# Privacy Preserving for Feature Selection in Data Mining Using Centralized Network

Hemanta Kumar Bhuyan[1], Maitri Mohanty[2], Smruti Rekha Das[3]

[1,2,3]Dept of CSE, Mahavir Institute of Engineering & Technology, Biju Patanaik University of Technology

Bhubaneswar, Odisha, India

**Abstract**

This paper proposed a feature selection with privacy preservation in centralized network. Data can be preserved for privacy by perturbation technique as alias name. In centralized data evaluation, it makes data classification and feature selection for data mining decision model which make the structural information of model in this paper. The application of gain ratio technique for better performance of feature selection has taken to perform the centralized computational task. All features don't need to preserve the privacy for confidential data for best model. The chi-square testing has taken for the classification of data by centralized data mining model using own processing unit. The alias data model for privacy preserving data mining has taken to develop the data mining technique to make best model without violating the privacy individuals. The proposed process of data miner task has made best feature selection and two type experimental tests have taken in this paper.

***Keywords:*** *Privacy preserving, data mining, perturbed data, feature selection*

## 1. Introduction

Data mining technology has emerged to identify the patterns, domains of data from huge amount of database. The aim of the data mining is that to collect the data from multiple organizations and gathering it into data miner's database for creating a special data model, algorithm and estimate the best accurate result for model. On the above concept, it will be tested by own gathering data and also multiple organization data. The classification of the predictive modeling task which predicts the value of a univariable based on the known values of other variables. The value of attribute is also predicting the value of other attribute to make best class for data model.

Albeit some organizations collect a lot of data for their own and collect similar data about different people from several area (i.e. horizontal partitioned data), otherwise the organization collect the different information from same set of people (i.e., vertical partitioned data). In SMC (Secure multiparty computation) mechanism, different party can perform on global computation jointly on their private data without any loss of data privacy. SMC mechanisms create a type of environment where large amount of work can be done and also provide the secure joint computations among mutually distributed entities. Yao presented the initial concept of SMC in the form of "Two party computations" [1]. Later, it was generalized to multiparty computation problems by Goldreich who made SMC in the form of secure solutions for any functionality [2].

Generally, the definitions of SMC exist for two adversary model: (1).Semi-honest. (2).Malicious. In the semi-honest, each party follows the protocol and they may attempt to compute additional information from the messages received during execution after completion of the protocol. In the malicious model, a party can diverge arbitrarily from normal execution of the protocol for privacy preserving information sharing. The honest or semi-honest parties are described in [3, 4, 5, 6]. There have been studied that tried to remove the semi-honest assumption [7, 8, 9].

The statistical analysis can help to create many models for privacy preservation of organization data i.e., tools, techniques, algorithms etc. As statistical analysis, the feature values can be checked for decision model. Hypothetical testing is also used for statistical analysis. It is not necessary to preserve all features for privacy purpose. As requirement of model, the feature can be preserved for privacy. This paper helps to create the best product model by using data mining model. The data mining researcher can access data only from perturbed data based by analyze or quarry process. Data mining researcher produce the results, rules or patterns after computational analysis of perturbed data.

The framework of centralized computation has taken for performing centralized data mining task, specially using gain ratio technique. Different party gives their data to center for getting the unique result, where nobody deviate their result. We consider the data collection by using alias name of each feature set and sub-feature values, where nobody know about the actual data and conversion data except data miner. The data miner task has been derived through some process for getting actual result. After finding the result, data miner sends their data to all

Copyright (c) 2012 International Journal of Computer Science Issues. All Rights Reserved.

participating parties. The experimental result value has already derived in experimental section.

In the next section, section 2, we survey the previous work that has been done related to this paper. In section 3, we create the problem statement for centralized data computation. In section 4, we describe the frame work of the model with privacy preservation of data. In section 5, we show the experimental result and in section 6, the analysis of the result for privacy preservation of data is described. Finally in section 7, we give our conclusion.

## 2. Related work

Privacy preserving data mining (ppdm) is regulated the confidential data towards social model where several database is distorted by attacking the adversary for individual identifiable information and commit to several agencies trustfully for global pattern [13]. The secure multiparty computation in a distributed system has been developed the secure protocols in the adversary model. The padded item-sets is more secure when collusion of item-sets available in database but for the strong privacy, it needed the level of security protocol in distributed data mining [14]. Both privacy and security are important for the participating parties in distributed data mining to discover the computational result correctly. For the classification problem, the privacy maintains the tight bound of constraint set [15]. The variation methods for approximate technique are used for graphical models with the help of probabilistic inference and approximate inference [16]. The distributed probabilistic inference has also extended work of variation approximation technique in sensor network. It compares this technique for exact and centralized local algorithm framework to develop about the communicational efficient with variation mean field approach [17]. The local distributed algorithm for multivariate regression like distributed inference, data compression, data modeling, data prediction, classification has developed as a powerful statistical and machine learning tool [18] with a low monitoring cost. The set of facilities open for clients to serve at a low cost that depend on peer's data for facility location problem [19].

Microdata(i.e., individual records or data vectors ) are grouped into small aggregates prior to publication [10,11,12]. The microdata set is a set of records or data vectors containing data of individual respondents who can be persons, companies etc. The individual data vector of a micro data set is stored in a micro data file. The privacy control methods for micro data sets may be used the substitution (perturbation) method. The values of each variable in the micro data set are perturbed and the perturbed values are used to replace the original values. The main idea of privacy control is to provide sufficient protection without seriously damaging the information contained in the original data.

The rationale behind micro data aggregation is that the confidential rules allow publication of microdata sets if the data vectors correspond to group or more individuals where no individual dominates the group. The Strict application of such confidential rules leads to replacing the individual values with computed on small aggregates (micro data aggregate) prior to publication. Micro data aggregation also creates the class distribution for different problems. For example, in a product evaluation data, the instances are grouped as micro data aggregation which will make the different kind of class distribution. Micro data aggregation makes a small group of feature values for special class distribution. Sensitive micro data group:- The group of individual confidential data which make the class distribution is called sensitive micro data . It is the minimum number of sensitive micro data for privacy preservation of data model.

## 3. Problem statement

Let n number of parties is participating in a centralized computation network. They trust only on trusted third party i.e., data miner who collect the data from different parties and provide the result of the computation only. The data miner classifies the data as data provider requirement by which each party satisfy the result of the computation. The question arise (1)" How to maintain the privacy for the parties data", when the data miner collect the data from each party and (2) how to find the sensitive and non-sensitive feature from data miner database.

## 4. Framework data mining model

Our model assumes that all party (data provider) sends their data for the same set of feature to data miner. The number of rows of several parties may or may not be same. The feature values of each instance may vary. All the n parties have their own datasets $D_1 \ldots\ldots D_n$ with same number of features name but the features value may be different. In some situation only a part of the feature set needs to be kept confidential. The confidential features of a feature set are known as sensitive feature and remaining feature can treat as usually for any computational treatment. When the individual party is not willing to disclose its value or an advisory can never know about the value of that feature then that feature is recognized as sensitive feature. All party want to jointly conduct data mining operation on a single database D which is formed by the union of all data sets $D_1 \ldots\ldots D_n$ to get better results.

Copyright (c) 2012 International Journal of Computer Science Issues. All Rights Reserved.

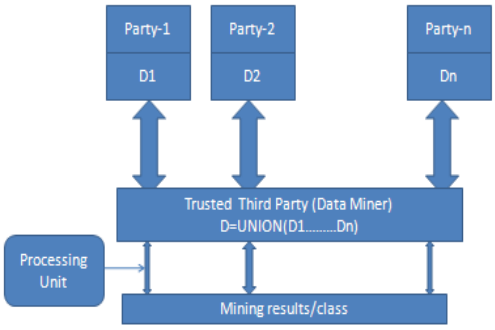


Fig -1:. The framework for centralized data mining model

In the above figure data miner collect all data from different parties and gathering the collecting data at data miner database for evaluation. The processing unit makes the computation for unique result which will be sent to all participating parties. The following section has derived for better performance of data mining model.

4.1 Data perturbation

The data miner works on perturbed data which can perform data mining task as if it works on the original data .But it can never interpret the result or class about the data. It can declare only the results to all participating party in the data sharing. Since all party only knows the result of their computation on their data, then it preserves their privacy on their data. The data miner collects the data with the help of alias name technique for privacy preservation of individual data.

4.2 Alias Name for Perturbed Data

Alias name is used in perturbed technique which is preserved by individual sites. So the converted categories data values contain only the alias name which is helpful to analyze the perturbed data with the third party i.e., data miner who cannot interpret any actual values. It introduces the noise of data means the original data converted to any other form of actual data or actual database is transformed into modified (perturbed) database. The perturbed data can be constructed by the data miner or data provider. The researchers make the method, techniques, and tools by analyzing the perturbed data, not original data. In table-1, the data has been shown both original data and transferred data (alias data).

Table-1: For Perturbed data

| S. N. | Original Feature set | Alias feature set | Original sub-feature value | Transferred data |
|---|---|---|---|---|
| **1** | Buying | A | v-high | CE1 |
| | | | high | CE2 |
| | | | med | CE3 |
| | | | low | CE4 |
| **2** | Maintainance | B | v-high | CE5 |
| | | | high | CE6 |
| | | | med | CE7 |
| | | | low | CE8 |
| **3** | Doors | C | 2 | CE9 |
| | | | 3 | CE10 |
| | | | 4 | CE11 |
| | | | 5-more | CE12 |
| **4** | Person | D | 2 | CE13 |
| | | | 4 | CE14 |
| | | | more | CE15 |
| **5** | .Lug-boots | E | Small | CE16 |
| | | | Med | CE17 |
| | | | Big | CE18 |
| **6** | Safety | F | Low | CE19 |
| | | | Med | CE20 |
| | | | high | CE21 |

Each feature set and its sub-feature values have derived with both alias feature set and alias sub-feature values. The fig.-2 shows the original data, perturbed data which was useful for research work. The original and perturbed database are both maintained by the system

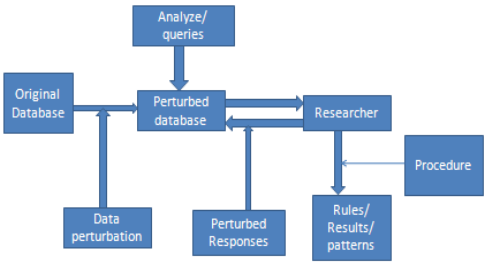


Fig-2: Perturbed Data Model

Copyright (c) 2012 International Journal of Computer Science Issues. All Rights Reserved.

4.3. Feature selection using Gain ratio

Let D is the Dataset and C is the class set of database. When the whole dataset divide into number of partition to recognize the unique identifier feature value then the information required to classify data set D based on this partitioning would be info (D) = 0. Since, information gain on this feature will be maximum on each pure partition then there is no required of classification by such partitioning.

The gain ratio which attempts to avoid this problem by using split information which is defined below. The split information of a feature X is recognized as splitinfo$_X$ which is derived by

$$\text{Splitinfo}_X(D) = -\sum_{i=1}^{p} \frac{|Di|}{|D|} \log_2\left(\frac{|Di|}{|D|}\right) \quad (1)$$

where $D_j$ is the jth partition of D which belong to feature X. The sum of each outcome of number of tuples in different partition is with respect to total number of tuples in D. It distinguishes the measures of classification information from information gain based on same partitioning. The gain ratio is defined as below.

$$\text{Gainratio}(X) = \frac{\boldsymbol{Gain\ (X)}}{\boldsymbol{splitinfo\ (X)}} \quad (2)$$

The splitinfo(X) has already given above, but the gain(X) is derived as below. Let $D_i$ are the partition data of D and $C_j$ are the distinct classes. Let $P_i$ is the probability of arbitrary tuples in D which belongs to class $C_j$. The expected information of D for several classes is defined by

$$\text{Info}(D) = -\sum_{i=1}^{all\ classes} P_i \log_2(P_i) \quad (3)$$

Here info (D) is the expected information of D to identify the class label. But the exact classification is measured by individual feature after partitioning of data as

$$\text{Info}_X(D) = \sum_{i=1}^{No.of\ partitions} \frac{|Di|}{|D|} * info(Di). \quad (4)$$

Where $\frac{|Di|}{|D|}$ is the weight of the $i^{th}$ partition and Info$_x$(D) is the expected information required to classify the tuples. The information gain is defined as the difference between info (D) and info $_X$(D) ie

$$\text{Gain}(X) = \text{info}(D) - \text{info}_x(Di) \quad (5)$$

It chooses the best classification for feature selection but it has some drawback which is explain in gain ratio. The feature having maximum gain ratio is selected as the splitting feature. But the split information should not be zero, because the ratio becomes unstable. So when the parties are putted their feature data, it may be violated the split information when the gain ratio is measured. The constraints for this measurement is that the information gain of the test selected must be large, i.e., at least as great as the average gain over all tests examined. So the splitinfo(X) should always greater than Gain (X). The privacy purpose it should maintain by each peer which are participating in the network. The different colluding parties may violate their measurement to select best feature.

4.4. Process of Task of Data Miner

1. Collect the instances from n parties with alias name.
2. Convert alias data to original data
3. As the class distribution, the instances are sent to individual class.
4. In each class, the #instances are checked as its own condition.
5. Data miner make own database with both feature set and classes.
6. Apply the gain ratio technique for feature selection
7. Ordering the gain ratio values for best feature selection.
8. Maintain the pure privacy for best feature data set.
9. Send best feature result to all participating party.

4.5 Statistical data evaluation for sensitive features

In database, we consider two type of feature set (1) sensitive features (2) non-sensitive feature sets. Both feature sets are evaluated by using chi-square statistical technique which is analysis in experimental section. The evaluation takes the item as (1) Class category, (2) observation value, (3) Expected value, (4) difference between them, (5) square the difference, (6) division of the square of difference by expected value. Also it uses the hypothetical test for level of significance of acceptance which is support the theory with experimental results.

## 5. Experiments

The data miner's job is to perform the union operation on the various sensitive attribute and alone can be used for only data mining task. We are taking the experiment on real data set and used the car evaluation database from UCI machine learning repository with 1728 records and 6 features which are recognized the class distribution as unacceptable, acceptable, good or very good car. The features are buying, maintenance; doors, persons, lugg-boot and safety are taken for analysis. The same features are received from different sites. We have considered two type of experimental test (1) the different sensitive feature for different class distribution using chi-square test. The sensitive feature preserved the privacy maintained by the data miner where any party cannot know about the data even own original data in data

Copyright (c) 2012 International Journal of Computer Science Issues. All Rights Reserved.

miner. (2) Using gain ratio technique for best feature selection from the own database of data miner.

5.1 Experimental Test using Chi-Square Value

The proportion of car evaluation instances in the four groups is 18: 6: 1: 1. In the experiment among 1728 instances, the numbers in the four groups were 1210, 384, 69, 65. We have to find out the expected values are
Unacc= (18/26) * 1728 =1196.3
Acc = (6/26) * 1728 = 398.7
Good = (1/26) * 1728 = 66.5
Vgood = (1/26) * 1728 =66.5

But observed values are unacc=1210, acc=384, good=69, vgood=65. The evaluation of chi-square is as follows.

Table 2: Chi-square valued.

| S. N. | Category | Observation value(O) | Expected value (E) | O - E | $(O-E)^2$ | $(O-E)^2/E$ |
|---|---|---|---|---|---|---|
| 1 | Unacc | 1210 | 1196.3 | 14 | 196 | 0.1638 |
| 2 | Acc | 384 | 398.7 | -15 | 225 | 0.5639 |
| 3 | Good | 69 | 66.5 | 2 | 4 | 0.0597 |
| 4 | Vgood | 65 | 66.5 | -1 | 1 | 0.0151 |
| 5 | Total | 1728 | 1728 | 0 | 426 | 1.34 |

So d.f =4-1=3 and tabulated $\chi^2_{0.05}$ for 3 d.f= 7.81. Since calculated value of $\chi^2$ is less than tabulated value, it is not significant. Hence null hypothesis may be accepted at 5% level of significance and we may conclude that the experimental results support the theory.

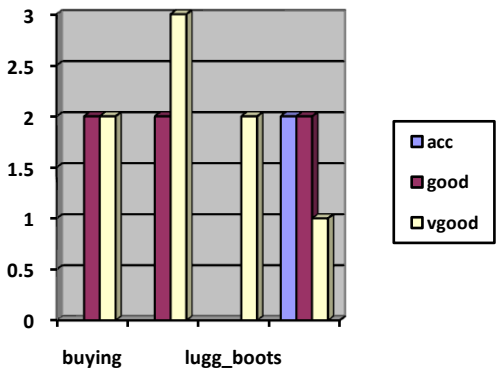


Fig- 3: Sensitive feature values of acc, good and vgood class

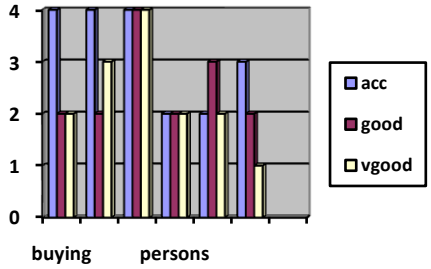


Fig- 4. Non-Sensitive feature values of acc, good and vgood class

The above fig -3 and fig-4 has shown for sensitive and non-sensitive feature sets using three classes.

5.2. Experimental result using gain ratio

In this experiment we have taken only three classes (i.e., acc, vgood, and good) and six feature sets (i.e, buying, maintenance, doors, persons, lug-boot, safety) with individual sub-feature values. We didn't consider unacceptability class because this class can neither be used for privacy preservation nor getting better production of this type of car. So we had taken 518 of data records which are involved with above three classes. The experimental results are derived in table-3 where $inf_x(D)$ is information of particular feature x (where x is vary for several features) and D is the data miner database from where data miner develop the model using several data. $G_x(D)$ is gain value, $SI_x(D)$ is splitinformation of D and $GR_x(D)$ is gain ratio. We have evaluated info(D) is 1.083.

Table-3: Feature evaluation using gain ratio

| Feature set(X) | $Info_X(D)$ | $G_X(D)$ | $SI_X(D)$ | $GR_X(D)$ |
|---|---|---|---|---|
| Buying | 0.871 | 0.211 | 1.921 | 0.109 |
| Maint | 0.928 | 0.505 | 1.932 | 0.080 |
| Doors | 0.578 | 0.154 | 1.990 | 0.253 |
| Persons | 1.082 | 0.0004 | 0.999 | 0.0004 |
| L-boot | 1.026 | 0.057 | 1.552 | 0.036 |
| Safety | 0.957 | 0.125 | 0.996 | 0.126 |

After feature evaluation using gain ratio we have ordered the feature set from the above table-3 as (1) Doors (2) Safety (3) Buying (4) Maintenance (5) Lug-boot (6) Persons. So we conclude the feature 'Doors' is the best feature among feature set. The experimental result has generated using Matlab-7.0. The figure-5 shows the ordering of feature set.

Copyright (c) 2012 International Journal of Computer Science Issues. All Rights Reserved.

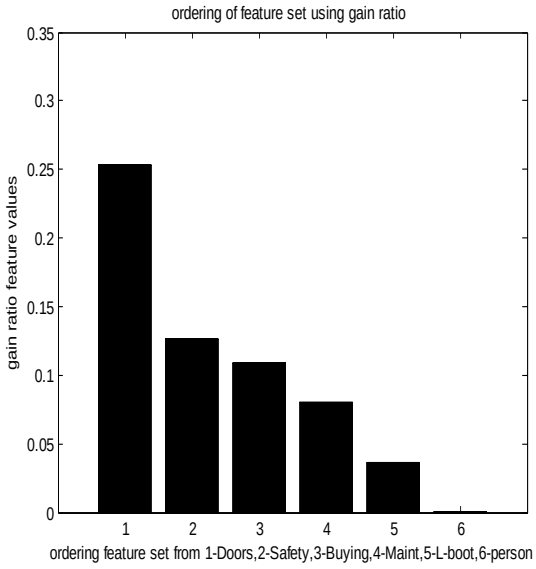


Fig-5: Ordering of feature set using gain ratio.

## 6. Privacy preserving analysis

The sensitive features preserved the privacy maintained by the data miner where any party cannot know about the data even own original data in data miner. It is not required to preserve the privacy for all features, but which are necessary that should be preserved for privacy. Since car class distribution are divided into 4 classes as unacceptability, acceptability, good and very good, it is not necessary to preserve the privacy for unacceptability class distribution because all type of feature values are involved in it. But in case of acceptable, good, very good class, it needs to preserve the privacy. The following features are preserved for privacy analysis purpose.

(1)Acceptability: -- (a) Persons, (b) Safety (2). Good: (a) Buying, (b) Maint, (c) Persons, (d) Safety (3)Very Good (a) Buying,( b).Maint, (c) Persons,(d)Lugg-boot, (e) Safety. The common sensitive feature is persons with feature values 4 and more .The other sensitive feature of each class is follows:

1. Acceptibility-1.Only safety, 2.good-1.Buying 2.Maintainace 3.Safety, 3.vgood-1.Buying, 2.Maint 3.Lug-boot 4.Safety

We have considered three cases of privacy of computation (1) Since we have considered the perturbed data collection from each party by alias name then it automatically preserve the privacy for all data (2) privacy for sensitive feature which are involved in chi-square test. (3) Privacy preservation for the data set of feature 'doors' by using gain ratio technique for best feature selection.

## 7. Conclusion

In centralized data mining model, the data miner can find huge amount of data for making classification model. The classification of individual instances usually preserves more information. The data mining processing work can derive the feature selection using gain ratio technique for best feature as framework. The ordering of feature set has made from data mining framework. The car evaluation model preserves the few feature of each class for better result of computation. The Microaggregation of data on the data set deal with each instances independently. The sensitive and non-sensitive feature help to derive the classification of data model for privacy preservation of data at data miner. The cases have taken for privacy preserving of data of this framework in this paper.

## References

[1] T.C. Yao, "How Generate and Exchange Secrets", In proceedings of the IEEE symposium on foundation of computer science IEEE, 1986 pages, 162-167.

[2] O.Goldreich, "Secure Multiparty computation", September 1998(woking draft online available on: http//www.wisdom.weizmann.ac.il/-oded/pp.html.

[3] R.Agrawal, A.Evfimievski, and R. Srikant, "Information Sharing Across Private Databases", Proc 22nd ACM SIGMOD Int'l Conf. Management of data 2003,pp.86-97.

[4] W.Jiag and C. Clifton,"Privacy Preserving distributed K-anonymity", Proc. 19th Ann. IFIP WG 11.3 Working Conf. Data and Applications Security 2005, pp. 166-177.

[5] M.Kantarcioglu and C.Clifton, "Privacy Preserving Distributed mining of Association Rules on Horizontally Partitioned Data", IEEE Trans. Knowledge and Data Engg. Vol.16, 2004, pp. 1026-1037.

[6] M.Kantarcioglu and J.Vaidya, "Privacy Preserving Naïve Bayes Classifier for Horizontally Partitioned Data", Proc. Workshop Privacy Preserving Data Mining 2003.

[7] M.J.Freedom, K.Nissim, and B.Pnkas, "Efficient private matcing and set intersection", Advances in Cryptography: Eurocrypt, 2004, pp 1-19.

[8] L.Kissner and D.X. Song, "Privacy Preserving Set Operations", Advances in Cryptography: Crypto. 2005, pp 241-257.

[9] Z.Yang, S. Zhang and R. N. Wright, "Anonymity-preserving Data Collection", Proc.11th ACM SIGKDD Int'l Con. Knowledge Discovery in Data Mining , 2005, pp.334-343.

Copyright (c) 2012 International Journal of Computer Science Issues. All Rights Reserved.

[10] L.Willenborg and T.D. Wall, "Statstical Disclosure Control in Practice", New York: Springer-Verlag 1996
[11] N.R.Adam and J.C.Wortmann, "Security- Control Methods for Statistical Databases: A Comparative Study", ACM Computing Surveys Vol 21, 1989, pp515-556.
[12] D.E. Denning, "Cryptography and Data Security", Reading Mass Addision-Wesley, 1982.
[13] Nabil R. Adam, John C. Wortmann. "Security-control Methods for Statistical Databases: A comparative Study", ACM Computing Surveys. Vol. 21, No. 4, December, 1989.
[14] Murat Kantarcıoˇglu and Chris Clifton. "Privacy preserving Distributed Mining of Association Rules on Horizontally Partitioned Data". IEEE, Transaction on Knowledge and Data Engineering, 16(9) 2004: 1026-1037.
[15] Jaideep Vaidya1 and Chris Clifton, S. A. Patterson. "Privacy- Preserving Decision Trees over Vertically Partitioned Data". ACMTKDD, 2(3):2008,1-27.
[16] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul. "An Introduction to Variational Methods for Graphical Models". *Machine Learning*, 37(2), Nov.1999, 183– 233.
[17] S. Mukherjee and H. Kargupta. "Distributed Probabilistic Inferencing in Sensor Networks using Variational Approximation". *J. Parallel Distrib. Comput.*, 68(1):2008, 78–92.
[18] K. Bhaduri and H. Kargupta. "An Scalable Local Algorithm for Distributed Multivariate Regression". *Statistical Analysis and Data Mining*, 1(3): Nov. 2008, 177–94.
[19] D. Krivitski, A. Schuster, and R. Wolff. "A Local Facility Location Algorithm for Large-Scale Distributed Systems". *Journal of Grid Computing*, 5(4), 2007,361–378.

**Hemanta Kumar Bhuyan** is a research scholar (PhD) in the Department of Computer Science and Engineering, Sikhya 'O' Anusandhan (SOA) University, Odisha, India. He obtained his M.Tech degree in Computer Science and Engineering from Utkal University, Odisha, India in 2005. He is currently working as an Assistant Professor in the department of computer science & engineering, Mahavir Institute of Engineering and Technology, Odisha, India. He has published many international and national journals and conference papers. His research interests include privacy preserving, distributed data mining and feature selection.

**Maitri Mohanty** has completed her BE in computer science from Biju Pattnaik university of Technology and M.Tech. from college of engg. and Technology(CET), Bhubaneswar, Biju Pattnaik University of Technology, Odisha,India. She is currently working as an Assistant Professor in the department of computer science & engineering, Mahavir Institute of Engineering and Technology, Odisha, India. She has published many international and national journals and conference papers. Her research interests include privacy preserving and distributed data mining, feature selection.

**Smruti Rekha Das** has completed Master of computer and Application (MCA) from Utkal University, Odisha, India. Her M.Tech degree in Computer Science and Engineering from Sikhya 'O' Anusandhan (SOA) University, Odisha,India. She is currently working as a Asst. Prof. in the department of computer science & engineering, Mahavir Institute of Engineering and Technology, Odisha, India. He has published many international and national journals and conference papers. Her research interests include privacy preserving, data mining and feature selection, clustering etc.

Copyright (c) 2012 International Journal of Computer Science Issues. All Rights Reserved.