Indian J.Sci.Res. 17 (1): 164-169, 2017 ISSN: 0976-2876 (Print)
ISSN: 2250-0138 (Online)

PRIVACY PRESERVING USING ANONYMIZATION AND PERTURBATION IN CLASSIFICATION

KIRAN ISRANI[a1], SHALU CHOPRA[b] AND KAJAL JEWANI[c]

[a]M.E. Student, [b]Associate Professor, [c]Assistant Professor

ABSTRACT

Data mining techniques are used for analysis purpose, but the data may contains sensitive information about individuals, which individuals don't want to be revealed, during data mining process. k anonymity is one of the technique, which is used for preserving privacy in Data mining. In k anonymity, k records will appear similar in quasi identifier attribute. For achieving this generalization or suppression can be used. In Generalization attribute values are replaced by less specific value and in suppression attribute values are suppressed by meaningless characters like '*' or '?'. In this paper we have proposed k anonymity using suppression, crossover and perturbation in classification tree. Original dataset will be input to our algorithm and anonymized data set is output, in anonymized data set number of tuple are same as original data set and we are comparing accuracy of original dataset with anonymized dataset. Accuracy of anonymized dataset is better as compared to original data set.

KEYWORDS: Anonymization, Classification, Crossover, Perturbation, Privacy Preserving, PPDM.

Data mining is a process of analyzing data for hidden patterns and information. While performing data mining it may happen sensitive information about the individual may get revealed, to overcome this new branch got emerged called as privacy preserving in data mining. In [Sweeney, 2002] L. Sweeney proposed k anonymity method, in which each record cannot be distinguished from k-1 records. K anonymity is achieved by generalization and suppression. Generalization is a process of replacing attributes values with less specific value. For example age attribute value is 40; it can be generalized to less than 45 or greater than 35value. Suppression is hiding values using '*' or '?'. Generalization and suppression is applied on quasi identifier attributes, these are the attributes whose values can be linked with other data base to re-identify the tuples identity. Drawback of generalization is it requires manual domain hierarchy for quasi- identifier attribute. In data perturbation the attribute values are perturbed by adding noise or by translation or rotation technique [Israni and Chopra, 2016].

In this paper we are proposing hybrid approach in PPDM which will generate anonymized data set from original data set. First classification tree is generated from original data set using C4.5 algorithm then anonymization of data is done by considering tuples at each leaf node. We have used k anonymity using suppression with crossover and perturbation to get better accuracy.

LITERATURE REVIEW

PPDM transforms the data so that privacy will be preserved while performing data mining task [Malik et. al., 2012]. There are various techniques in PPDM such as anonymization perturbation randomization, condensation and cryptography etc. but there is no single technique available which can balance between disclosure and utility of data. Now a day's hybrid approaches are also getting developed. Hybrid approaches in PPDM combines two or more of above techniques.

Anonymization can be done by following techniques k anonymization using generalization and suppression, p sensitive k anonymity, (α, k) anonymity, t closeness.

K-anonymity using generalization and suppression protects from identity disclosure but fails to protect from attribute disclosure [Israni and Chopra, 2016]. Due to this p sensitivity k anonymity technique got evolved which protects from identity disclosure and sensitive attribute disclosure.

In p sensitivity k anonymity a group of records which satisfies k anonymity will have distinct confidential attribute value at least p times in that group, to overcome attribute disclosure problem k must be greater than p value [Truta and Vinay, 2006].

(α, k)anonymity [Wong et. al., 2007] they have shown two types of generalization global recording and local recording. Global recording loose more information than local recording. They suggested local recording method for generalization is better than global generalization.

l diversity group of tuples which satisfies with k anonymity will have l diverse sensitive attribute values. In [Machanavajjhala et. al., 2007] they have shown 4 anonymous 3 diverse table and it will be difficult for attacker to identify sensitive attribute value for a particular record.

[1]Corresponding author

t closeness is enhancement on l diversity method, In [Aggarwal and Yu] they suggest that attribute value corresponding to disease is more sensitive when positive than negative and distribution of sensitive attribute value in anonymized group should not differ from global distribution by more than t threshold.

Perturbation original values are altered by synthetic values. It can be done by additive noise or data swapping or synthetic data generation or multiplicative perturbation this can be done by rotation or projection.

In Randomization [Aggarwal and Yu], data is altered by using probability distribution. This technique is simple to use and does not require knowledge of distribution of other records. But this technique considers all records equally irrespective of local density.

Condensation technique uses sudo (dummy) data rather than modified data so it will be difficult for attacker to identify the actual data.

A cryptography technique is used when multiple parties are involved for giving input without actually sharing their data with each other [Israni and Chopra, 2016].

Baghel and Dutta, 2013 have used modified C4.5 algorithm on unrealized perturbed datasets and performed experiments on unrealized dataset and original dataset using C4.5 and modified C4.5 algorithm, in results they showed performance of modified C4.5 on unrealized data set is better.

Xu et al., 2014 have identified users involved in process of data mining such as provider, data collector, data miner and decision maker. They identified privacy concers of users and methods that can be adopted to protect sensitive information.

Taneja et al., 2014 proposed encryption and perturbation in clustering. Encryption of sensitive attributes using ASCII code and special character. For primary attribute C Tree and perturbation technique is used.perturbation will not reveal ones identity and original dataset can be reconstructed from perturbed data.

Saranya et al., 2015 have given survey on PPDM techniqes in classification, clustering, and association rule mining with their merits and demerits.

Lohiya and Ragha, 2012 proposed a hybrid technique in which they used randomization and generalization. In this approach first they randomize the data and then generalized the randomized data. This technique protects private data with better accuracy; also it can reconstruct original data and provide data with no information loss.

Deivanai et. al., 2011 proposed a method by using k anonymization using suppression. They performed suppression only on certain records depending upon other attribute values; there method identifies attributes which have less influence on classification of data records, and those values are suppressed. This method shows a higher predictive performance when compared to existing methods. Limitation of this method is the data loss due to suppression. The suppressed data does not contribute to complete mining. Thus the accuracy of data will be comparatively lower.

Kisilevich et. al., 2010 proposed a k-anonymity classification tree based suppression (kACTUS). They created decision tree form original dataset. It then uses this tree to apply k-anonymity to the dataset while maintaining balance between k-anonymity constraints and classification quality. The resultant anonymous dataset can be given to an external user, who can use any classification algorithm for training using anonymous data set.

METHOD

The objective of this paper is to get anonymized dataset. First Classification tree is generated using original data set using C4.5 algorithm. Classification tree will be input to kactus algorithm. Each node will have some records associated with it depending upon splitting criteria of parent node. Node having k or more than k instances is complying node otherwise it will be considered as non-complying node [Kisilevich et. al., 2010]. Leaf nodes will be having subset of original dataset. If we combine all leaf node instances that will be original data set.

We check at leaf nodes, if it contains k or more than k instances then we transfer these instances to anonymized data set. If it contains less than k then we perform anonymization by suppression, crossover and perturbation techniques. Performance of classifier trained on anonymized data set is better as compared to trained original data set. Below are algorithms for kactus, anonymization, crossover suppression and perturbation.

Kactus Algorithm

Input: Classification tree, k-anonymity threshold, set of quasi-identifiers.

Output: Instance in anonymized data set.

Step 1: Iterate over the classification tree while it has at least one root node.

Step 2: find the longest path from it to the leaf node.

Step 3: If the longest path is of a height greater than or equal to 1 it means that the root node has children then call Anonymization method, otherwise check how many instances are associated with the root node.

Step 4: If the number of instances with longest node is greater or equal to the k-anonymity threshold then we move the instances to the anonymized dataset. And then remove leaf nodes of longest node from the classification tree.

Step 5: If the number of instances with longest node is less than the k-anonymity threshold then add the node to a set of non complying nodes and check how many instances in total are associated with the non complying nodes stored in the non-complying set.

Step 6: If the total number of instances is greater or equals to the k-anonymity threshold then we move root of non complying node instances to the anonymized dataset.

Anonymization

Input: set of instances with node.

Output: Instance in anonymized data set.

Step 1: check how many instances are associated with the longest node. That is count all the instances of its child nodes.

Step 2: If the total number of instances with longest node is less than the k threshold then we perform perturbation on the child nodes, and remove child nodes. Otherwise find complying and non-complying leaf nodes (children of the longest node).

Step 3: If non-complying leafs set is empty then just move all instances associated with the complying leaf node and remove all the children nodes.

Step 4: Otherwise call crossover method and Suppression method for complying and non complying nodes.

Crossover

Input: complying and non complying node.

Output: Instance in anonymized data set.

Step 1: For each non-complying node, calculate how many instances are required in order to make the non-complying nodes compliant. Then calculate the required-ratio as required instances divided by K Threshold and compared to the crossover threshold (CoT).

Step 2: Perform crossover only if the ratio of required instances is less than the predefined CoT.

Step 3: For every non-complying node, search for best complying node from available complying nodes using entropy of each complying nodes.

Step 4: If the best complying node is not found, perform perturbation, otherwise move the instances associated with the non-complying node to the anonymized dataset and perform crossover.

Step5: In crossover instances from complying node are moved to non complying node.

Suppression

Input: complying and non complying node.

Ouput: Instance in anonymized data set.

Step1: For each complying node, calculate how many instances; it is capable of compensating to non-complying nodes.

Step2: If the number of instances which the complying node can compensate is greater or equal to the number of required instances then compensation is possible, otherwise perform perturbation.

Step3: In compensation move required number of instances from the complying node to non-complying node. Then non complying node instances will be moved to anonymized dataset with the quasi attribute values suppressed and the remaining instances of the complying node will be moved to the anonymized dataset.

Perturbation

Input: non complying node.

Output: Instance in anonymized data set.

Step1: For each non-complying node, find the parent node splitting attribute value.

Step2: perturbate by adding parent node attribute value with half of non complying node instance value.

Step3: Move instances to the anonymized dataset with the quasi attribute values suppressed.

IMPLEMENTATION

The algorithm is implemented using Java in Net Beans Environment. We have used Heart stat log, sonar data set and Haberman dataset from the UC Irvine machine learning repository. In Heart stat log Data set

Number of Instances is 270, with 13 attributes and one class attribute. For testing we have used age, sex, resting blood pressure attributes as quasi attributes. In Sonar data set there are 60 attributes, one class attribute and 208 instances. All the attributes are real. For quasi identifiers we are using any attribute number. In Haberman's survival data set there are 4 attributes along with class attribute and 306 are total instances. We have used age of patient at time of operation (numerical) and patient's year of operation as quasi attribute.

Implementation of proposed model is done in java, original data set is given to C4.5 classifier to generate classification tree and this tree is given as input to our model which is generating anonymous data set. For testing we are checking for accuracy in weka tool. Accuracy of anonymized data set is compared with original dataset that is if original data set is given to classifier and anonymized data set is given to classifier, the accuracy of anonymized data set is better than original dataset. Below tables 1 show the 10 tuples of heart statlog data set and table 2 shows anonymized data set of those 10 tuples. Figure 1 show accuracy of original data set and Figure 2 shows accuracy of anonymized data set.

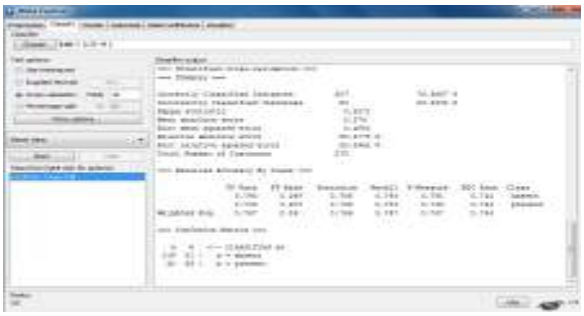

Figure 1: Accuracy Of Heart Stat Log Original Dataset

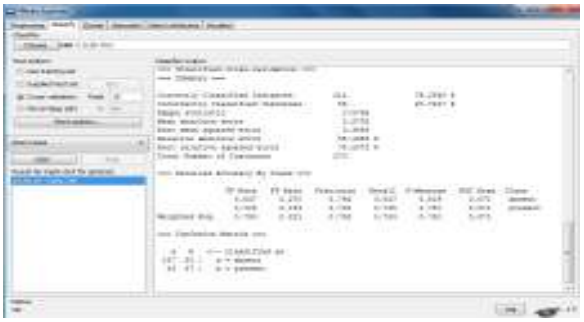

Figure 2: Accuracy Of Heart Stat Log Anonymized Dataset

Table 1: Heart Statlog Original Dataset

Age	Sex	Chest	Resting blood pressure	Serum cholesterol	Fasting blood sugar	Resting electrocar diographic	Maximum heart rate achieved	Exercise induced angina	Old peak	Slope	Number of major vessels	Thal	Class
70	1	4	130	322	0	2	109	0	2.4	2	3	3	P
67	0	3	115	564	0	2	160	0	1.6	2	0	7	A
57	1	2	124	261	0	0	141	0	0.3	1	0	7	P
64	1	4	128	263	0	0	105	1	0.2	2	1	7	A
74	0	2	120	269	0	2	121	1	0.2	1	1	3	A
65	1	4	120	177	0	0	140	0	0.4	1	0	7	A
56	1	3	130	256	1	2	142	1	0.6	2	1	6	P
59	1	4	110	239	0	2	142	1	1.2	2	1	7	P
60	1	4	140	293	0	2	170	0	1.2	2	2	7	P
63	0	4	150	407	0	2	154	0	4	2	3	7	P

In table 1 the original data set is shown which is given to classifier C4.5 which gives classification tree that will be input to our proposed method. For implementation we have given input for k value as 5. Age, sex and thal attributes as quasi identifiers, algorithm find best splitting attribute and it splits according to age attribute and resting blood pressure value as less than or equal to or greater than split value.

From table 2 it can be observed that there are 5 anonymized instances from which 4 instances hold class value as absent and 1 instance holds class value as present, in this case 4 instances with node form a non complying node therefore to make it as complying node. 1 instance from sibling node has been taken to make this node as complying node. Complying node instances are directly transferred to anonymized data set.

Table 2: Heart Statlog Anonymized Dataset

Age	Sex	Chest	Resting blood pressure	Serum cholesterol	Fasting blood sugar	Resting electrocardiographic	Maximum heart rate achieved	Exercise induced angina	Old peak	Slope	Number of major vessels	Thal	Class
?	?	3	115	564	0	2	160	0	1.6	2	0	?	Absent
?	?	4	128	263	0	0	105	1	0.2	2	1	?	Absent
?	?	2	120	269	0	2	121	1	0.2	1	1	?	Absent
?	?	4	120	177	0	0	140	0	0.4	1	0	?	Absent
?	?	4	130	322	0	2	109	0	2.4	2	3	?	Present
57	1	2	124	261	0	0	141	0	0.3	1	0	7	Present
56	1	3	130	256	1	2	142	1	0.6	2	1	6	Present
59	1	4	110	239	0	2	142	1	1.2	2	1	7	Present
60	1	4	140	293	0	2	170	0	1.2	2	2	7	Present
63	0	4	150	407	0	2	154	0	4	2	3	7	Present

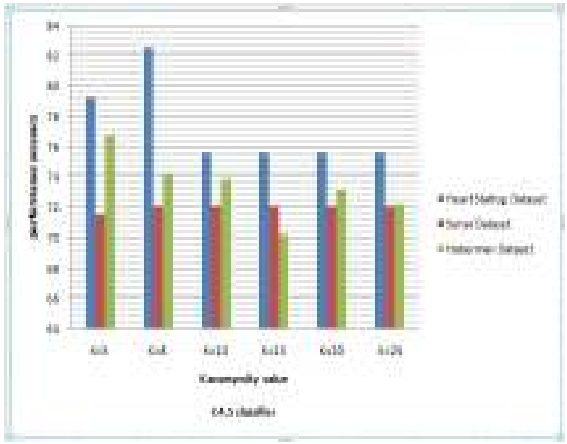

Figure 3: Accuracy on different values of k for datasets in C4.5 Classifier

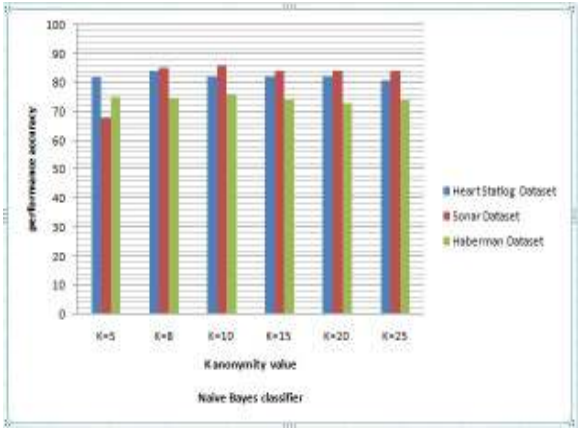

Figure 4: Accuracy on different values of k for datasets in C4.5 Classifier

Table 3: Performance Accuracy of k on Datasets

Data set	Inducer	K- anonymity					
		K=5	K=8	K=10	K=15	K=20	K=25
Heart -statlog	C 4.5	79.25	82.59	75.55	75.55	75.55	75.55
	NB	81.85	84.07	82.22	82.22	82.22	80.74
Sonar	C 4.5	71.63	72.11	72.11	72.11	72.11	72.11
	NB	67.78	85.09	86.05	84.13	84.13	84.13
Haberman	C 4.5	76.79	74.18	73.85	70.26	73.20	72.22
	NB	75.16	74.83	75.81	74.50	72.87	73.85

We have used three data sets heart statlog sonar and Haberman survival. Accuracy is checked using weka tool in J48 accuracy of original data sets for heart-statlog is 76.66 for sonar data set is 71.15 and for Habermans survival is 71.89.In Naïve Bayes classifier accuracy of heart statlog is 83.70, for sonar data set is 67.78 and for Habermans survival is 74.83. From the table 3 we observe that accuracy of our model is better and it is also observed that accuracy with increase in k value remains constant. Figure 3 and 4 shows graph of accuracy on different k values for c4.5 classifier and Naïve Bayes classifier.

CONCLUSION AND FUTURE WORK

In this paper we have presented a hybrid approach for PPDM for classification task using k anonymity, crossover and perturbation. Previously existing models [Deivanai et. al., 2011][Kisilevich et. al., 2010] has limitation with data loss and accuracy of data was comparatively lower. To overcome this limitation, we have combined perturbation and k anonymity for classification tree. Our approach mainly focuses on avoiding the data loss (in terms of instance loss) and

improving the accuracy of anonymized data. Our method can be extended to be used with clustering and we have used k anonymization technique instead of other anonymity technique can also be used.

REFERENCES

Sweeney L., 2002. "Achieving k-anonymity privacy protection using generalization and suppression". International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, **10**(5):571-588.

Aggarwal C.C. and Yu P.S., "A General Survey of Privacy-Preserving Data Mining Models and Algorithms", in springer ISBN 978-0-387-70991-8, e-ISBN 978-0-387-70992-5,DOI 10.1007/978-0-387-70992-5 .

Israni K. and Chopra S., 2016. "Survey on Anonymization Technique for Privacy Preserving Data Mining (PPDM)" International Journal of Innovative Research in Computer and Communication Engineering, ISSN (Online): 2320-9801, **4**(11).

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

Truta T. and Vinay B., 2006. "Privacy Protection: p-Sensitive k-Anonymity Property", In Proceedings of the 22nd International Conference on Data Engineering Workshops, pp. 94-103.

Wong R., Liu Y., Yin J., Huang Z., Fu A.W.-C and Pei J., 2007. "(α, k)-anonymity Based Privacy Preservation by Lossy join", APWeb/WAIM'07 Proceedings of the joint 9th Asia-Pacific web and 8th International Conference on Web-age Information Management Conference.

Machanavajjhala A., Kifer D., Gehrke J. and Venkitasubramaniam M., 2007. "ℓ-Diversity: Privacy Beyond k-Anonymity", ACM Transactions on Knowledge Discovery from Data (TKDD), **1**(1):1–47.

Lohiya S. and Ragha L., 2012. "Privacy Preserving in Data Mining Using Hybrid Approach", in proceedings of 2012 Fourth International Conference on Computational Intelligence and Communication Networks, IEEE.

Deivanai P., Nayahi J.J.V. and Kavitha V., 2011. "A Hybrid Data Anonymization integrated with Suppression for Preserving Privacy in mining multi party data" in proceedings of International Conference on Recent Trends in Information Technology, IEEE.

Kisilevich S., Rokach L., Elovici Y. and Shapira B., 2010. "Efficient Multi-Dimensional Suppression for K-Anonymity", IEEE Transactions on Knowledge and data Engineering, **22**(3):334 -347.

Baghel R. and Dutta M., 2013. "Privacy Preserving Classification By Using Modified C4.5" IEEE International conference on Data Mining (ICDM), August 2013.

Xu L., Jiang C., Wang J., Yuan J. and Ren Y., 2014. "Information Security in Big Data: Privacy and Data Mining". Date of publication October 9, 2014, date of current version October 20, 2014. Digital Object Identifier 10.1109/ACCESS. 2014.2362522.

Taneja S., Khanna S., Tilwalia S. and Ankita, 2014. A Hybrid C- Tree Algorithm for Privacy Preserving Data Mining", International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, **4**:21-24.

Saranya K., Premalatha K. and Rajasekar S.S., 2015. "A Survey on Privacy Preserving Data Mining", IEEE sponsored 2nd international conference on electronics and communication system (icecs 2015).