ENHANCING THE EFFECTIVENESS OF PRIVACY PRESERVING DATA MINING (PPDM) BY USING CORRELATION BASED TRANSFORMATION STRATEGY

Aryan Grover

ABSTRACT

Preservation of security is a huge part of information mining. The primary goal of PPDM is to cover up or give security to certain touchy data with the goal that they can be shielded from unapproved gatherings or unauthorised. Though security is accomplished by concealing the touchy or private information, it will influence the information mining calculations in information extraction, so a compelling strategy or methodology is required to give security to the information and at the same time ensuring the nature of information mining calculations. Rather than expelling or scrambling touchy or private information, we utilize information change methodologies that keep the factual, semantic and heuristic nature of information while securing the delicate or private information. In this paper, we considered the specialized possibility of acknowledging Privacy-Preserving Data Mining. In the proposed work, Correlation Based Transformation Strategy for Privacy Preserving Data Mining is utilized for ordinal information. We apply the technique on a few datasets to be specific soybean, Breast Cancer, Nursery dataset and Car dataset. We classify the final products applying the proposed system on both the first and the changed dataset and watch connection distinction, Information Entropy and Classification Accuracy with various AI calculations and Clustering Quality. As an improvement, the proposed work can be stretched out by utilization of vector stamping methods where these strategies help in expanding the productivity by maintaining a strategic distance from unapproved access to the data.

1. INTRODUCTION

Data Mining is extensively used in varied areas like financialdata analysis, retail industry, biological data analysisand many more. However, it has got its downsides. Oneof the key issues raised by data mining technology is nota business or technology one, but a social one. It is theprivacy of an individual or a company. Data Miningmakes it achievable to evaluate everyday businesstransactions and gather a considerable quantity of informationabout individuals buying habits and preferences. Many companies are making fortune aggregating petitepieces of information about people and putting scrapstogether to build a digital profile. Most of the times the information collected will be used to sell stuff, which isuseful. However, the information extracted can be usedfor privacy violating purposes. Agencies, hospitals andother organizations often need to publish micro datafor research and other purposes. However, the informationextracted can be used for privacy violatingpurposes. As explained in1 micro data is usually storedin the form of table where each row represents an individual.

Here the table has three types of attributes:

1.Identity attribute (To uniquely identify an individuallike name),

2. Quasi identifier (which includes demographicattributes),

3. Sensitive attributes (whichinclude confidential information like diseases).

Quasiidentifiers attributes may be merged with other publicdatabases to uniquely identify the individual andtheir sensitive data (Linking attack). Thus privacy isbecoming a critical issue which led to a new researchfield called Privacy Preserving Data Mining (PPDM)2.PPDM comes into picture in the situations like theone described above. PPDM helps to perform datamining efficiently while preserving the private dataor information about an individual or a company.Instead of hiding or encrypting, PPDM transformsthe sensitive data to some other form while preservingthe usefulness of the data. Many strategies have beenproposed for PPDM, one of such is Correlation BasedTransformation Strategy (CBTS) which is used onnumerical data. The datasets likewise contain ordinal and ostensible information; the need is to change over the ordinal and ostensible information to numerical information by saving the information utility, with the goal that the calculation can be applied proficiently. In this paper, we propose a CBTS which can be applied to ordinal qualities. We depict a system to change over both ordinal and ostensible information to numerical information on which the CBTS can be applied. We measure the Information Entropy estimations of both Original information and Transformed information and the outcomes are practically identical and furthermore we measure Cluster Misclassification Error and demonstrate the mistake is less in our methodology. The paper is sorted out as pursues: Section 2 depicts Related Work. Segment 3 clarifies Problem Definition. Design is introduced in Section 4. The result is explained in Section 5. We close this paper with future work in Section 6.

2. RELATED WORK

In3, the creators have utilized a system called altered information transitive strategy in which the delicate numerical information thing is to be secured by changing the first information thing. There is an examination between the changed information transitive method and the perturbative concealing procedures, for example, added substance clamour, adjusting and miniaturized scale conglomeration and exhibitions are investigated and results are drawn by closing with the acceptable outcomes utilizing the transitive systems.

In4 authors proposed a new approach which involvesin preserving sensitive information using fuzzy logic.Clustering is done, in which the original dataset i.e.numerical data is transformed into fuzzy data and thennoise is added to the numeric data using an S shape fuzzymembership function.The Clusters which are generated using the fuzzyfieddata is similar to the original cluster and privacy is alsoachieved.In5 proposed a system which makes use of a perturbativesystem where encryption technique is applied tosensitive data items. The information has to be changedto a considerable extent before it is made available to thepublic for safe guarding the confidentiality of the sensitiveinformation. The proposed data transformation techniqueprotects categorical sensitive data which is modifiedusing advanced data transformation technique includingcryptography technique which prevents sensitive itemsfrom public disclosure. This system gives greater resultswhile preventing sensitive data from unauthorized disclosureand should not affect the importance of the originalobjective of data mining.In6 and7, the authors have proposed distortionbased techniques to meet the privacy requirements.In the former randomized distortion technique isapplied only on confidential categorical attribute. Inlatter probabilistic distortion method is used on originaldata before using frequent item set mining on thedata.In8 and9, the authors have used correlation basedtechniques to achieve privacy in huge datasets. Inpaper10 authors proposed a work which concentrates onfinding an efficient solution for the classification problemover encrypted data in cloud. This work protects theprivacy of sensitive data of users query and data accesspatterns. A k-NN classifier is developed firstly on a realworlddataset for different parameters and the efficiencyis resolved.

Authors of11 proposed a new patient centric clinicaldecision support system, which is of a great help for aclinician complementary in diagnosing the risk of patient'sdisease without compromising its privacy. This methodportrays correlation by spatial proximity. It involves thefollowing methodologies which can handle categoricaland numerical variables.Authors of12 proposed various methods and possiblerisks by the method of Random Projection. Itdefines a number of reconstruction techniques over thedata.

In paper13 authors concentrate on decision treelearning, without accompanying loss of accuracy. Thismethod strives at preserving the privacy of data whichare partially lost. This deals with the production of a setof unreal datasets which can be obtained as a result ofconversion of original dataset. Such that, redesigningof original samples without the entire group of unrealdatasets becomes impossible. From these datasets thedecision tree is built precisely. And also this method iscongruent with that of the other approaches which preservethe privacy of sensitive data and thereby ensuringhigher protection of data.In paper14 the authors have proposed a method whichprovides an excellent spatial transformation method toprotect the privacy concerns in cloud computing andthis method also provides considerably good results withrespect to the communication cost.In15 presents an erratic system based chaotic signalgenerator. Due to the characteristics of chaotic signal,estimators find it hard to estimate original data sincethey work on noise Probability Distribution Function(PDF). The issue of maintaining data privacy whilepublishing is resolved. Data Perturbation level depends on trust on which the data is to be generated. Due tothe different levels of trusts or same levels of trust ofsame data, a problem on security of data arises and maycause estimation of accurate data copies by Linear LeastSquares Error (LLSE), which is an advance computationalalgorithm.

3. PROBLEM DEFINITION

Given large structured data constituting of sensitiveinformation of ordinal nature, the objective isto preserve privacy by transforming the ordinal datainto an equivalent numeric representation whileretaining the original statistical nature with minimalentropy.

4. ARCHITECTURE

Given a huge data containing ordinal sensitive information,our solution first converts the ordinal and nominal data tonumerical data and transforms the resultant numeric datain such a way that it retains the correlation structure amongthe data values preserving its usefulness and maintaining thelevel of privacy. The conversion of ordinal data is done bytaking input for each data value from the concerned user andthe conversion of nominal data is done by assigning randomnumbers to each nominal data value. The numeric attributesare retained. We consider a dataset containing mixture ofordinal, nominal and numerical data attributes, in whichmany attributes are private and sensitive. The dataset is subjectedto clustering method like Simple K Means to groupthe similar rows and classification algorithm like J48. Theobjective of this paper is to convert and transform the ordinalsensitive data such that the correctly classified instancesand the decision trees of original data and transformed dataare comparable.For the given dataset with numerical sensitive information,authors in paper16 proposed CBTS for numericaldata. Given a dataset comprising sensitive and private data,CBTS produces an outcome comprising of the subset ofvectors correlated to sensitive data and produces equivalentcomponents as substitutes. CBTS uses Pearson'scorrelation coefficient.

$$X^2 = \sum_{K=1}^{n} \frac{(O_k - E_k)^2}{E_k}$$

Ok - Observed frequency.
Ek - Expected frequency.

The subsets generated are subjectedto transformation strategies that tend to converge on theobtained similarity forming new components. Hence thecomponents obtained are a mathematical representationof the sensitive data and used instead of sensitive data fordata mining. Figure 1 gives the Architecture of CBTS forNumerical data.Existing transformation methods PCA, SVD andNNMF have been used prior in PPDM by17–19 demonstratethe required property of convergence. The method was ableto remove the highly correlated sensitive data and transformthe non-correlated sensitive data. CBTS is appliedto datasets which has numerical values, the informationentropy values are compared for the original data and thetransformed data and the results are obtained. Thoroughexperiment analysis proved the proposed dataset transformationmethod has low clustering misplacement error andminimal deviation in classifier accuracies.In this paper we are extending CBTS to support ordinaldata. The proposed architecture is shown in Figure 2.Our method first converts both ordinal and nominal datato equivalent numerical data. The conversion step has two sub-steps. Initially, thedataset is parsed to extract the unique data values in eachcolumn which is given to next step. In the next step, basedon the type of the data values of the column, conversion isdone. When the column has ordinal data values, they areconverted to numerical values based on the user providedordering.In this work we have assumed all the nominal data tohave some ordinal nature. Nominal data are substituted byunsupervised statistical methods. Correlation coefficient iscalculated for the respective values against the data vectors.If there exists a strong correlation, then they are convertedto random numbers. If the correlation is weak, then theconversion is done by substituting categories with closeranged numbers to avoid and minimize error bias. Chisquaredtest is done to determine the correlation betweennominal data values.

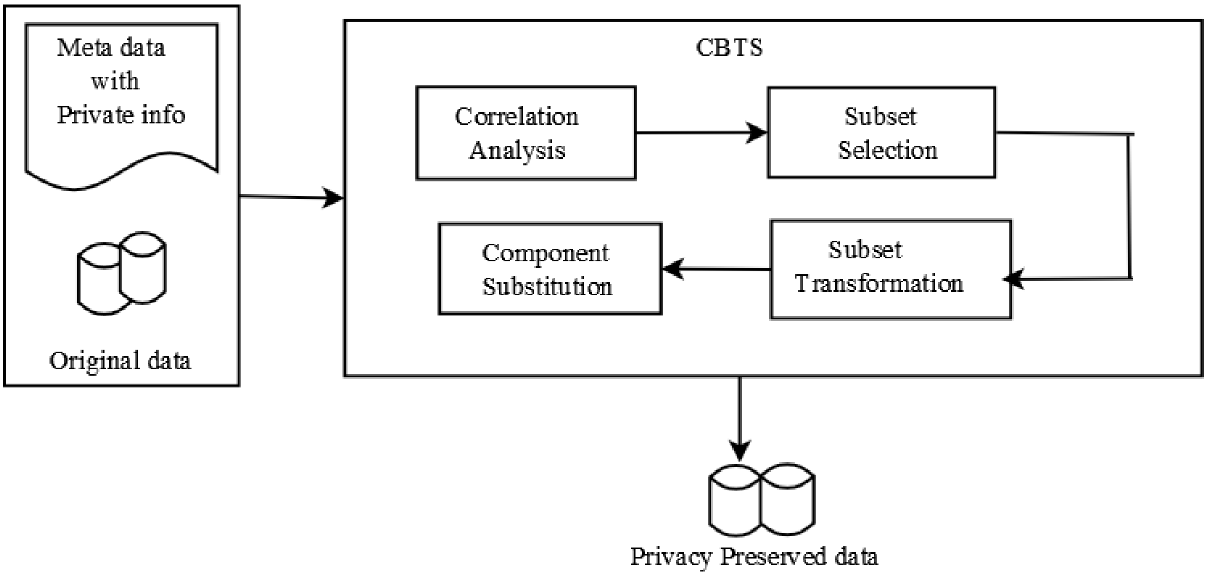

Figure 1. Architecture of CBTS for numerical data.

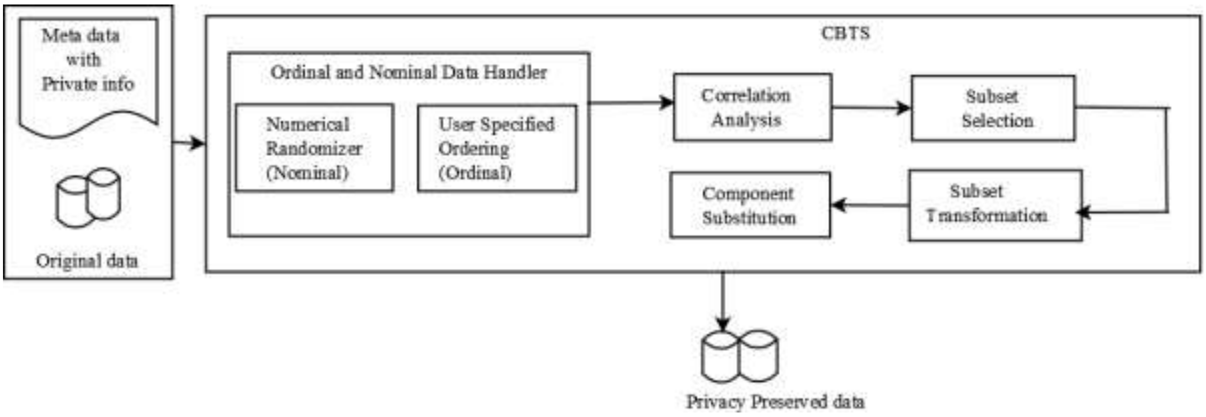

Figure 2. Architecture of CBTS for ordinal and nominal data.

(IJIEST) 2018, Vol. No. 4, Jan-Dec e-ISSN: 2454-9584; p-ISSN: 2454-8111

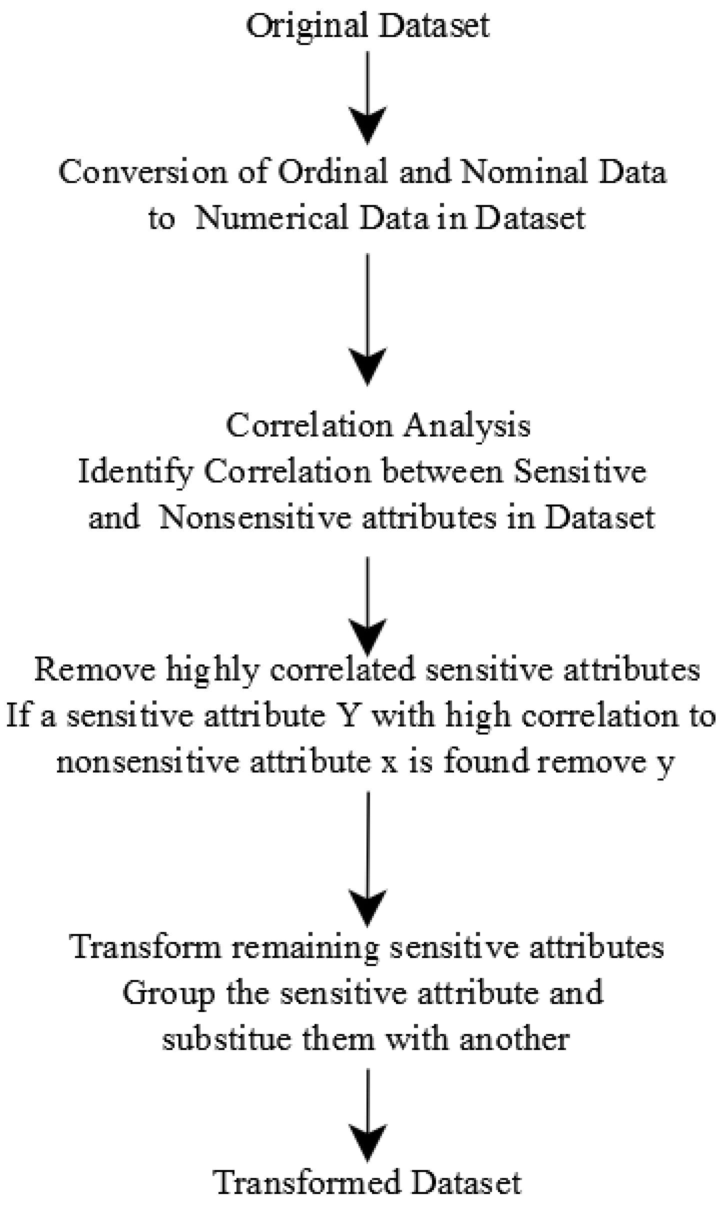

Figure 3. Transformation method.

5. RESULT

The datasets used in this paper are Soybean and BreastCancer. Both the datasets are taken from UCI MachineLearning Repository. Soybean dataset is a dataset with307 instances and 35 attributes. Among 35 attributessome are ordinal and some are nominal. Breast Canceris another dataset with ordinal, nominal and numericalattributes. There are 286 instances and 10 attributes in thisdataset. This dataset contains two classes and among 286instances, 201 belong to one class and the other 85 belongto another class.Data Entropy of unique information against bothered information utilizing CBTS for Ordinal information with change techniques is condensed in Table 1. We can gather from the table that deviation in Information Entropy is least utilizing the proposed CBTS strategy against utilizing change strategies alone. Table 2 gives the examination of classifier correctness's for different AI calculations utilizing CBTS against unique information. It is plainly detectable from the outcomes the classifier execution is similar to the first information. Table 3 shows the Misclassification Error ME esteems with k-implies grouping.

Table 1. Comparison of information entropy

Types of Data	Original Entropy	Information Entropy(I_E) Using CBTS Method/Using existing Methods		
		PCA	SVD	NNMF
Soybean (683x36)	3.317	3.30/10.25	3.41/5.12	3.25/9.26
Car (1729x6)	2.31	2.28/5.16	2.37/2.58	2.22/9.14
Nursery Dataset (12960x7)	1.88	1.88/4.6	1.95/2.8	1.86/11.0
Breast Cancer (286x9)	3.02	3.7/6.39	3.5/3.8	3.39/8.04

6. CONCLUSION AND FUTURE WORK

CBTS achieves accountable privacy by applying correlationtransformation based methods. CBTS hasapplications over varied areas involving huge data.Joined with the CBTS we have introduced a method for change by changing over delicate ordinal and ostensible information to numerical information of a considered dataset all the while saving the protection and the information utility of the equivalent. The proposed work can be stretched out by utilization of vector checking systems where these methods help in expanding the productivity by maintaining a strategic distance from unapproved access to the data.

Table 2. Comparison of various machine learning algorithms using CBTS (ME)

Types of Data	Machine Learning Algorithms	Observed Classifier Accuracy (%)				
		Ordinal and Nominal Data	Numerical Data	Transformation using CBTS		
				PCA	SVD	NNMF
Soybean (683x36)	Decision Tree	97.0	96.3	97.6	97.0	97.0
	Multilayer Perceptron	99.8	93.3	94.8	95.0	95.0
	Naïve Bayes	93.7	82.1	82.5	81.8	81.8
Breast Cancer(286x9)	Decision Tree	81.4	81.4	81.4	81.4	81.4
	Multilayer Perceptron	84.6	84.6	84.2	84.6	84.6
	Naïve Bayes	73.4	73.4	73.4	73.4	73.4

Table 3. Cluster misclassification error (ME)

Types of Data	Clusters (k)	M_E(with CBTS)			M_E(without CBTS)		
		PCA	SVD	NNMF	PCA	SVD	NNMF
Soybean(683x36)	2	0.253	0.455	0.248	0.999	0.999	1.0
	3	1.22	0.88	0.74	1.09	2.6	0.9
Breast Cancer (286 x 9)	2	0.017	0.7	0.7	1.3	1.60	1.50
	3	0.7	0.74	0.74	0.5	1.91	1.54