ISSN 2277-3061

PRIVACY PRESERVING CLUSTERING BASED ON LINEAR APPROXIMATION OF FUNCTION

Rajesh Pasupuleti[1], Narsimha Gugulothu[2]
Department of CSE, Vasireddy Venkatadri Institute of Technology, Guntur, India
rajesh.pleti@gmail.com
Department of CSE, Jawaharlal Nehru Technological University, Hyderabad, India
narsimha06@gmail.com

ABSTRACT

Clustering analysis initiatives a new direction in data mining that has major impact in various domains including machine learning, pattern recognition, image processing, information retrieval and bioinformatics. Current clustering techniques address some of the requirements not adequately and failed in standardizing clustering algorithms to support for all real applications. Many clustering methods mostly depend on user specified parametric methods and initial seeds of clusters are randomly selected by user. In this paper, we proposed new clustering method based on linear approximation of function by getting over all idea of behavior knowledge of clustering function, then pick the initial seeds of clusters as the points on linear approximation line and perform clustering operations, unlike grouping data objects into clusters by using distance measures, similarity measures and statistical distributions in traditional clustering methods. We have shown experimental results as clusters based on linear approximation yields good results in practice with an example of business data are provided. It also explains privacy preserving clusters of sensitive data objects.

Indexing terms/Keywords

Clustering, Privacy preserving, Principal component analysis, Regression, Self organization mapping.

Academic Discipline And Sub-Disciplines

Computer Science and Engineering;

SUBJECT CLASSIFICATION

Data Mining; Information Security; Privacy

TYPE (METHOD/APPROACH)

Theory; Experimental Analysis

INTRODUCTION

Finding the underlying structure of relationship among objects is essential for data analysis. Clustering is the process of grouping set of data objects into classes which are having the similar characteristic behavior. Cluster analysis is a significant tool used to make decisions and finding exploratory pattern analysis in every human activity of large datasets and it has addressed a lot of consideration in the past literature. This knowledge is used either implicitly or explicitly in real time analysis. Clustering algorithms are mostly suitable for achieving data abstraction [1], [2] . A single clustering algorithm is not sufficient to solve all clustering problems because data items need to be clustered differently for diverse applications. Many clustering methods and algorithms have been proposed in different numerous applications such as data analysis, image processing, statistics, data mining, machine learning, pattern recognition, biology, marketing, business analysis, text mining, medical diagnosis, web analysis, CRM (customer relationship management) and information retrieval. Many clustering algorithms needs domain knowledge explicitly by passing user defined parameters, such as random selection of initial clusters, similarity measures, where distance is measured with respect to all available variables. Instead of grouping the objects which are having similar abstractions by their similarity measures, we proposed new cluster algorithm based on linear approximation of function by getting over all idea of behavior knowledge of clustering function. Linear approximation of clustering based approach will predict the continuous values of patterns from datasets efficiently. The future of data mining lies in predictive analytics.

Predictive analysis is a branch of statistical analysis that deals with extracting information from data and using it to predict future trends and behavior patterns [3]. Predictive analytics should be used to identify predict customer behavior and forecast product demand and related market dynamics. Estimated market for data mining is $500 millions. Data mining automates the process of discovering predictive information in a huge database. Data mining is an application of the mathematical system of statistics used for data extraction, classification, transformation, grouping, organization and analysis to identify patterns in order to make predictions. Regression models are the basis and most suitable for predictive analysis.

Regression analysis aims at predicting the value of unknown variable from the known value of other variable. It is widely used statistical tool in almost all branches of sciences. It is specially used in business and economics to study the functional relationship between two or more variables that are related. The estimation of future production, sales, return on investments, income, profits, prices, and consumption etc., are consider important to business analysis [4]. Correlation does not specify cause and effect relationship between the variables under study. However, regression analysis clearly specifies the cause and effect relationship between the variables. Regression coefficients byx, bxy are absolute measures representing the change in variable y, for a unit change in the value of variable x. Regression analysis has wider applications as it studies linear as well as non linear relationship between the variables.

The related work of clustering based on regression is explained as follows. H. Bertan Ari, Altay Guvenir [5] presented a new algorithm CLR (clustered linear regression) that improves the accuracy of linear regression by clustering training spaces of data sets to improve the accuracy of local linear regression. CLR can make good linear approximations only on large data sets. it can give accurate results for non linear regression functions. Vladik Kreinovich and Yeung Yam [6] have given a theoretical explanation on the choice of an optimal clustering method is important because different clustering methods lead to results of different quality, so it is extremely important to find the best clustering technique. Balazs Feil [7] introduced a clustering based approach is applied to the product space of input and output values of nonlinear variables in order to reduce the computational time for apply non linear models construction tools for the selection of the proper model order.

In this paper, we proposed a new approach privacy preserving clustering based on linear approximation of function. This approach holds two modules. Module 1 describes how to form clusters based on linear approximation of function. Module 2 explains privacy preserving of cluster objects of sensitive data. The outline of this paper structured in the following way.

We explained our proposed clustering algorithm description steps and pseudo code of algorithm in section 2. It also includes precisely defining clusters by choosing points on the linear approximation line as initial seeds and then performing cluster operations. Section 3, describes experimental results, how the proposed method works in practice with an example of business data are provided. In section 4, provides privacy preserving of cluster objects of sensitive data. Finally Section 5 consists of conclusion and future scope.

PROPOSED APPROACH

Clustering methods are an exploratory data analysis tool. Clustering is an unsupervised learning, so it is very essential to know about clustering approximation functional behavior instead of grouping data objects of similar characteristics into a cluster. Analyzing this kind of functional behavior of cluster objects will be suitable for any kind of data to make good clusters using linear approximation. Finding the linear group based structured clustering is the natural way assessment. In linear approximation output is directly propositional to input data objects. In linear approximation a family of linearly independent solutions can be used to construct general solution as superposition principle. If the data objects are non linear then we can transform non linear functions into linear. Once nonlinear data are transformed into a linear model, the result is generally biased. With the intention of explaining privacy preserving clustering based on linear approximation of function, we provided an example of sensitive company business dataset consist of dimensions as name of the company, Investments, net worth, rate of growth, year, reserves, total asserts, total liabilities.

The key idea is to form clusters of business data objects which companies are having similar rate of growth by the linear approximation based clustering method. Not only that we can consider any one of the dimension in company business dataset to form clustered (form the clusters by similar investments of companies, according to year wise and total asserts etc). In that scenario, simultaneous regression performs multiple regression over all input variables. Without loss of generality, we consider simultaneous regression in our proposed method instead of multiple linear regression. Each attribute or dimension (dependent variable) is predicted by the remaining variables (independent variables). By the method of clustering based on linear approximation of function, we can predict the behavior analysis of a company's outlook. Finally, scrutinizing the result analysis and providing privacy preserving for the sensitive values of business data objects.

```
Algorithm 1.    Clustering Based on Linear Approximation Function
Input : Instance huge set of multi dimensional data  objects(S).
Output: Form clusters.
repeat
(1)         Clusters ←Empty;
(2)         Apply PCA (principle component analysis) on S to reduce into two dimensional PCA axes;
(3)         Perform the simultaneous regression on available dimensions including PCA object  space;
(4)         Select the points on the straight as initial seeds for clustering;
(5)         Then perform SOM clustering method;
(6)         Result analysis;
Until making good clusters.
```

In proposed method, first we consider the multi dimensional data objects of various companies business dataset according to dimensional variables. Then by using principle component analysis (PCA), reduced the multi dimensional data objects into two dimensional PCA axes [8]. Represent all the data objects into a scatter plot of two dimensional PCA axes. After that constructs a linear approximation function line using regression technique. Add simultaneous regression to identify for a model of the association between a target variable and a set of other input variables (attributes or dimensions, features). Then choose the points on the linear approximation line as initial seeds for clustering analysis unlike grouping data objects of similar characteristics into clusters by taking random initial seeds as clusters, distance parameter as similarity measure. Then perform clustering by neural networks based algorithm using self organization mapping (SOM) to form clusters. A most popular neural network algorithm and efficient for clustering is the self-organizing map (SOM) [9], [10], [11]. Linear approximation based type clusters will present information about continuously increasing data objects among linear approximation of straight line.

Linear approximation is best suitable model for identifying the overall functional behavior of multi dimensional data objects of business dataset. Companies, which are in clusters of small, medium and large scaled are used to make decisions to increase services, productions, sales of company and cross examine the competitor growth rate. Clustering based on linear approximation will also helpful to make decisions to procure small companies by large scaled companies. Analyzing this kind of functional behavior of cluster objects will be suitable for any kind of data to make good clusters using linear approximation. The architecture of clustering based on linear approximation function is given below.

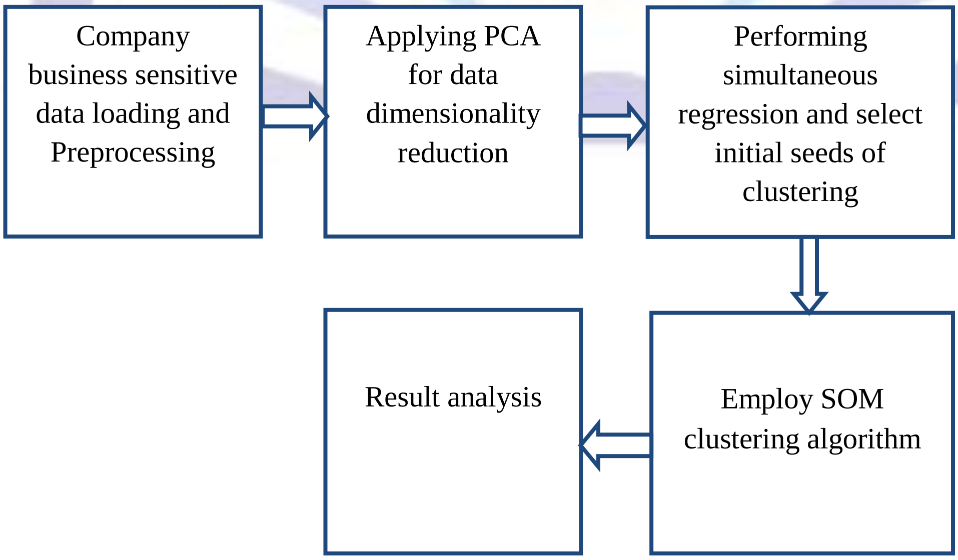
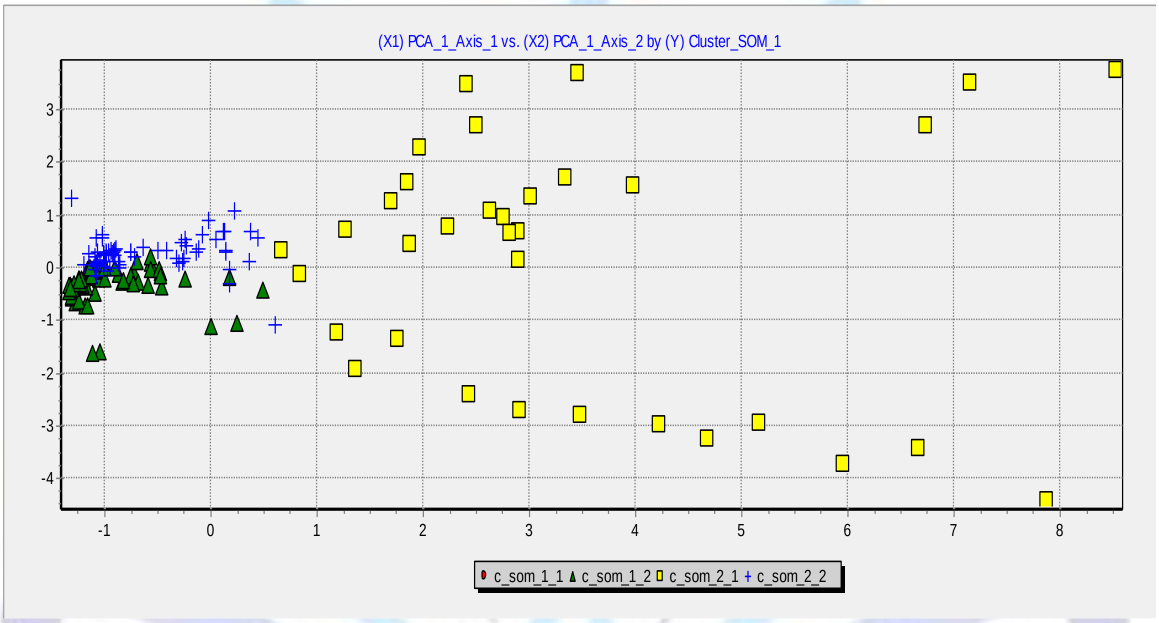

Company business sensitive data loading and Preprocessing
Applying PCA for data dimensionality reduction
Performing simultaneous regression and select initial seeds of clustering
Employ SOM clustering algorithm
Result analysis

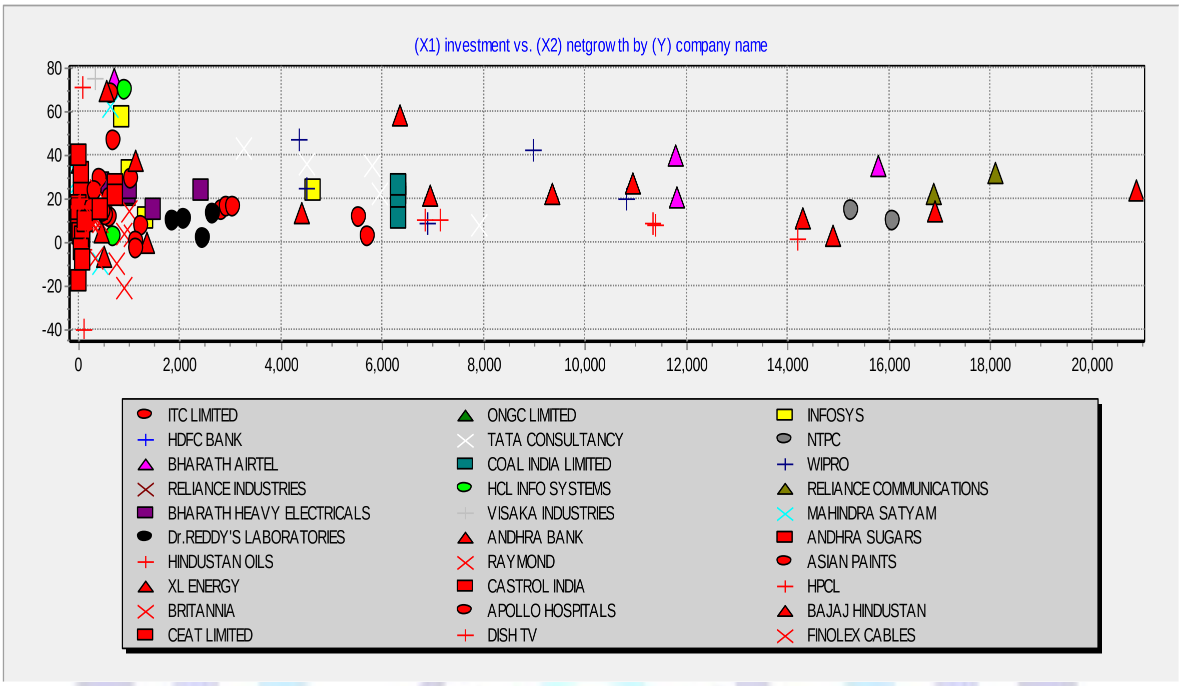
Fig1: System architecture for clustering by linear approximation

EXPERIMENTAL RESULTS

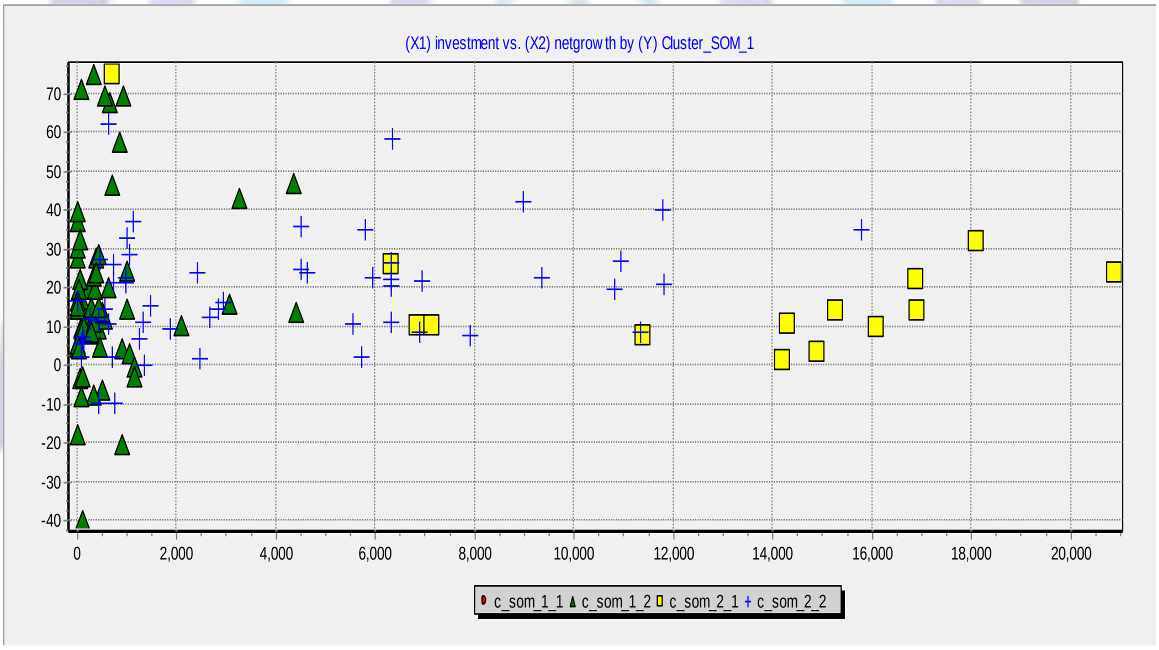
Data mining has grown to be a admired term among numerous business companies and organizations to facilitate analysis of data. We exploit TANAGRA tool for implementing privacy preserving clustering based on linear approximation of function. Tanagra is a data mining open source application tool allows analyzing and searching for interesting patterns in large data bases. It can be pedagogical tool for learning data mining process techniques. The data set describes 176 examples of business sensitive company's data [12]. With the intention of find unbiased estimate of performance of clustering, we divide the data set using sampling into two parts: 100 instances as training set and 76 instances for testing for the preparation phase. After that we applied the analysis on whole data set. We insert DEFINE STATUS component from tool bar to define target attributes and input attributes. Then we add Principal Component Analysis (PCA) from factor analysis tab. Principal Component Analysis (PCA) is an unsupervised technique that transforms a set of correlated variables into a new set of uncorrelated variables. Using principal component analysis, given information can be modeled by a linear or a nonlinear technique. The first four factors correspond to the 99.07% available information. Apply simultaneous regression component from regression tab. The R2 values of attributes investment, reserves, net worth, total asserts, total liabilities are 0.9061, 0.9965, 0.9966, 1.0, 1.0. R2 values of attributes indicates that 90% of variability in response variables is described by descriptive variables. Then selected the points on the regression line from scatter plot of data visualization tab as initial seeds for clustering. Once seeds are obtained, perform self organization map (SOM) clustering method. SOM is popularly used in statistical data mining, the learning process of SOM takes place by the winning neuron. We click on the VIEW menu in order to obtain the results. The subsequent figure shows clusters with respect to principle component axis.



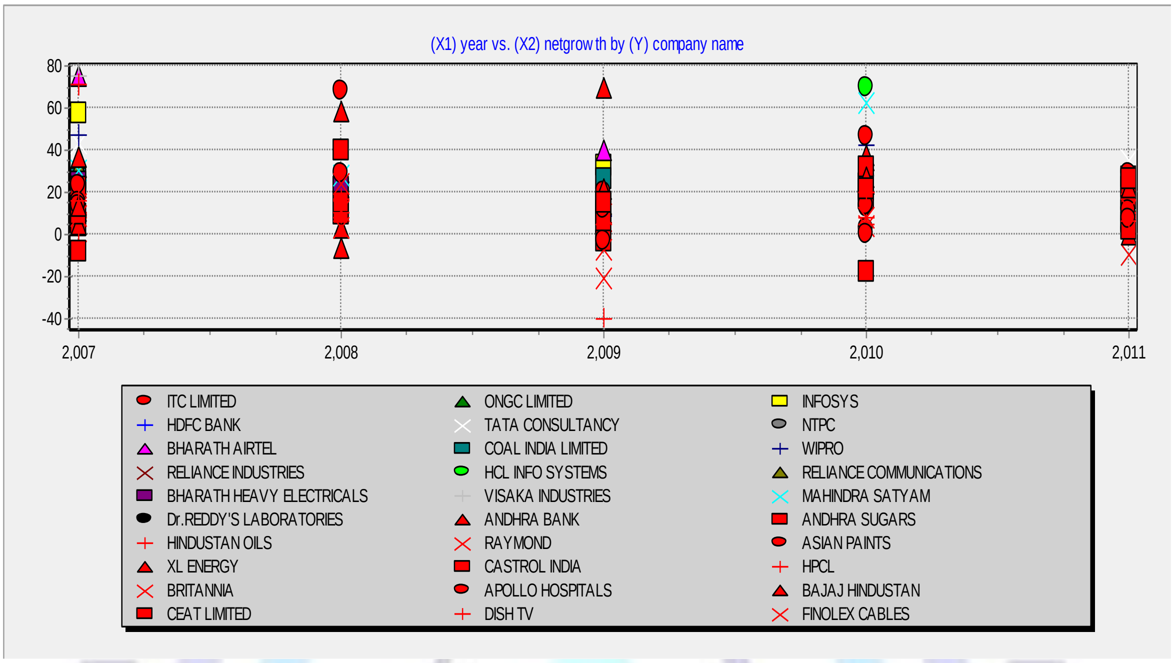
The following diagram describes about companies invested amount according to their net growth rate. Reliance communications and Andhra bank has put highest investments.



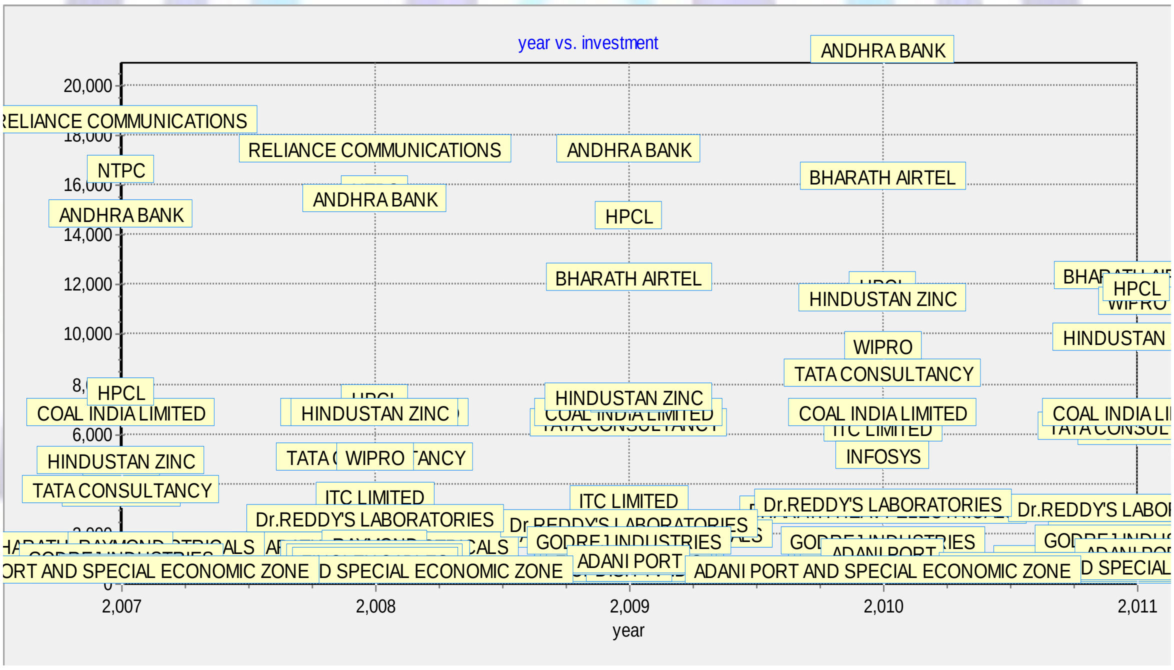
The following figure shows group of cluster companies which are having similar investments and their net growth.



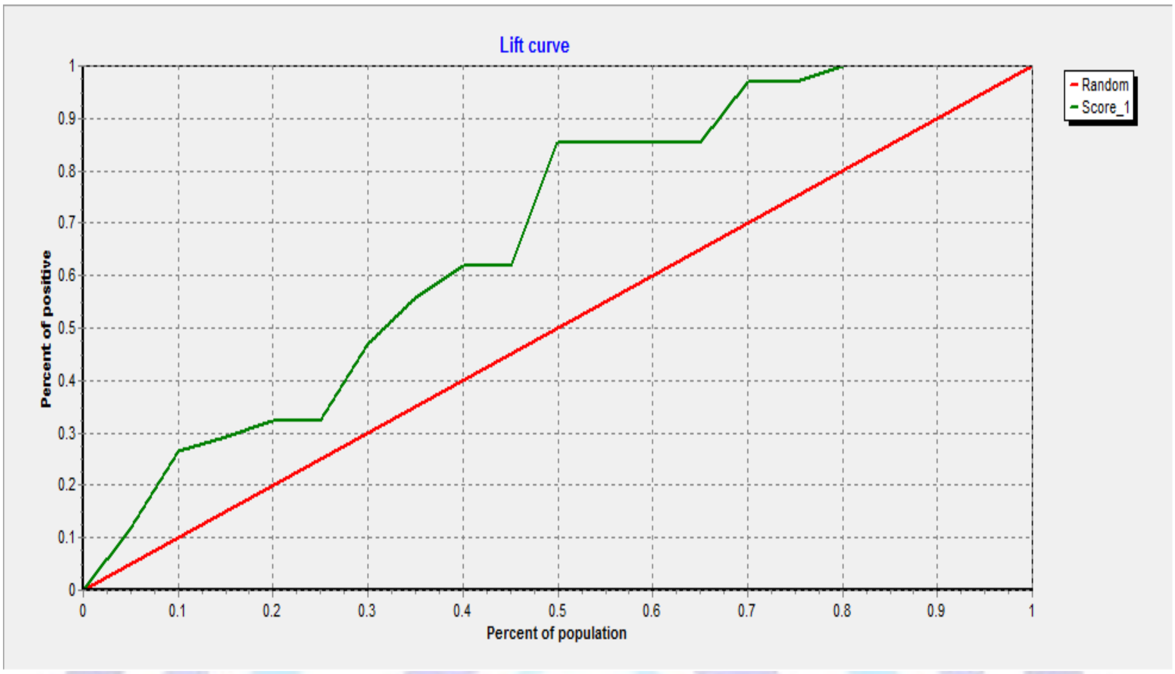
The subsequent diagram illustrate which companies having highest net growth with respected to year wise.



The next figure gives information about companies' investments according to year wise.



The supervised clustering method produces a good analytics rather than unsupervised clustering techniques. We tried to estimate the accuracy of unsupervised clustering by supervised linear discriminant analysis of scoring function. Lift curve shows that 50% of population of the data has 87.57% positive values (assigned data objects to their respective clusters).



PRIVACY PRESERVING OF SENSITIVE CLUSTER DATA OBJECTS

Nowadays information is becoming increasing important and in fact information is a key part in decision making in an organization. We are in the world of information era. Data is the major valuable resource of any enterprise. There is an incredible amount of sensitive data produced by day-to-day business operational applications [13], [14]. An attractive novel trend for data mining research is the development of techniques that integrates privacy concerns [15], [16], [17], [18]. Two current central categories to perform data mining tasks without compromising privacy are Perturbation method [19], [20], [21] and the secure computation method [22], [23], [24], [25]. By including any privacy preserving technique to data mining, the communication and computation cost will not be increased. In the privacy preserving clustering based on linear approximation function, we address only the perturbation of senstitive values of clustered data objects. We provided privacy preserving of sensitive information of data by replacing the original data objects to the object lie on linear approximation line nearest to it. So the real values of the data objects are ambiguous. This type of privacy preserving randomness approach of original data values, will work for data objects scattered thick among the linear approximation line is easier than the objects far away from line. Precisely privacy preserving clustering based linear approximation of function efficiently supports to continuously increasing data objects behavior among linear approximation of straight line

CONCLUSION AND FUTURE SCOPE

Every business or organization must needs data mining to increase their effectiveness and economical growth rate in the competitive business world. By employing data mining techniques in business effectively analyze the future knowledge, quickly make decisions and future predictions. In this paper, "clustering by linear approximation" we tried to find out continues, sophisticated extraction of analysis from company data sets which includes form of clusters of companies having similar rate of growth, similar investments, clusters according to year wise, profit sales of a company and total asserts .We also provided privacy preserving of sensitive information of data by replacing the original data objects to the object lie on linear approximation line nearest to it(randomness). Data mining has wide application domains in almost every industry. So that, these kind of technique "privacy preserving clustering based on linear approximation of function" can be applied to most promising interdisciplinary developments in Information Technology include retail stores, hospitals, banks, and insurance companies, Industries , business, real estates, agriculture, imports and exports, sales. Data mining is considered as one of the most important frontiers in database and information systems. Privacy preserving data mining field is expected to flourish.

ACKNOWLEDGMENTS

The authors would like to extend their gratitude to the anonymous reviewers and who continuously supported to bring out this technical paper. I would like to thank my advisor, Dr. A. Damodaram, Director of Academic Audit Cell, Jawaharlal

Nehru Technological University Hyderabad , for his professional assistance and remarkable insights in research. I would like to express my gratitude to Prof. G. Narsimha for helping me take the first steps in the research area. And finally, special thanks to my mother and father for providing the moral support to me. we are pleased to acknowledge our sincere thanks to our beloved chairman sri.V. Vidya Sagar and Director Prof. S.R.K. Paramahamsa for valuable cooperation in developing my career.

REFERENCES

[1] Usama, Fayyad, M. 1996. Data-Mining and Knowledge Discovery: Making Sense Out of Data. Microsoft Research IEEE Expert, 11, 5, 20-25.

[2] Han, J., Kamber, M. 2000. Data Mining: Concepts and Techniques Morgan Kaufman, New York.

[3] Weiss, Sholom, M. 1998. Predictive Data-Mining: A Practical Guide. San Francisco, Morgan Kaufmann.

[4] Giudici, P. 2003. Applied Data-Mining: Statistical Methods for Business and Industry West Sussex, England, John Wiley and Sons.

[5] Ari Bertan, Guvenir H. Altay. " Clustering linear regression", In the proceedings of ELSEVIER journal knowledge based systems, 15, 169-175, 2002.

[6] Kreinovich Vladik, and Yam Yeung. "Why clustering in function approximation? theoretical explanation", In the proceedings of International Journal of Intelligent System, 15, 10, 959-966, 2000 .

[7] Feil Balazs, Abonyi Janos, and Szeifert Ferenc. " Model Order Selection of Nonlinear Input-Output Models- A Clustering Based Approach", In the Proceeding of journal of process control, 14, 6, 593-602, 2004.

[8] Jolliffe, I.T. 1986. Principal Component Analysis. In proceedings of Springer, Verlag, New York.

[9] Kohonen, T. 2001. Self-Organizing Maps, Springer.

[10] Giraudel, J.L. , Lek, S. 2001. A comparison of self-organizing map algorithm and some conventional statistical methods for ecological community ordination, Ecol. Modelling, 146(1-3), 329-339.

[11] Vesanto, J. 1997. Using the SOM and local models in time-series prediction: In Proceedings of WSOM'97, Workshop on Self-Organizing Maps, Helsinki University of Technology, Neural Networks Research Centre, Espoo, Finland.

[12] Google, India Financial Portal. http://www.moneycontrol.com/stocksmarketsindia.

[13] European Union. Directive on privacy protection (October 1998).

[14] Saltelli, A. 2000. What is Sensitivity Analysis. John Wiley & Sons, Ltd., 3-13.

[15] Agrawal, R., Srikant, R. 2000. Privacy preserving data mining: In Proceedings. of the ACM SIGMOD Conference on Management of Data, Dallas, Texas, 439-450.

[16] Derosa, M. 2005. Data mining and Data analysis for counterterrorism, Center for Strategic and International studies.

[17] Agrawal, R., Srikant, R. , and Eufimieuski, A. 2003. Information Sharing across Private Databases, In Proceedings of ACMSIGMOD International Conference on Management of Data, 86-97.

[18] Dinur, I., Nissim, K. 2003. Revealing information while preserving privacy, In proceedings of the 22nd ACM SIGMOD-SIGACT-SIGART Symposium on principles of database systems, 202-210.

[19] Adam, N.R. , Wortman, J.C. 2006. Security-control methods for statistical databases, A comparative study on ACM Computing Surveys, 515-556.

[20] Agarwal, D., Aggarwal, C.C. 2001. In the Design and Quantification of Privacy Preserving Data Mining Algorithms, In Proceedings of the 20th ACM Symposium on Principles of Database Systems 247-255, Santa Barbara, California.

[21] Conway, R., Strip, D. 1976. Selective partial access to a database, In proceedings of ACM Annual conference, 85-89.

[22] AnandSharma, VibhaOjha. " Implementation of Cryptography for Privacy Preserving Data mining", In Proceedings of ITDMS, (2), 3, 2010.

[23] Canetti, R., Ishai, Y., Kumar, R., and Reiter, M.K. 2001. Selective Private Function Evaluation with Application to Private Statistics, In 20th PODC, 243-304.

[24] Canetti, R. "Security and Composition of multiparty Cryptographic Protocols", In Journal of Cryptology, 2000. 143-202.

[25] Lindell, Y., Pinkas, B. 2000. Privacy preserving data mining, In CRYPTO, 2000, 36- 54.

Author' biography with Photo

P.Rajesh received the M.Tech degree in computer science and engineering (CSE) from Jawaharlal Nehru Technological University Hyderabad in 2009. He is currently pursuing Ph.D degree in the department of computer science and engineering from Jawaharlal Nehru Technological University Hyderabad and working as an Assistant professor in CSE department at Vasireddy Venkatadri Institute of technology, Guntur, Andhra Pradesh. His research interests are in the area of Data mining, Information security, Privacy preserving data publishing and sharing.

Dr.G.Narshima received Ph.D degree from osmania university, Hyderabad. He is having thirteen years of teaching experience and having seven years of research experience in various prestigious institutions. He is currently working as an Associate professor in the department of computer science and engineering from Jawaharlal Nehru Technological University Hyderabad. He has enormous research and teaching learning experience in various prestigious universities. His research interests are in databases, data privacy, Data mining, Information security, Information networks, Mobile communications, Image processing, Privacy preserving data publishing and sharing.