IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 4, April 2016

A Privacy Preserving K-Nearest Neighbor Classification Protocol Over Secure Encrypted Data

Dr. V. Selvi[1], K. Rajalakshmi[1]

Department of Computer Science, Mother Teresa Women's University[1]

Abstract: Data mining has wide variety of real time application in many fields such as financial, telecommunication, biological, and among government agencies. Classification is the one of the main tasks in data mining. For the past few years, due to the increment in various privacy problems, many conceptual and feasible solutions to the classification problem have been proposed under different certainty prototype. With the increment of distributed computing users have an opportunity to offload the data and processing in the cloud, in an encrypted form. The data in the distributed are in encrypted form, existing privacy preserving classification systems are not relevant. To perform privacy preserving k-NN classification over encrypted data. The recommended protocol preserves the privacy of data, protect the user query, and hide the access mode.

Keywords: Security, K-NN Classifier, Outsourced databases, Encrypted,Privacy Preserving classification.

I. INTRODUCTION

Data mining involves the use of sophisticated data analysis tools to discover previously unknown, valid pattern and relationships in large data set. These tools include statistical models, mathematical algorithms, and machine learning methods. Consequently, data mining is the task of collecting and managing data which also includes analysis and prediction.

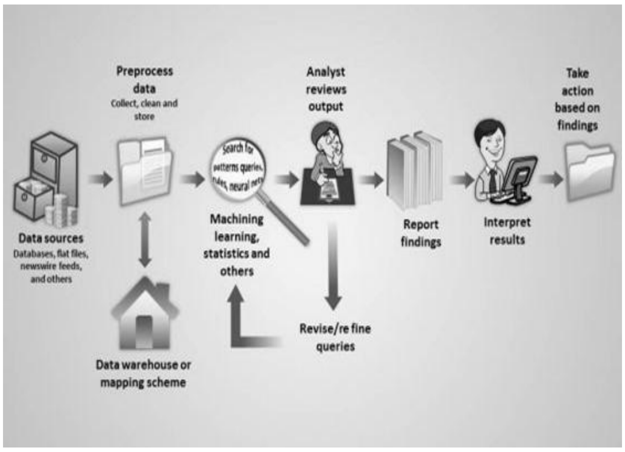
Preprocess data
Collect, clean and store
Analyst reviews output
Take action based on findings
Search for patterns queries, rules, neural nets
Data sources
Databases, flat files, newswire feeds, and others
Machining learning, statistics and others
Report findings
Interpret results
Revise/re fine queries
Data warehouse or mapping scheme

Figure 1 Data Mining Process

1.1 What is Data Mining

Data mining is a larger process known as knowledge discovery in databases (KDD). There are many other terms carrying a similar or slightly different meaning to data mining, such as knowledge mining from database, knowledge extraction, data pattern analysis. Data mining treat as synonym for another popularity used term, knowledge discovery in database, or KDD.

KDD has consisted the following steps to process it. There are as follows:

- Data Cleaning
- Data Integration,
- Data Selection
- Data Transformation
- Data Mining
- Pattern Evaluation
- Knowledge, Presentation

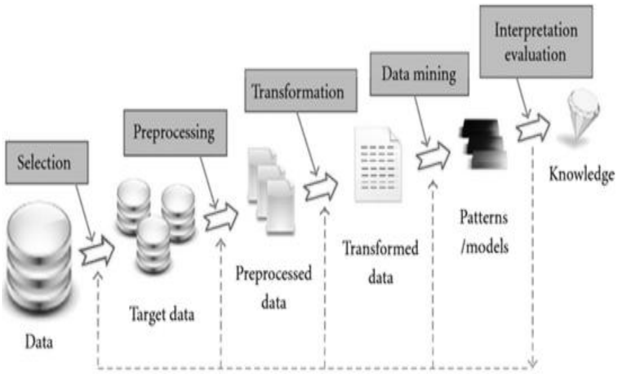

Figure 2 for KDD process

Data mining can be performed on data represented in quantitative, textual, or multimedia forms. Data mining application can use a variety of parameters to examine the data. They include association, sequence or path analysis, classification, clustering, and forecasting.

Classification of Data Mining Systems

The classifications of Data Mining Systems are as follows:

- Kinds of Database Mined
- Kinds of Knowledge Mined
- The kinds of techniques Utilized
- Applications Adapted

DOI 10.17148/IJARCCE.2016.54279

1.2 Why Use Data Mining

Data might be one of the most valuable assets of an organization, corporation – but if know how to reveal valuable knowledge hidden in raw data. Data mining allows extracting diamonds of knowledge from historical data and predicting outcomes situations. It will help to optimize business decisions, increase the value of each customer and communication, and improve the satisfaction of the customer with services.

Data analysis differs for companies in different industries. Examples include,

- Sales and contacts, histories
- Call support data
- Demographic data on customers and prospects
- Patient diagnoses and prescribed drugs data
- Click stream and transaction data from website

In all these cases, data mining can help to knowledge hidden in data and turn this knowledge into a crucial competitive advantage. Today increasingly more companies acknowledge the value of this new opportunity and turn to megaputer for leading edge data mining tools and solutions that help optimizing their operations.

1.3 Data Mining Applications

- Medicine – drug side effects, hospital cost analysis, genetic sequence analysis, prediction ect.
- Finance – stock market prediction, credit assessment, fraud detection etc.
- Marketing/sales – product analysis, buying patterns, sales prediction, target mailing, identifying 'unusual behavior' etc.
- Knowledge Acquisition
- Scientific discovery – superconductivity research, etc.
- Engineering – automotive diagnostic expert systems, fault detection ect.

1.4 Data Mining Functionalities

Data mining functionalities are used to specify the kind of patterns to be found in data mining tasks. In general, data mining tasks can be classified into two categories: Descriptive and Predictive. Descriptive mining tasks characterize the general properties of the data in the database. Predictive mining tasks perform inference on the current data in order to make predictions.

In some cases, there is no idea, which kinds of patterns in the databases may be interesting, and hence may like to search for several different kinds of patterns in parallel. Thus, it is important to have a data mining system that can mine multiple kinds of patterns to accommodate different user expectations or applications. Furthermore, data mining systems should be able to discover patterns at various granularities. Data mining functionalities, and the kinds of patterns can discover, are described here.

Concepts/Class Description: Characterization and Discrimination

Data can be associated with classes or concepts. It can be useful to describe individual classes and concepts in summarized, concise, and yet precise terms. Such descriptions. There are several methods for effective data summarization and characterization. The output data characterization can be presented in various forms, include pie charts, bar charts, multidimensional data cubes, and multidimensional tables, including cross tabs.

Data discrimination is a comparison of the general features of target class data objects with the general features of objects from one or set of contrasting classes. The user can specify the target and contrasting classes, and the corresponding data objects retrieved through database queries.

Association Analysis

Association analysis is the discovery of association rules showing attribute value conditions that occur frequently together in a given set or data. Association analysis is widely used for market basket or transaction data analysis.

Classification and Prediction

Classification is the process of finding a set of models that describe and distinguish data classes or concepts. The derived model is based on the analysis of a set or training data. Classification can be used for predicting the class label of data objects. However, in many application, users many wish to predict some missing or unavailable data values rather than class lables. This is usually the case when the predicted values are numerical data and is often specifically referred to as prediction.

Cluster Analysis

Unlike classification and prediction, which analyse class labeled data objects without consulting a known class model. In general, the class labels are not present in the training data simply because they are not known to begin with. Clustering can be used to generate such labels. The objects are dustered or group based on the principle of maximizing the interclass similarity.

A cluster has high similarity in comprarision to one another, but is very dissimilar to objects in other clusters. Each cluster that is formed can be viewed as a class of objects, from which rules can be derived. Clustering can also facilitate taxonomy formation, that is, the organization of observation into a hierarchy of classes that group simimar event6s together.

Outlier Analysis

A database may obtain data objects that do not comply with the general behavior or model of the data. These data objects are outliers. Most data mining methods discard outlier as noise or exception. However, in some applications such as fraud detections, the rare events can be more interesting than the more regularly occurring ones. The analysis of outlier data is referred to as outlier mining.

Outlier can be caused by measurement or esxecution error. The outlier themselves may be of particular interest, such as in the case of fraud detection, where outliers may indicate fraudulent activity. Thus, outlier detection and analysis is an interesting data mining task, referred to as outlier mining.

DOI 10.17148/IJARCCE.2016.54279

Evolution Analysis
Data evolution analysis describes and models regularities nor trends for objects whose behavior changes over time. Although this may include characterization, classification, discrimination, or clustering of time – related data, distinct features of such an analysis include time – series data analysis, sequence or periocity pattern matching and similarity – based data analysis.

1.5 Major Issues in Data Mining
Major issues in data mining regarding mining methodology, user interaction, performance and diverse data types. These issues are,

Mining Methodology and User Interaction Issues

- Interactive mining of knowledge at multiple levels of abstraction
- Incorporation of background knowledge
- Data mining query languages and ad hoc data mining
- Presentation and visualization of data mining results
- Handling noisy or incomplete data
- Pattern evaluation – the interestingness problem

Performance Issues
Efficiency and scalability of data mining algorithms

- Parallel, distributed, and incremental mining algorithms

Issues Relating to the Discoversity of Database Types
Handling of relational and complex types of data

- Mining different data from heterogeneous databases and global information systems

The above issues are considered major requirements and challenges for further evolution of data mining technology.

II. LITERATURE SURVEY

P. Williams, R. Sion, and B. Carbunar
“Building castles out of mud: practical access pattern privacy and correctness on untrusted storage”
We introduce a new practical mechanism for remote data storage with efficient access pattern privacy and correctness. A storage client can deploy this mechanism to issue encrypted reads, writes, and inserts to a potentially curious and malicious storage service provider, without revealing information or access patterns. The provider is unable to establish any correlation between successive accesses, or even to distinguish between a read and a write. Moreover, the client is provided with strong correctness assurances for its operations illicit provider behavior does not go undetected. We built a first practical system -- orders of magnitude faster than existing implementations that can execute over several queries per second on 1Tbyte+ databases with full computational privacy and correctness.

P. Paillier, Cryptography Department “Public-Key Cryptosystems Based on Composite Degree Residuosity Classes”
This paper investigates a novel computational problem, namely the Composite Residuosity Class Problem, and its applications to public-key cryptography. We propose a new trapdoor mechanism and derive from this technique three encryption schemes: a trapdoor permutation and two homomorphic probabilistic encryption schemes computationally comparable to RSA. Our cryptosystems, based on usual modular arithmetic’s, are provably secure under appropriate assumptions in the standard model.

Craig Gentry, Stanford University and IBM Watson “Fully Homomorphic Encryption Using Ideal Lattices”
We propose a fully homomorphic encryption scheme – i.e., a scheme that allows one to evaluate circuits over encrypted data without being able to decrypt. Our solution comes in three steps. First, we provide a general result – that, to construct an encryption scheme that permits evaluation of arbitrary circuits, it suffices to construct an encryption scheme that can evaluate (slightly augmented versions of) its own decryption circuit; we call a scheme that can evaluate its (augmented) decryption circuit boots trappable. Next, we describe a public key encryption scheme using ideal lattices that is almost bootstrappable. Lattice-based cryptosystems typically have decryption algorithms with low circuit complexity, often dominated by an inner product computation that is in NC1. Also, ideal lattices provide both additive and multiplicative homeomorphisms (modulo a public-key ideal in a polynomial ring that is represented as a lattice), as needed to evaluate general circuits. Unfortunately, our initial scheme is not quite bootstrappable – i.e., the depth that the scheme can correctly evaluate can be logarithmic in the lattice dimension, just like the depth of the decryption circuit, but the latter is greater than the former. In the final step, we show how to modify the scheme to reduce the depth of the decryption circuit, and thereby obtain a bootstrappable encryption scheme, without reducing the depth that the scheme can evaluate. Abstractly, we accomplish this by enabling the encrypted to start the decryption process, leaving less work for the decrypted, much like the server leaves less work for the decrypted in a server-aided cryptosystem.

Y. Lindell and B. Pinkas “Privacy preserving data mining”
In this paper we address the issue of privacy preserving data mining. Specifically, we consider a scenario in which two parties owning confidential databases wish to run a data mining algorithm on the union of their databases, without revealing any unnecessary information. Our work is motivated by the need to both protect privileged information and enable its use for research or other purposes. The above problem is a specific example of secure multi-party computation and as such, can be solved using known generic protocols. However, data mining algorithms are typically complex and, furthermore, the input usually consists of massive data sets. The generic protocols in such a case are of no practical use and therefore more efficient protocols are required. We focus on the problem of decision tree learning with the popular ID3 algorithm. Our protocol is considerably more efficient

DOI 10.17148/IJARCCE.2016.54279

than generic solutions and demands both very few rounds of communication and reasonable bandwidth.

C. Gentry and S. Halevi "Implementing gentry's fully-homomorphic encryption scheme"
We describe a working implementation of a variant of Gentry's fully homomorphic encryption scheme (STOC 2009), similar to the variant used in an earlier implementation effort by Smart and Vercauteren (PKC 2010). Smart and Vercauteren implemented the underlying "somewhat homomorphic" scheme, but were not able to implement the bootstrapping functionality that is needed to get the complete scheme to work. We show a number of optimizations that allow us to implement all aspects of the scheme, including the bootstrapping functionality.
Our main optimization is a key-generation method for the underlying somewhat homomorphic encryption, that does not require full polynomial inversion. This reduces the asymptotic complexity from Õ (n2.5) to Õ(n1.5) when working with dimension-n lattices (and practically reducing the time from many hours/days to a few seconds/minutes). Other optimizations include a batching technique for encryption, a careful analysis of the degree of the decryption polynomial, and some space/time trade-offs for the fully-homomorphic scheme.

We tested our implementation with lattices of several dimensions, corresponding to several security levels. From a "toy" setting in dimension 512, to "small," "medium," and "large" settings in dimensions 2048, 8192, and 32768, respectively. The public-key size ranges in size from 70 Megabytes for the "small" setting to 2.3 Gigabytes for the "large" setting. The time to run one bootstrapping operation (on a 1-CPU 64- bit machine with large memory) ranges from 30 seconds for the "small" setting to 30 minutes for the "large" setting.

III.PROPOSED METHOD

Focus on solving the classification problem over encrypted data. In particular, we propose a secure k-NN classifier over encrypted data in the cloud. The proposed protocol protects the confidentiality of data, privacy of user's input query, and hides the data access patterns. To the best of our knowledge, our work is the first to develop a secure k-NN classifier over encrypted data under the semi-honest model. Also, we empirically analyse the efficiency of our proposed protocol using a real-world dataset under different parameter settings. Proposed novel methods to effectively solve the DMED problem assuming that the encrypted data are outsourced to a distributed system. Specifically, we focus on the classification problem since it is one of the most common data mining tasks. Because each classification technique has their own advantage, to be concrete, concentrates on executing the k-nearest neighbour classification method over encrypted data in the cloud computing environment.

Advantages of proposed system

- Data records correspond to the k-nearest neighbors and the output class label are not known to the cloud.
- We significantly ease the parameterization complexity for duplicate detection in general and contribute to the development of more user interactive applications.
- A secure k-NN classifier over semantically secure encrypted data.
- Privacy Preserving Approach.

Objective of the Project
Data Mining has wide applications in many areas such as banking, medicine, scientific research and among government agencies. Classification is one of the commonly used tasks in data mining applications. For the past decade, due to the rise of various privacy issues, many theoretical and practical solutions to the classification problem have been proposed under different security models. However, with the recent popularity of cloud computing, users now have the opportunity to outsource their data, in encrypted form, as well as the data mining tasks to the cloud. Since the data on the cloud is in encrypted form, existing privacy preserving classification techniques are not applicable.

IV. RESULT AND DISCUSSION

The new privacy preservation protocol implementation for the input query record classification over the encrypted database in the cloud is carried by the steps,

- Secure Data Upload
- Query Processing
- Secure KNN query process

ADMIN

List all Documents
In this module, the admin can view list all documents. If the admin click on the list all documents button, then the admin will get list of all documents with their tags such as author name, document type, date, location storm name, storm category, warnings, file name, ranks and admin can view the document.

List of searched History
In this module, the admin can view all searched history. If the admin clicks on search history button, then the admin will get all searched history details with their tags such as user name, key word1 used, key word2 used, key word3 used, time and date.

AUTHOR / SERVICE PROVIDER
The Author has to login by using valid user name and password. After login successful he can do some operations such as view my details, upload documents, request for secret key, and view my documents, Check Duplication and logout.

Upload Document
The Author or Service provider has already to register itself. The Author or Service provider is uploading the data to the distributed server, before it should be encrypted. The author can upload number of documents. Before uploading any documents, the author should request secret

DOI 10.17148/IJARCCE.2016.54279

key to admin, then the admin will provide a secret key, after getting a secret key, the author can enter document type, date, location, storm Name, storm category, warnings, attach file related to document and submit. After submit he will get response. Asymmetric –key algorithms require the use of asymmetric key pairs, consisting of a private key and corresponding public key. The key to be used for each operation depends on the cryptographic process being performed each public / private key pair is associate with only one entity; this entity is known as the key-pair owner. The public key may be known by anyone, whereas the private key must be known and used only by the key-pair owner. Key pairs are generated by the key-pair owner

View Document
The author can view all uploaded documents. If the author clicks on the view document button, then author will get all documents with their tags such as document type, date, location, storm Name, storm category, warnings, document file and rank. The author can download the document file.

B. USER
There are n numbers of users present. User should register before doing any operations. After registration successful he has to login by using authorized user name and password. After logged in he will do some operations such as view my details, search on queries, search on content, request for secret key, view my search history and logout. If user clicks on my details button, then the user will get all details with their tags such as user ID, User name, DOB, E-mail, Mobile, Location, gender and pin code.

Query Processing
After the user login to access the normal query window. In query process window, user to select the database name, table name and data owner access code from the database. In this process to protects the confidentiality of the data, user's input query, and hides the data access pattern. The user's input query will encrypted and pass to the cloud database. The cloud will classify label to corresponding query record. The query can retrieve the data from the cloud and show the encrypted and decrypted data in the output window.

Secure KNN query process
An authorized user sends the encrypted query to cloud server. The proposed PPKNN protocol is to classify user's query record using encrypted database in a privacy preserving manner. The PPKNN protocol has: PPKNN (Encrypted Database (D1),Query (Q))->Class Label(Cq). Where Cq denotes the class label for Q after applying k-NN classification method on D1 and Q. The KNN classification algorithm is a machine learning algorithm. It is a method for classifying objects based on closest training samples in the feature space. KNN is a type of instance-based learning; many test records will not be classified because they do not exactly match any of the training records. A more sophisticated approach, k-nearest neighbor (kNN) classification, finds a group of k objects in the training set that are closest to the test object, and bases the assignment of a label on the predominance of a particular class in this neighbourhood. There are three key elements of this approach: a set of labelled objects, e.g., a set of stored records, a distance or similarity metric to compute distance between objects, and the value of k, the number of nearest neighbors. To classify an unlabelled object, the distance of this object to the labelled objects is computed, its k-nearest neighbors are identified, and the class labels of these nearest neighbors are then used to determine the class label of the object.

THE PROPOSED PPKNN PROTOCOL MAINLY CONSISTS OF TWO STAGES
PROCEDURE
The proposed PPkNN protocol mainly consists of the following two stages:

Stage 1: Secure Retrieval of k-Nearest Neighbors (SRkNN)

- In this stage, User initially sends his query q (in encrypted form) to C1.
- After this, C1 and C2 involve in a set of sub-protocols to securely retrieve (in encrypted form) the class
- Labels corresponding to the k-nearest neighbors of the input query q.
- At the end of this step, encrypted class labels of k-nearest neighbors are known only to C1.

Stage 2: Secure Computation of Majority Class (SCMCk)

- C1 and C2 jointly compute the class label with a majority voting among the k-nearest neighbors of q.
- At the end of this step, only User knows the class label corresponding to his input query record q.

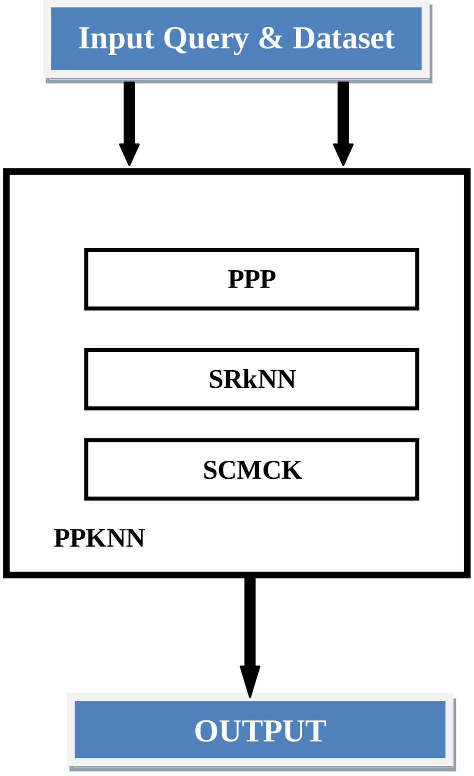

Figure 3 Procedural flow diagram

DOI 10.17148/IJARCCE.2016.54279

First of all, we emphasize that the outputs in the above mentioned protocols are always in encrypted format, and are known only to P1. Also, all the intermediate results revealed to P2 are either random or pseudo-random. Since the proposed SMIN protocol (which is used as a sub-routine in SMINn) is more complex than other protocols mentioned above and due to space limitations, we are motivated to provide its security proof rather than providing proofs for each protocol.

TABLE 1 COMPARISON OF VARIOUS ENCRYPTION SCHEME WITH K-NEAREST NEIGHBOR CLASSIFICATION

Methods	Advantages	Disadvantages
k-Nearest Neighbor Classification with Traditional Encryption Scheme	1. Data are secured 2. Classification error is very less due to decryption	1. If the data size is large then processing speed will become slow 2. Encryption and Decryption over head is very high
Privacy Preserving k-Nearest Neighbor Classification (PPkNN)	1. K-Nearest Neighbor classification to be carried out the encrypted data set 2. Only encryption method is used, is to reduce the overhead 3. Sensitive data are more secured	1. Encryption is done by using only partial scheme

IV. CONCLUSION

To ensure client security, different protection safeguarding arrangement systems have been proposed over the previous decade. The current strategies are not relevant to outsourced database situations where the information lives in encoded structure on an outsider server. This system will give novel protection safeguarding k-NN characterization convention over scrambled information in the database. Our system will secure the information's privacy, client's info inquiry, and conceals the information access designs. We likewise assessed the execution of our convention under distinctive parameter settings. Since enhancing the effectiveness of SMINn is a critical first stride for enhancing the execution of our PPkNN convention, we plan to research elective and more proficient answers for the SMINn issue in our future work. Additionally, we will examine and extend our exploration to other order calculations.

Future Work-Since enhancing the effectiveness of SMINn is an imperative initial step for enhancing the execution of our PPkNN convention, we plan to examine option and more proficient answers for the SMINn issue in our future work. Likewise, we will examine and extend our exploration to other characterization calculations. The performance of the proposed protocol depends on the efficiency of the SMINn protocol. Improving the SMINn will be the first scope of future work. Implementing this new privacy preserving protocol algorithm in the other classification methods and comparing the performance of those classification methods with current KNN classification method will be the second scope of future work.

REFERENCES

1. B. K. Samanthula, Y. Elmehdwi, and W. Jiang, “k-nearest neighbor classification over semantically secure encrypted relational data.” eprint arXiv:1403.5001, 2014.
2. Dharmendra Thakur and Prof. Hitesh Gupta,” An Exemplary Study of Privacy Preserving Association Rule Mining Techniques”, P.C.S.T., BHOPAL C.S Dept, P.C.S.T., BHOPAL India, International Journal of Advanced Research in Computer Science and Software Engineering ,vol.3 issue 11,2013.
3. R. Agrawal and R. Srikant, “Privacy-preserving data mining,” in ACM Sigmod Record, vol. 29, pp. 439–450, ACM, 2000.
4. A. Evfimievski, J. E. Gehrke, and R. Srikant. Limiting Privacy Breaches Privacy Preserving Data Mining. In Proceedings of the 22nd ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems (PODS 2003). San Diego, CA. June 2003.
5. M. Prakash, G. Singaravel, “A New Model for Privacy Preserving Sensitive Data Mining”, in proceedings of ICCCNT Coimbatore, India, IEEE 2012.
6. Y. Lindell and B. Pinkas, “Privacy preserving data mining,” in Advances in Cryptology (CRYPTO), pp. 36–54, Springer, 2000.
7. P. Zhang, Y. Tong, S. Tang, and D. Yang, “Privacy preserving naive bayes classification,” ADMA, pp. 744–752, 2005.
8. A. Evfimievski, R. Srikant, R. Agrawal, and J. Gehrke, “Privacy preserving mining of association rules,” Information Systems, vol. 29, no. 4, pp. 343–364, 2004.
9. R. J. Bayardo and R. Agrawal, “Data privacy through optimal k-anonymization,” in IEEE ICDE, pp. 217–228, 2005.
10. C.V.Nithya and A.Jeyasree,”Privacy Preserving Using Direct and Indirect Discrimination Rule Method”, Vivekananda College of Technology for Women Namakkal India, International Journal of Advanced Research in Computer Science and Software Engineering, vol.3 issue 12,2013.

DOI 10.17148/IJARCCE.2016.54279