CLASSIFICATION AND EVALUATION THE PRIVACY PRESERVING DATA MINING USING ANONYMIZING

T SOWJANYA 1*, L NARENDRA 2*

1. *PG Scholar, Dept of MCA, Lakireddy Balireddy College Of Engineering, Mylavaram.*
2. *Asst.Prof, Dept of MCA, Lakireddy Balireddy College Of Engineering, Mylavaram.*

Abstract: Data mining procedures are utilized for investigation reason, yet the information may contains touchy data about people, which people would prefer not to be uncovered, amid data mining process. k obscurity is one of the methods, which is utilized for safeguarding security in Data mining. In k secrecy, k records will seem comparable in semi identifier trait. For accomplishing this speculation or concealment can be utilized. In Generalization characteristic esteems are supplanted by less particular esteem and in concealment property estimations are smothered by good for nothing characters like '*' or '?'. In this paper we have proposed k anonymity utilizing concealment, hybrid and irritation in arrangement tree. Unique dataset will be contribution to our algorithm and anonymized informational collection is output, in anonymized informational index number of tuple are same as unique informational collection and we are contrasting precision of unique dataset and anonymized dataset. Exactness of anonymized dataset is better when contrasted with unique informational index.

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

1. Introduction

Data mining is a procedure of breaking down information for shrouded examples and data. While performing data mining it might happen touchy data about the individual may get uncovered, to beat this new branch got rose called as protection saving in data mining. In [Sweeney, 2002] L. Sweeney proposed k anonymity strategy, in which each record can't be recognized from k-1 records. K secrecy is accomplished by speculation and concealment. Speculation is a procedure of supplanting properties esteems with less particular esteem. For instance age quality esteem is 40; it can be summed up to under 45 or more prominent than 35value. Concealment is concealing esteems utilizing '*' or '?'. Speculation and concealment is connected on semi identifier characteristics, these are the qualities whose esteems can be connected with other information base to re-recognize the tuples personality. Downside of speculation is it requires manual area chain of command for semi identifier quality. In information annoyance the characteristic esteems are irritated by including clamor or by interpretation or pivot strategy [Israni and Chopra, 2016]. In this paper we are proposing cross breed approach in PPDM which will create anonymized informational index from unique informational collection. To begin with arrangement tree is produced from unique informational collection utilizing C4.5 algorithm then anonymization of information is finished by thinking about tuples at each leaf hub. We have utilized k secrecy utilizing concealment with hybrid and annoyance to show signs of improvement precision.

2. Literature Review

PPDM changes the information with the goal that protection will be saved while performing data mining assignment [Malik et. al., 2012]. There are different methods in PPDM, for example, anonymization bother randomization, buildup and cryptography and so on yet there is no single procedure accessible which can adjust amongst exposure and utility of information. Presently a day's cross breed approaches are additionally getting created. Half breed approaches in PPDM joins at least two of above strategies. Anonymization should be possible by following methods k anonymization utilizing speculation and concealment, p delicate k secrecy, (α, k) anonymity, t closeness. K-anonymity utilizing speculation and concealment shields from character revelation however neglect to shield from property exposure [Israni and Chopra, 2016]. Because of this p affectability k secrecy system got advanced which shields from personality revelation and touchy quality exposure. In p affectability k anonymity a gathering of records which fulfills k secrecy will have unmistakable secret property estimation in any event p times in that gathering, to conquer quality exposure issue k must be more prominent than p esteem [Truta and Vinay, 2006]. (α, k)anonymity [Wong et. al., 2007] they have demonstrated two sorts of speculation worldwide account and nearby chronicle. Worldwide chronicle free more data than nearby account. They proposed nearby chronicle strategy for speculation is superior to worldwide speculation.

Irritation unique esteems are modified by engineered esteems. It should be possible by added substance clamor or information swapping or engineered information age or multiplicative irritation this should be possible by turn or projection. In Randomization [Aggarwal and Yu], information is modified by utilizing likelihood appropriation. This system is easy to utilize and does not require information of circulation of different records. In any case, this strategy considers all records similarly independent of neighborhood thickness. Buildup procedure utilizes sudo (sham) information as opposed to altered information so it will be troublesome for assailant to recognize the genuine information. A cryptography system is utilized when various gatherings are included for giving contribution without really imparting their information to each other [Israni and Chopra, 2016]. Baghel and Dutta, 2013 have utilized changed C4.5 algorithm on undiscovered annoyed datasets and performed probes hidden dataset and unique dataset utilizing C4.5 and altered C4.5 algorithm, in comes about they indicated execution of adjusted C4.5 on undiscovered informational collection is better. Xu et al., 2014 have distinguished clients associated with procedure of data mining, for example, supplier, information gatherer, information excavator and chief. They distinguished security concers of clients and techniques that can be embraced to ensure touchy data. Taneja et al., 2014 proposed encryption and irritation in grouping. Encryption of touchy characteristics utilizing ASCII code and unique character. For essential trait C Tree and annoyance strategy is used. Perturbation won't uncover ones character and unique dataset can be recreated from bothered information. Saranya et al., 2015 have given study on PPDM techniques in order, bunching, and affiliation administered mining with their benefits and faults. Lohiya and Ragha, 2012 proposed a crossover strategy in which they utilized randomization and speculation. In this approach first they randomize the information and afterward summed up the randomized information. This strategy ensures private information with better precision; additionally it can recreate unique information and furnish information with no data misfortune.

3. Method

The goal of this paper is to get anonymized dataset. In the first place Classification tree is created utilizing unique informational collection utilizing C4.5 algorithm. Characterization tree will be contribution to kactus algorithm. Every hub will have a few records related with it relying on part criteria of parent hub. Hub having k or more than k examples is agreeing hub else it will be considered as non-consenting hub [Kisilevich et. al., 2010]. Leaf hubs will have subset of unique dataset. In the event that we consolidate all leaf hub occurrences that will be unique informational collection. We check at leaf hubs, on the off chance that it contains k or more than k cases then we exchange these cases to anonymized informational collection. On the off chance that it contains not as much as k then we perform anonymization by concealment, hybrid and annoyance systems. Execution of classifier prepared on anonymized informational index is better when contrasted with prepared unique informational collection. The following are algorithms for kactus, anonymization, hybrid concealment and annoyance.

Kactus Algorithm

Input: Classification tree, k-obscurity limit, set of semi identifiers.

Output: Instance in anonymized informational index.

Stage 1: Iterate over the order tree while it has no less than one root hub.

Stage 2: locate the longest way from it to the leaf hub.

Stage 3: If the longest way is of tallness more noteworthy than or equivalent to 1 it implies that the root hub has kids at that point call Anonymization strategy, generally check what number of occasions are related with the root hub.

Stage 4: If the quantity of cases with longest hub is more noteworthy or equivalent to the k-secrecy limit then we move the cases to the anonymized dataset.

Stage 5: If the quantity of occasions with longest hub is not as much as the k-anonymity edge at that point add the hub to an arrangement of non consenting hubs and check what number of cases altogether are related with the non going along hubs put away in the non-going along set.

Stage 6: If the aggregate number of examples is more noteworthy or equivalents to the k-obscurity edge then we move base of non going along hub cases to the anonymized dataset.

Anonymization

Input: set of occurrences with hub.

Output: Instance in anonymized informational index.

Stage 1: check what number of cases is related with the longest hub. That is check every one of the occasions of its youngster hubs.

Stage 2: If the aggregate number of cases with longest hub is not as much as the k limit then we perform annoyance on the tyke hubs, and evacuate kid hubs. Generally find agreeing and non-consenting leaf hubs (offspring of the longest hub).

Stage 3: If non-agreeing leafs set is unfilled then simply move all occurrences related with the going along leaf hub and evacuate every one of the kid's hubs.

Stage 4: Otherwise call hybrid strategy and Suppression technique for agreeing and non going along hubs.

Crossover

Input: going along and non consenting hub.

Output: Instance in anonymized informational index.

Stage 1: For each non-going along hub, figure what number of examples is required so as to make the non-agreeing hubs consistent. At that point figure the required-proportion as required occurrences separated by K Threshold and contrasted with the hybrid edge.

Stage 2: Perform hybrid just if the proportion of required occasions is not exactly the predefined CoT.

Stage 3: For each non-agreeing hub, scan for best consenting hub from accessible going along hubs utilizing entropy of each going along hubs.

Stage 4: If the best consenting hub isn't found, perform irritation, generally move the examples related with the non-going along hub to the anonymized dataset and perform hybrid.

Stage5: In hybrid examples from going along hub are moved to non agreeing hub.

Suppression

Input: agreeing and non going along hub.

Output: Instance in anonymized informational collection.

Stage1: For each consenting hub, compute what number of occurrences; it is equipped for remunerating to non-agreeing hubs.

Stage2: If the quantity of cases which the consenting hub can remunerate is more noteworthy or equivalent to the quantity of required occasions then pay is conceivable, generally perform annoyance.

Stage3: In remuneration move required number of occasions from the consenting hub to non-going along hub. At that point non going along hub occasions will be moved to anonymized dataset with the semi quality esteems stifled and the rest of the occurrences of the going along hub will be moved to the anonymized dataset.

Perturbation

Input: non agreeing hub.

Output: Instance in anonymized informational collection.

Stage1: For each non-agreeing hub, discover the parent hub part property estimation.

Stage2: Perturbate by including guardian hub quality incentive with half of non going along hub case esteem.

Stage3: Move occasions to the anonymized dataset with the semi characteristic esteems smothered.

4. Implementation

The algorithm is executed utilizing Java in Net Beans Environment. We have utilized Heart detail log, sonar informational collection and Haberman dataset from the UC Irvine machine learning archive. In Heart detail log Data setNumber of Instances is 270, with 13 traits and one class characteristic. For testing we have utilized age, sex, resting pulse qualities as semi traits. In Sonar informational index there are 60 traits, one class quality and 208 cases. Every one of the properties are genuine. For semi identifiers we are utilizing any trait number. In Haberman's survival informational index there are 4 qualities alongside class characteristic and 306 are add up to occasions. We have utilized period of patient at time of task (numerical) and patient's time of activity as semi property. Usage of proposed show is done in java, unique informational index is given to C4.5 classifier to produce grouping tree and this tree is given as contribution to our model which is creating mysterious informational collection. For testing we are checking for precision in weka apparatus. Precision of anonymized informational collection is contrasted and unique dataset that is if unique informational index is given to classifier and anonymized informational index is given to classifier, the exactness of anonymized informational index is superior to unique dataset. Beneath tables 1 demonstrate the 10 tuples of heart statlog informational collection and table 2 indicates anonymized informational index of those 10 tuples.

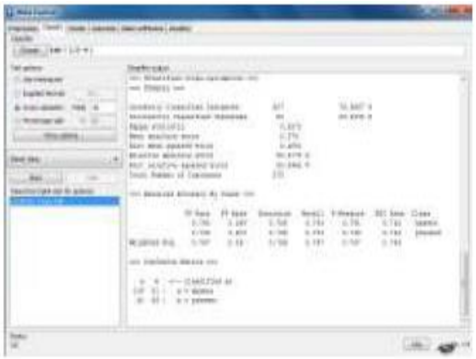

Figure 1: Accuracy of Heart Stat Log Original Dataset

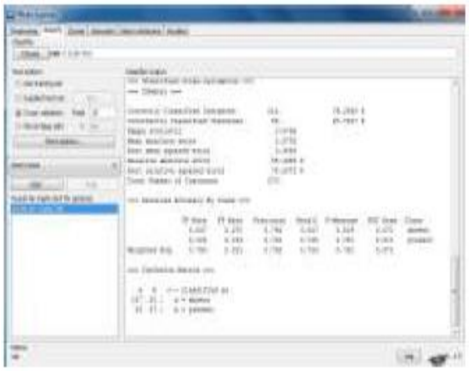

Figure 2: Accuracy of Heart Stat Log Anonymized Dataset.

Figure 1 indicates precision of unique informational collection and Figure 2 demonstrates exactness of anonymized informational collection.

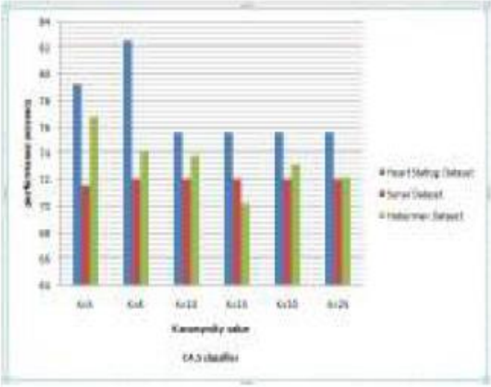

Figure 3: Accuracy on different values of k for datasets in C4.5 Classifier

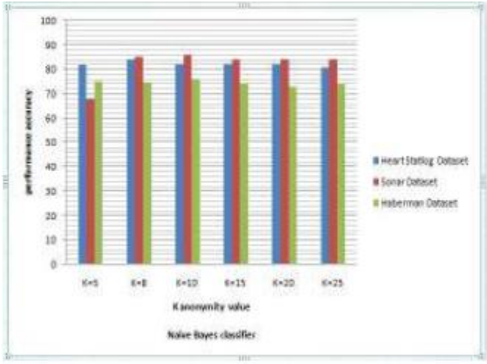

Figure 4: Accuracy on different values of k for datasets in C4.5 Classifier

Table 1: 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 utilized three informational collections heart statlog sonar and Haberman survival. Precision is checked utilizing weka device in J48 exactness of unique informational indexes for heart-statlog is 76.66 for sonar informational collection is 71.15 and for Habermans survival is 71.89.In Naive Bayes classifier exactness of heart statlog is 83.70, for sonar informational collection is 67.78 and for Habermans survival is 74.83. From the table 3 we watch that precision of our model is better and it is additionally watched that exactness with increment in k esteem stays steady. Figure 3 and 4 indicates chart of exactness on various k esteems for c4.5 classifier and Naive Bayes classifier.

5. Conclusion

In this paper we have exhibited a half and half approach for PPDM for grouping assignment utilizing k obscurity, hybrid and irritation. Beforehand existing models [Deivanai et. al., 2011][Kisilevich et. al., 2010] has constraint with information misfortune and precision of information was relatively lower. To defeat this confinement, we have consolidated irritation and k obscurity for characterization tree. Our approach mostly centers around staying away from the information misfortune (as far as example misfortune) and enhancing the exactness of anonymized information. Our strategy can be stretched out to be utilized with bunching and we have utilized k anonymization method rather than other obscurity procedure can likewise be utilized.

6. References

[1] L. Sweeney, "Replacing personally-identifying information medical records, the scrub system." in AMIA Fall Symposium, 1996, p. 333.

[2] V. T. Chakaravarthy, H. Gupta, P. Roy, M. K. Mohania, "Efficient techniques for document sanitization," in ACM Conference on Information and Knowledge Management, 2008, pp. 843–852.

[3] M. Tambe, Security Game Theory: Algorithms, Deployed Systems, and Lessons Learned. Cambridge University Press, 2011.

[4] Z. Wan, Y. Vorobeychik, W. Xia et al., "A game theoretic framework analyzing re-identification risk," PLoS One, 2015, in press.

[5] M. Bruckner, T. Scheffer, "Stackelberg games for adversarial ̈prediction problems," in ACM International Conference on Knowledge Discovery, Data Mining, 2011, pp. 547–555.

[6] N. Dalvi, P. Domingos, Mausam, S. Sanghai, D. Verma, "Adversarial classification," in ACM International Conference on Knowledge Discovery, Data Mining, 2004, pp. 99–108.

[7] M. Kantarcioglu, B. Xi, C. Clifton, "Classifier evaluation and ˇattribute selection against active adversaries," Data Mining and Knowledge Discovery, vol. 22, pp. 291–335, 2011.

[8] B. Li, Y. Vorobeychik, "Feature cross-substitution adversarial classification," in Neural Information Processing Systems, 2014, pp. 2087–2095.

[10] W. Jiang, M. Murugesan, C. Clifton, L. Si, "t-plausibility: semantic preserving text sanitization," in International Conference on Computational Science and Engineering, vol. 3, 2009, pp. 68–75.

[11] G. Szarvas, R. Farkas, R. Busa-Fekete, "State-of-the-art anonymization of medical records using an iterative machine learning framework," Journal of the American Medical InformaticsAssociation, vol. 14, no. 5, pp. 574–580, 2007.

[12] O. Uzuner, T. C. Sibanda, Y. Luo, P. Szolovits, "A deidentifier for medical discharge summaries," Artificial Intelligence in Medicine, no. 1, pp. 13–35, 2008.

[13] X. Xiao, Y. Tao, "M-invariance: towards privacy preserving re-publication of dynamic datasets," in ACM International Conference Management of Data, 2007, pp. 689–700.

[14] R. J. Bayardo, R. Agrawal, "Data privacy through optimal k-anonymization," International Conference on Data Engineering, 2005, pp. 217–228.

[15] X. He, A. Machanavajjhala, B. Ding, "Blowfish privacy: tuning privacy-utility trade-offs using policies," in ACM International Conference on Management Data, 2014, pp. 1447–1458.

[16] D. Kifer, A. Machanavajjhala, "Pufferfish: A framework for mathematical privacy definitions," in ACM Transactions on Database Systems, 2014, p. 3.

About Authors:

T.Sowjanya is currently pursuing at MCA Department, Lakireddy Balireddy College Of Engineering,Mylavaram,A.P .

L.Narendra is currently working as an Assistant Professor in MCA Department, Lakireddy Balireddy College Of Engineering , Mylavaram.A.P He research includes data mining.