Privacy Preservation in Data Mining

Gauri Deo[1], Dr. Prof. Girish Kulkarni.[2]

[1] *PG Student, dept. of Communication Engineering, S.S.G.B.C.O.E.T., Maharashtra, India.*
[2] *Assistant professor, dept. of E & TC, S.S.G.B.C.O.E.T, Maharashtra, India.*

ABSTRACT

The growing popularity and development of data mining technologies bring serious threat to the security of individual's sensitive information. An emerging research topic in data mining, known as privacy preserving data mining (PPDM), has been extensively studied in recent years. The basic idea of PPDM is to modify the data in such a way so as to perform data mining algorithms effectively without compromising the security of sensitive information contained in the data. In this paper we discussed the k nearest neighbor algorithm which is used for privacy preserving data mining.

Keyword: - *Data Mining, PPDM, Privacy Preservation, k nearest neighbor.*

1. Introduction

In modern days organizations are extremely dependent on Data Mining results to provide better service, achieving greater profit, and better decision-making. For these purposes organizations collect huge amount of data. Data Mining deals with automatic extraction of previously unknown patterns from large amounts of data sets. This data includes sensitive data about Individuals or organizations. While running Data Mining algorithm against such data, the algorithm not only extracts the knowledge but it also reveals the information which is considered to be private. The real threat is that once information gets exposed to unauthorized party, it will be impractical to stop misuse. Privacy can for instance be threatened when Data Mining techniques uses the identifiers which themselves are not very sensitive, but are used to connect personal identifiers such as addresses, names etc., with other more sensitive personal information. Privacy is very important for trusted collaboration and interactions. Because of these privacy and data security concerns in data mining, the data owner hesitates while sharing data for data mining activities. And this creates barrier in data mining task. Privacy preserving data mining technique gives new direction to solve this problem [1].

Solution to this problem is provided by Privacy preserving in data mining (PPDM). PPDM is a specialized set of Data Mining activities where techniques are evolved to protect privacy of the data, so that the knowledge discovery process can be carried out without barrier. The objective of PPDM is to protect sensitive information from leaking in the mining process along with accurate Data Mining results. In PPDM, new techniques are invented to provide privacy for the knowledge discovered in Data Mining. It also takes care that knowledge discovery process should not be banned because of privacy reason [1].

2. Literature Review

In this paper [2], discussed an innovative protocol. This protocol used both actual and idyllic model. By using both the models, were providing more security and privacy. The data blocks were broke into segments and redistribute the segments among all the parties. The key idea was that, whatever computed by a party participating in the protocol, computation based on its input and output only.

In this paper [3], presented a sanitization algorithm with the consideration of four side effects based on multi-objective PSO and hierarchical clustering methods to find optimized solutions for PPDM. Experiments

showed that compared to existing approaches, the designed sanitization algorithm based on the hierarchical clustering method achieves satisfactory performance in terms of hiding failure, missing cost, and artificial cost.

This paper [4], addressed some of the basic techniques: randomization, k-anonymity, distributed privacy preserving and application effectiveness downgrading. This paper illustrated the application of certain techniques for preserving privacy on experimental dataset, and reveals the effects that their use has on the results.

In this paper [5], dissected the characteristics of emotional intelligence data, analysed 804 samples of 5 Chongqing vocational Institute using K-means cluster analysis method.

This paper [6], stated that the vertically partitioned data available with the parties involved cannot provide accurate mining results when compared to the collaborative mining results. To overcome the privacy issue in data disclosure this paper described a Key Distribution-Less Privacy Preserving Data Mining (KDLPPDM) system in which the publication of local association rules generated by the parties was published.

In this paper [7], the most relevant PPDM techniques surveyed from the literature and the metrics used to evaluate such techniques and presents typical applications of PPDM methods in relevant fields. Furthermore, the current challenges and open issues in PPDM are discussed.

This paper [8], proposed a method called Hybrid approach for privacy preserving. First randomizing the original data. Then by applying generalization on randomized or modified data. This technique protected private data with better accuracy, also it can reconstruct original data and provide data with no information loss, makes usability of data.

In this paper [9], stated a survey for the models and techniques that were used for publishing data about patients. The results showed that the algorithm is scalable and was doing better than existing methods.

In this paper [10], stated that the efficient privacy preserving decision tree constructed by reducing communication and computation cost while performing secure cardinality of scalar product during tree induction. The algorithm scaled well with more number of parties.

In this paper [11], provided user anonymous access to cloud services and shared storage servers. This means that users' personal attributes (age, valid registration, successful payment) can be proven without revealing users' identity. Thus, users can use services without any threat of profiling their behaviour. On the other hand, if users break provider's rules, their access rights are revoked.

In this paper [12], stated an approach which accepts the user queries consisting of different attributes and identifies the sensitive attributes whose values are to be scrambled depending on threshold value. The threshold value was calculated depending on the different weights assigned to individual attributes. The information under those particular attributes whose total weights exceeds the threshold values is scrambled.

In this paper [13], a set of fuzzy based mapping techniques was compared in terms of their privacy preserving property and their ability to retain the same relationship with other fields.

In this paper [14], stated that the k-anonymity is a significant method for protecting privacy in micro-data release or publishing. k-anonymity protect micro-data table released be indistinguishably related to no fewer than k respondents. Partition in k-anonymity are single dimensional. This paper proposed a new multidimensional model, which provides better k-anonymity. A multidimensional k-anonymity with nearest neighbourhood strategy introduced and experimental results show that it performs better ink-anonymity

In this paper [15], proposed an improved algorithm based on the FDM algorithm. In the process, it computes the total support count with the privacy-preserved method, meanwhile ensures the source of every local large item-set and local support count is covered, so it reduced the time spent on communication and preserved the privacy of the data distributed at each site. The experimental evaluations show that the proposed algorithm was efficient and rather suitable for the practical application fields.

This paper [16], specifically aimed to solve the problem of secured outsourcing of frequent item-set mining on multi-cloud environments. For satisfying the k-support anonymity, adopted the taxonomy-based anonymization technique to build a taxonomy tree with the items of complete support and include the items of partial support as noise.

This paper [17], provided a comprehensive review of the current state-of-the-art PPUM algorithms, including the preliminaries of utility mining and PPUM, evaluation criteria for PPUM, details of existing PPUM algorithms (e.g., techniques, advantages and disadvantages). Also highlighted some important open challenges and opportunities of this topic that need to be further developed in the future.

3. Research Gap Identified

In this paper [2], the protocol is different from the existing hybrid secure sum protocol because of redistribution of data blocks. Thus, there is no chance of attack because of this. This protocol provides a zero leakage chance, but the complexity of the protocol is very high.

In this paper [3], the results showed that the designed algorithm has achieved satisfied performance especially when compared to other approaches with respect to hiding failure. Another approach can combined to obtain better results in the future work.

In this paper [4], said that although the filters in Weka tool do not provide an absolute privacy protection, users can benefit from their correct and combined utilization. To determine real benefit utilization of different techniques for privacy preserving brings, it is necessary to measure the achieved degree of privacy preserved in data and the amount of data and information that is lost in this process.

In this paper [6], stated that the acceptance of the proposed KDLPPDM system heavily relies on the accuracy of the data mining algorithm adopted in it. In this paper, the KDLPPDM system is developed with two decision tree algorithms namely the C4.5 and C5.0. The data mining efficiency of classification is dependent on the rules generated using both the data mining algorithm and is evaluated on the basis of the Receiver Operating Characteristics (ROC).

In this paper [8], stated that anonymity technique gave privacy protection and usability of data but it had drawback of homogeneity and background attack. Random perturbation technique did not provided usability of data. Blocking method gave information loss. Cryptography technique gives privacy protection but did not provided usability of data and it required more computational overhead. Condensation and randomized response technique preserve privacy but they gave information loss.

In this paper [9], the drawback of the proposed algorithm was that the production of artificial data for wide range health database became inadequate when the domain size was huge.

In this paper [11], the verification process was directly related to the blacklist, which contain revoke value. So as the blacklist contents increases verification process time also increases.

In this paper [12], the identification of sensitive attribute was based on threshold limit of sensitivity which was related to each attribute property. To preserve the privacy of sensitive information the data owner modifies the value under identified sensitive attributes using swapping technique. If attribute did not achieve the certain threshold then it will not be changed and information might reveal.

In this paper [13], two types of tests were conducted to evaluate the effect of mapping on data mining. 1] Similarity measures 2] Effect on derived association rules. Similarity measures compared with other similarity measures - Jaccard similarity (J), Dice similarity (D), and Cosine similarity (C) all the proposed similarity measures show results similar to the correlation coefficient measure

In this paper [14], tried to find out which algorithms were better in divide and conquer and nearest neighborhood techniques with the help of different measuring attribute. The nearest neighborhood techniques gave better k-anonymity as compare to divide and conquer techniques. The selection for quasi-identifiers in this paper has cause of an information loss by anonymization.

In this paper [17], stated the problem of secured outsourcing of frequent item-set mining on the multi-cloud environments. Concerning the challenges in big data analysis, suggested to partition the data into several parts, and outsourced each part independently to different cloud.

In this paper [18], discussed about four different types of users involved in data mining applications, namely, data provider, data collector, data miner, and decision maker. For each type of user, discussed their privacy concerns and the methods that can be adopted to protect sensitive information and briefly introduced the basics of related research topics, review state-of-the-art approaches.

4. K nearest neighbor algorithm

The k-nearest neighbors (KNN) algorithm is a simple, easy-to-implement supervised machine learning algorithm that can be used to solve both classification and regression problems. K Nearest Neighbor is a simple algorithm that stores all the available cases and classifies the new data or case based on a similarity measure. It is mostly used to classify a data point based on how its neighbors are classified. K-nearest neighbors (KNN) algorithm is a type of supervised ML algorithm which can be used for both classification as well as regression predictive problems. However, it is mainly used for classification predictive problems in industry. The following two properties would define KNN well –

- Lazy learning algorithm − KNN is a lazy learning algorithm because it does not have a specialized training phase and uses all the data for training while classification.

- Non-parametric learning algorithm − KNN is also a non-parametric learning algorithm because it doesn't assume anything about the underlying data.

A supervised machine learning algorithm (as opposed to an unsupervised machine learning algorithm) is one that relies on labelled input data to learn a function that produces an appropriate output when given new unlabeled data.

The KNN algorithm assumes that similar things exist in close proximity. In other words, similar things are near to each other. 'K' in KNN is a parameter that refers to the number of nearest neighbors to include in the majority of the voting process. 'k' in KNN algorithm is based on feature similarity choosing the right value of K is a process called parameter tuning and is important for better accuracy. Finding the value of k is not easy.

For defining value of k-

1. There is no structured method to find the best value for "K". We need to find out with various values by trial and error and assuming that training data is unknown.
2. Choosing smaller values for K can be noisy and will have a higher influence on the result.
3. Larger values of K will have smoother decision boundaries which mean lower variance but increased bias. Also, computationally expensive.
4. Another way to choose K is though cross-validation. One way to select the cross-validation dataset from the training dataset. Take the small portion from the training dataset and call it a validation dataset, and then use the same to evaluate different possible values of K. This way we are going to predict the label for every instance in the validation set using with K equals to 1, K equals to 2, K equals to 3.. and then we look at what value of K gives us the best performance on the validation set and then we can take that value and use that as the final setting of our algorithm so we are minimizing the validation error .
5. In general, practice, choosing the value of k is k = sqrt(N) where N stands for the number of samples in your training dataset.
6. Try and keep the value of k odd in order to avoid confusion between two classes of data

The following are the different boundaries separating the two classes with different values of K.

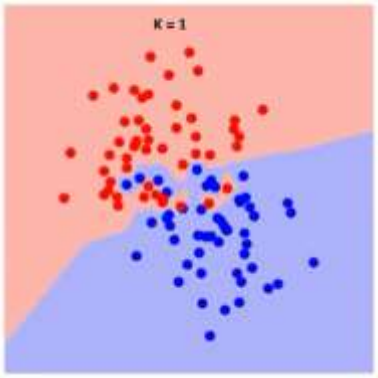

Fig-1

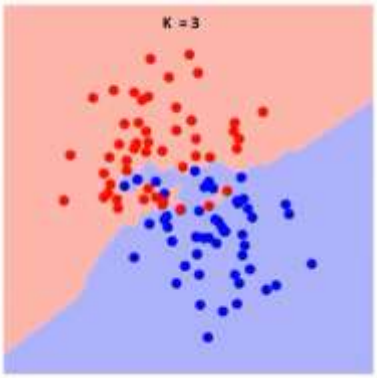

Fig-2

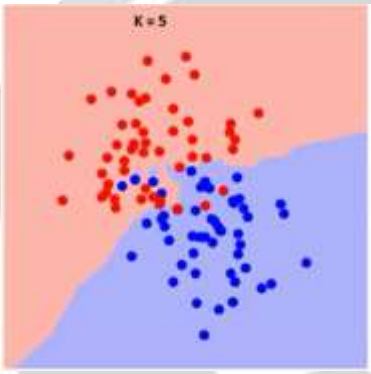

Fig-3

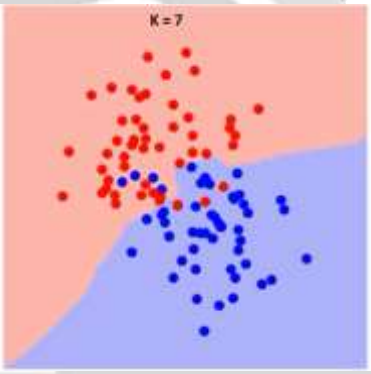

Fig-4

If you watch carefully, you can see that the boundary becomes smoother with increasing value of K. With K increasing to infinity it finally becomes all blue or all red depending on the total majority. The training error rate and the validation error rate are two parameters we need to access different K-value. Following is the curve for the training error rate with a varying value of K:

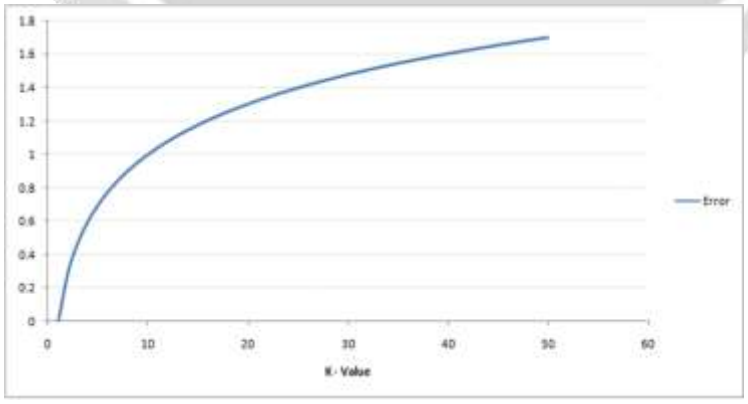

Chart-1

As you can see, the error rate at K=1 is always zero for the training sample. This is because the closest point to any training data point is itself. Hence the prediction is always accurate with K=1. If

validation error curve would have been similar, our choice of K would have been 1. Following is the validation error curve with varying value of K:

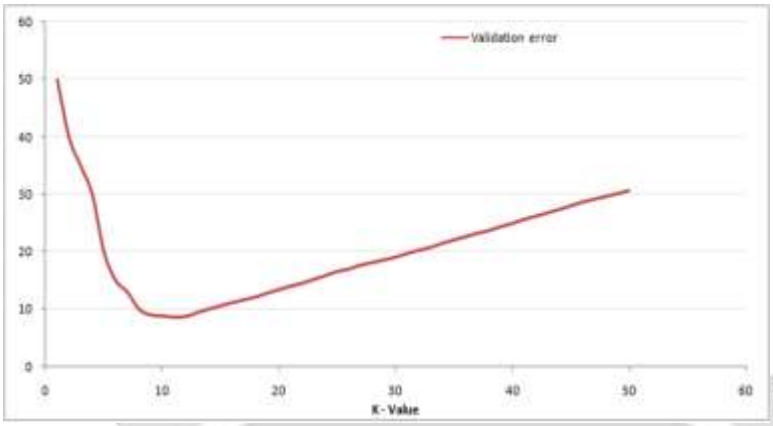

Chart-2

At K=1, we were over fitting the boundaries. Hence, error rate initially decreases and reaches a minimal. After the minima point, it then increase with increasing K. To get the optimal value of K, you can segregate the training and validation from the initial dataset. Now plot the validation error curve to get the optimal value of K. This value of K should be used for all predictions.

4.1 Working of KNN Algorithm

K-nearest neighbors (KNN) algorithm uses 'feature similarity' to predict the values of new data points which further means that the new data point will be assigned a value based on how closely it matches the points in the training set. We can understand its working with the help of following steps –

Step 1 − For implementing any algorithm, we need dataset. So during the first step of KNN, we must load the training as well as test data.

Step 2 − Next, we need to choose the value of K i.e. the nearest data points. K can be any integer.

Step 3 − For each point in the test data do the following −

3.1 − Calculate the distance between test data and each row of training data with the help of any of the method namely: Euclidean, Manhattan or Hamming distance. The most commonly used method to calculate distance is Euclidean.

3.2 − Now, based on the distance value, sort them in ascending order.

3.3 − Next, it will choose the top K rows from the sorted array.

3.4 − Now, it will assign a class to the test point based on most frequent class of these rows.

Step 4 – End

Example

The following is an example to understand the concept of K and working of KNN algorithm −

Suppose we have a dataset which can be plotted as follows −

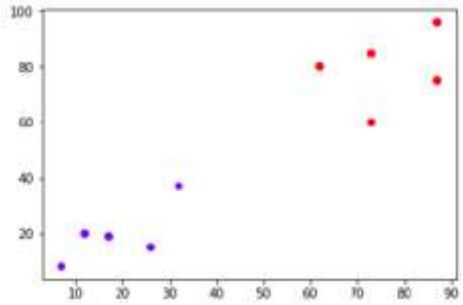

Chart-3

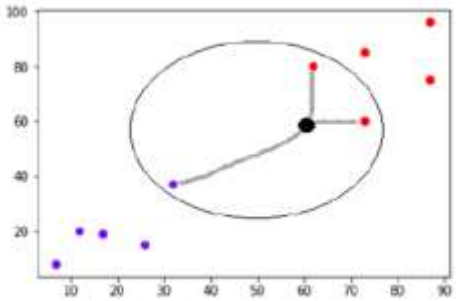

Chart-4

Now, we need to classify new data point with black dot (at point 60,60) into blue or red class. We are assuming K = 3 i.e. it would find three nearest data points. It is shown in the next diagram −

We can see in the above diagram the three nearest neighbors of the data point with black dot. Among those three, two of them lies in Red class hence the black dot will also be assigned in red class.

Advantages

1. The algorithm is simple and easy to implement.
2. There's no need to build a model, tune several parameters, or make additional assumptions.
3. The algorithm is versatile. It can be used for classification, regression, and search (as we will see in the next section).

Disadvantages

1. The algorithm gets significantly slower as the number of examples and/or predictors/independent variables increase.

4. Outcome & Discussion

In this project we applied this algorithm with the java application. So we can do data mining securely. It can prevent user private information from being leaked in the released environment, ensure the authenticity of the published data. In this application as the data increases mining time also increases.

5. CONCLUSIONS

The main purpose of Data mining analytics is to gain useful information from a large volume of heterogeneous data. However, having access to large-scale, distributed datasets presents certain privacy and security concerns which we have discussed briefly in this paper. We also discussed various method available for data mining and various gaps present in them. We have also seen the k nearest neighbor algorithm which is one of the basic and simplest algorithm available for data mining.

6. REFERENCES

[1] International Journal of Engineering Research & Technology (IJERT) ISSN: 2278-0181 IJERTV4IS100473 www.ijert.org (This work is licensed under a Creative Commons Attribution 4.0 International License.) Vol. 4 Issue 10, October-2015

[2] [IEEE 2013 Nirma University International Conference on Engineering (NUiCONE) - Ahmedabad, India (2013.11.28-2013.11.30)] 2013 Nirma University International Conference on Engineering (NUiCONE) - An efficient method for privacy preserving data mining in secure multiparty computation

[3] [IEEE 2019 IEEE Congress on Evolutionary Computation (CEC) - Wellington, New Zealand (2019.6.10-2019.6.13)] 2019 IEEE Congress on Evolutionary Computation (CEC) - A Swarm-based Data Sanitization Algorithm in Privacy-Preserving Data Mining

[4] [IEEE 2011 IEEE 9th International Symposium on Intelligent Systems and Informatics (SISY 2011) - Subotica, Serbia (2011.09.8-2011.09.10)] 2011 IEEE 9th International Symposium on Intelligent Systems and Informatics - Privacy preserving in data mining - Experimental research on SMEs data.

[5] [IEEE 2010 International Conference on Artificial Intelligence and Education (ICAIE) - Hangzhou, China (2010.10.29-2010.10.30)] 2010 International Conference on Artificial Intelligence and Education (ICAIE) - Application of Data Mining for emotional intelligence based on cluster analysis

[6] [IEEE 2014 Recent Advances in Engineering and Computational Sciences (RAECS) - Chandigarh, India (2014.03.6-2014.03.8)] 2014 Recent Advances in Engineering and Computational Sciences (RAECS) - Association rule sharing model for privacy preservation and collaborative data mining efficiency

[7] M. B. Malik, M. A. Ghazi, and R. Ali, ''Privacy preserving data mining techniques: Current scenario and future prospects,'' in Proc. 3rd Int. Conf. Comput. Commun. Technol. (ICCCT), Nov. 2012, pp. 26–32.

[8] [IEEE 2012 4th International Conference on Computational Intelligence and Communication Networks (CICN) - Mathura, Uttar Pradesh, India (2012.11.3-2012.11.5)] 2012 Fourth International Conference on Computational Intelligence and Communication Networks - Privacy Preserving in Data Mining Using Hybrid Approach

[9] [IEEE 2016 7th International Conference on Computer Science and Information Technology (CSIT) - Amman, Jordan (2016.7.13-2016.7.14)] 2016 7th International Conference on Computer Science and Information Technology (CSIT) - Privacy preserving data mining on published data in healthcare: A survey
[10] M. A. Sheela and K. Vijayalakshmi, ''A novel privacy preserving decision tree induction,'' in Proc. IEEE Conf. Inf. Commun. Technol. (ICT), Apr. 2013, pp. 1075–1079.
[11] Malina L, Hajny J (2013) Efficient security solution for privacy-preserving cloud services. In: 36th international conference on telecommunications and signal processing (TSP), pp 23–27. http://doi.org/10.1109/TSP.2013.6613884
[12] Kamakshi P (2012) Automatic detection of sensitive attribute in PPDM. In: IEEE international conference on computational intelligence & computing research (ICCIC)
[13] Mukkamala R, Ashok VG (2011) Fuzzy-based methods for privacy-preserving data mining. In: IEEE eighth international conference on information technology: new generations (ITNG)
[14] Patil BB, Patankar AJ (2013) Multidimensional k-anonymity for protecting privacy using nearest neighborhood strategy. In: IEEE international conference on computational intelligence and computing research (ICCIC)
[15] Wang H, Hu C, Liu J (2010) Distributed mining of association rules based on privacy-preserved method. In: Internationalsymposium on information science and engineering (ISISE), pp 494–497. http://doi.org/10.1109/ISISE.2010.125
[16] Tai C-H, Huang J-W, Chung M-H (2013) Privacy preserving frequent pattern mining on multi-cloud environment. In: 2013 international symposium on biometrics and security technologies (ISBAST)
[17] W. Gan, J. C. W. Lin, H. C. Chao, S. L. Wang, P. S. Yu, "Privacy preserving utility mining: a survey," IEEE International Conference on Big Data, pp. 2617–2626, 2018.
[18] L. Xu, C. Jiang, J. Wang, J. Yuan, and Y. Ren, ''Information security in big data: Privacy and data mining,'' IEEE Access, vol. 2, pp. 1149–1176, 2014