represent the accuracy increase in each group represented by node point value of the decision tree.

5. Conclusion

Conclusion is that when we are increasing the randomizing the privacy level is increasing but the accuracy decreases as the values of the randomized data does not match with original data set. In this paper, we are proposing a hybrid data mining technique for enhanced accurate grouping in which the accuracy is increasing. Because in the previous work done have only used randomized response technique and ID3 algorithm and having a predefined threshold which is not efficient for the accurate classification of data sets. That's why; we are implementing the ID3 algorithm along with the Genetic algorithm which decides at the runtime that what would be the threshold value for the cutoff and also calculate the threshold for every elementary data sets followed by k-mean algorithm because of a very sophisticated architecture is drawn within a single data set. Due to this, at different states of the value, threshold value is different and hence it is easy to compute the rational values of the computing environment so that it should give us an increased accuracy.

6. Future Scope

If final accuracy can be computed with the help of neural classifier then it would be an advantage in the accuracy gain. The researcher will have to keep it in mind; the neural classification cannot be done till the data sets are not categories appropriately.

References

[1] Zhijun Zhan, Wenliang Du, "Privacy-Preserving Data Mining Using Multi-Group Randomized Response Techniques," 2003.
[2] Zhouxuan Teng, Wenliang Du, "A Hybrid Multi-Group Privacy-Preserving Approach for Building Decision Trees," United States National Science Foundation, 2007.
[3] Zhengli Huang, Wenliang Du, "Optimizing Randomized Response Schemes for Privacy-Preserving Data Mining," IEEE, ICDE 2008.
[4] Chen Jin, Luo De-lin,Mu Fen-xiang, "An Improved ID3 Decision Tree Algorithm," In Proceedings of the IEEE 4th International Conference on Computer Science & Education, pp. 127-130, 2009.
[5] Ning Zhang, Ming Li,Wenjing Lou, "Distributed Data Mining with Differential Privacy," In Proceedings of the IEEE ICC, 2011.
[6] MohammadReza Keyvanpour, Somayyeh Seifi Moradi, "Classification and Evaluation the Privacy Preserving Data Mining Techniques by using a Data Modification–based Framework," International Journal on Computer Science and Engineering (IJCSE), pp. 862-870, 2011.

[7] Xingwen Liu*, Dianhong Wang, Liangxiao Jiang, Fenxiong Chen and Shengfeng Gan, "A Novel Method for Inducing ID3 Decision Trees Based on Variable Precision Rough Set," In Proceedings of the IEEE Seventh International Conference on Natural Computation, pp. 494-497, 2011.

[8] Imas Sukaesih Sitanggang, Razali Yaakob, Norwati Mustapha, Ahmad Ainuddin B Nuruddin, "An Extended ID3 Decision Tree Algorithm for Spatial Data," IEEE, 2011.

[9] V.Narmada, B. Narasimha Swamy, D. Lokesh Sai Kumar, "An enhanced security algorithm for distributed databases in privacy preserving data bases," International Journal of Advanced Engineering Sciences and Technologies (IJAEST), pp. 219-225, 2011.

[10] Nishant Mathur, Sumit Kumar, Santosh Kumar, and Rajni Jindal, "The Base Strategy for ID3 Algorithm of Data Mining Using Havrda and Charvat Entropy Based on DecisionTree," International Journal of Information and Electronics Engineering, pp. 253-258, 2012.

[11] Hem Jyotsana Parashar, Singh Vijendra, and Nisha Vasudeva, "An Efficient Classification Approach for Data Mining," International Journal of Machine Learning and Computing, pp. 446-448, 2012.

[12] Majid Bashir Malik, M. Asger Ghazi, Rashid Ali, "Privacy Preserving Data Mining Techniques: Current Scenario and Future Prospects," In Proceedings of the IEEE Third International Conference on Computer and Communication Technology, pp. 26-32, 2012.

[13] Savita Lohiya, Lata Ragha, "Privacy Preserving in Data Mining Using Hybrid Approach," In Proceedings of the IEEE Fourth International Conference on Computational Intelligence and Communication Networks, pp. 743-746, 2012.

[14] Hiroaki Kikuchi, Anna Mochizuki, "Privacy-Preserving Collaborative Filtering using Randomized Response," In Proceedings of the IEEE Sixth International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing, pp. 671-676, 2012.

[15] Thanveer Jahan, Dr. G. Narsihma, Dr. C. V Guru Rao, "Data Perturbation and Feature Selection in Preserving Privacy," IEEE, 2012.

Author Profile

Dilbahar Singh received the B.Tech degree in Computer Science and Engineering from CCS University Meerut in 2011 and pursuing M.Tech in Computer Science Engineering from Lovely Professional University Phagwara, Punjab.

Sumit Kumar Yadav received the B.Tech degree in Computer Science Engineering from UPTU in 2008 and M.S degree in Information Security from IIIT, Allahabad in 2010. Currently, he is working as the Assistant Professor of Computer Science and Engineering in Lovely Professional University Phagwara, Punjab.