Volume 6, Issue 9, September 2016 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

# Knowledge Acquisition and Privacy Preserving in Cloud using Simultaneous Diagonalization Algorithm

**Dr. Neelu Khare, Kumaran U, M. Mohan Vamsi**
SITE, VIT University, Vellore, Tamilnadu,
India

***Abstract— In the recent times the WWW(World Wide Web) has transformed from static collection of HTML data to a dynamic system that offered a platform for cloud technologies and distributed information systems. This paper describes how data mining is used in cloud computing while preserving its privacy. Data Mining is used for extracting potentially useful information from raw data. The integration of data mining techniques into normal day-to-day activities has become common place. Every day people are confronted with targeted advertising, and data mining techniques help businesses to become more efficient by reducing costs. Data mining techniques and applications are very much needed in the cloud computing paradigm. The implementation of data mining techniques through Cloud computing will allow the users to retrieve meaningful information from virtually integrated data warehouse that reduces the costs of infrastructure and storage.***

***Keywords—Privacy Preserving, Knowledge Acquisition, Data Mining, Cloud Computing, Simultaneous Diagonolization***

## I. INTRODUCTION

Internet is a vital tool in both professional and personal life.Most of the businesses are increasingly conducted over internet. One of the most revolutionary concepts of internet over recent years is cloud computing.

Cloud computing involves computing resources- hardware and software - that are delivered as a service over internet.

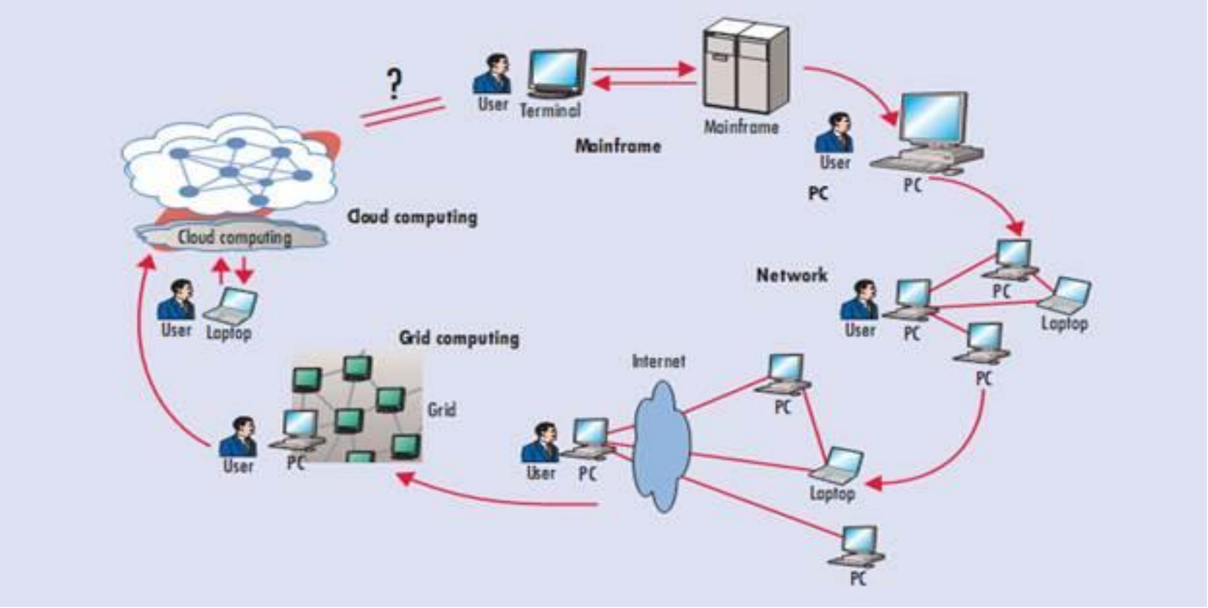

Figure 1 : Cloud Paradigm Shift of the last half century.

## II. LITERATURE SURVEY

**Top Cloud Computing Companies and its key features [1]**

| Cloud Name | Key Features |
|---|---|
| Sun Micro systems Cloud | More available application than any other open Operating Ststem |
| IBM Cloud Engine | Integrated power management helps us plan, predict, monitor and actively manage power consumption of BladeCenter server. |
| Amazon EC2 | Designed to make webscale computing easier for developers. |
| Google App Engine | No limit to the free trial period if you do not exceed the quota allotted. |

| | |
|---|---|
| Microsoft Azure | Currently offering a “development accelerator” discount plan. 15-30 % discount off consumption charges for first 6 months. |
| AT&T Synaptic Hosting | Use fully on-demand infrastructure or combine it with dedicated components to meet specialized requirements. |
| GoGrid Cloud Computing | Free load balancing and free 24/7 support. |
| Salesforce | Offers cloud solutions for automation, customer service and platform, respectively. Transparency through real-time information on system performance and security at trust.salesforce.com |

**Data Mining Techniques in Cloud [2]**

| Technique Name | Applicability |
|---|---|
| Clustering | Useful for exploring data and finding natural groupings. Members of a cluster are more like each other than they are like members of a different cluster. Common examples include finding new customer segments and life sciences discovery. |
| Classification | Most commonly used technique for predicting a specific outcome such as response / noresponse, high / medium / low value customer, likely to buy / not buy. |
| Association | Find rules associated with frequently cooccurring items, used for market basket analysis, cross-sell, root cause analysis. Useful for product bundling, in store placement, and defect analysis. |
| Regression | Technique for predicting a continuous numerical outcome such a customer lifetime value, house value, process yield rates. |
| Attribute Importance | Ranks attributes according to strength of relationship with target attribute. Use cases include finding factors most associated with customers who respond to an offer, factors most associated with healthy patients. |
| Anomaly Detection | Identifies unusual or suspicious cases based on deviation from the norm. Common examples include health care fraud, expense report fraud, and tax compliance. |
| Feature Extraction | Produces new attributes as linear combination of existing attributes. Applicable for text data, latent semantic analysis, data compression, data decomposition and projection, and pattern recognition. |

## III. PROBLEM STATEMENT

**Feature Extraction in Knowledge Discovery System**

Generally, feature extraction for classification can be seen as a search process among all possible transformations of the feature set for the best one, which preserves class separability as much as possible in the space with the lowest possible dimensionality.[3].

In other words we are interested in finding a projection w:

Y=w^T . X

where y is a p'×1 transformed data point (presented using p' features), w is a p× p' transformation matrix, and x is a p ×1 original data point (presented using p features).

The conventional PCA gives high weights to features with higher variabilities irrespective of whether they are useful for classification or not. This may give rise to the situation where the chosen principal component corresponds to the attribute with the highest variability but having no discriminating power.[4].

A usual approach to overcome the above problem is to use some class separability criterion, e.g. the criteria defined in Fisher linear discriminant analysis and based on the family of functions of scatter matrices:

$$J(\mathbf{w}) = \frac{\mathbf{w}^T \mathbf{S}_B \mathbf{w}}{\mathbf{w}^T \mathbf{S}_W \mathbf{w}}$$

where SB is the between-class covariance matrix that shows the scatter of the expected vectors around the mixture mean, and SW is the within-class covariance, that shows the scatter of samples around their respective class expected vectors.

## IV. PROPOSED SOLUTION

**Managing Feature Extraction and Privacy Preservation**

Currently, as far as we know, there is no feature extraction technique that would be the best for all data sets in the classification task. Thus the adaptive selection of the most suitable feature extraction technique for a given data set needs further research. Currently, there does not exist canonical knowledge, a perfect mathematical model, or any relevant tool to select the best extraction technique. Instead, a volume of accumulated empirical findings, some trends, and some dependencies have been discovered.

We consider a possibility to take benefit of the discovered knowledge by developing a decision support system based on the methodology of expert system design in order to help to manage the data mining process. The main goal of the system is to recommend the best-suited feature extraction method and a classifier for a given data set. Achieving this goal produces a great benefit because it might be possible to reach the performance of the wrapper type approach by using the filter approach. In the wrapper type approach the interaction between the feature selection process and the construction of the classification model is applied and the parameter tuning for every stage and for every method is needed.[5]. In the filter approach the evaluation process is independent from the learning algorithm and the methods, and their parameters' selection process is performed according to a certain set of criteria in advance. However, the additional goal of the prediction of model's output performance requires also further consideration.

## V. EXPERIMENTS AND RESULTS

Generally, the knowledge base is a dynamic part of the decision support system that can be supplemented and updated through the knowledge acquisition and knowledge refinement processes. Potential contribution of knowledge to be included into the KB might be found discovering a number of criteria from the experiments conducted on artificially generated data sets with pre-defined characteristics.

The results of experiments can be examined looking at the dependencies between the characteristics of a data set in general and the characteristics of every local partition of the instance space in particular. Further, the type and parameters of the feature extraction approach best suited for the data set will help to define a set of criteria that can be applied for the generation of rules of KB.

The results of our preliminary experiments support that approach. The artificially generated data sets were manipulated by changing the amount of irrelevant attributes, the level of noise in the relevant attributes, the ratio of correlation among the attributes, and the normality of the distributions of classes. In the experiments, supervised feature extraction (both the parametric and non parametric approaches) performed better than the conventional PCA when noise was introduced to the data sets. [6].

The similar trend was found with the situation when artificial data sets contained missing values. The finding was supported by the results of experiments on the LED17, Monk-3 and Voting UCI data sets (Table 1) that are known as ones that contain irrelevant attributes, noise in the attributes and a plenty of missing values.

Thus, this criterion can be included in the KB to be used to give preference to supervised methods when there exist noise or missing values in a data set. Nonparametric feature extraction essentially outperforms the parametric approach on the data sets, which include significant non-normal class distributions and are not easy to learn.

This initial knowledge about the nature of the parametric and nonparametric approaches and the results on artificial data sets were supported by the results of experiments on Monk-1 and Monk-2 UCI data sets.

| Dataset | PCA | Par | NPar | Plain |
|---|---|---|---|---|
| LED17 | .395 | .493 | .467 | .378 |
| MONK-1 | .767 | .687 | .972 | .758 |
| MONK-2 | .717 | .654 | .962 | .504 |
| MONK-3 | .939 | .990 | .990 | .843 |
| Voting | .923 | .949 | .946 | .921 |

## VI. CONCLUSION

We considered the goals of such a system, the basic ideas that define its structure and methodology of knowledge acquisition ,privacy preservation and validation. The Knowledge Base is the basis for the intelligence of the decision support system. That is why we recognised the problem of discovering rules from the experiments of an artificially generated data set with known predefined simple, statistical, and information theoretic measures, and validation of those rules on benchmark data sets as a prior research focus in this area. It should be noticed that the proposed approach has a serious limitation. Namely the drawbacks can be expressed in the terms of fragmentariness and incoherence (disconnectedness) of the components of knowledge to be produced.

## VII. FUTURE WORK

We do not claim the completeness of our decision support system. Otherwise, certain constrains and assumptions to the domain area were considered, and the limited sets of feature extraction methods, classifiers and data set characteristics were considered in order to guarantee the desired level of confidence in the system when solving a bounded set of problems.Also this privacy preservation measures can be integrated with the cloud framework and can be extended to Artificial Intelligent and Intelligence Systems like speech recognition,finger print scanner,face detection algorithms etc…

## REFERENCES

[1] Bhagyashree Ambulkar and Vaishali Borkar, “Data Mining in Cloud Computing”, MPGI National Multi Conference 2012 (MPGINMC-2012), 7- 8 April 2012

[2] ORACLE, “Oracle Data Mining Mining Techniques and Algorithms”, Link:http://www.oracle.com/technetwork/database/options/advancedanalytics/odm/odm-techniquesalgorithms-097163.html

[3] Bhushan Lal Sahu Rajesh Tiwari, “ A Comprehensive Study on Cloud Computing”, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 2, Issue 9, September 2012 ISSN: 2277 128X.

[4] Zhifeng Xiao and Yang Xiao, “Security and Privacy in Cloud Computing”, IEEE communications surveys & tutorials, vol. 15, no. 2, second quarter 2013.

[5] Vijay Varadharajan, and Udaya Tupakula, “Security as a Service Model for Cloud Environment”, IEEE transactions on network and service management, vol. 11, no. 1, march 2014.

[6] Mladen A. Vouk, “Cloud Computing Issues, Research and Implementations”, Journal of Computing and Information Technology - CIT 16, 2008, 4, 235–246.

[7] Rafael Moreno-Vozmediano, Rubén S. Montero, and Ignacio M. Llorente, “Key Challenges in Cloud Computing” Enabling the Future Internet of Services. IEEE, Volume: 17 , Issue: 4

[8] Harjit Singh, “Current Trends in Cloud Computing A Survey of Cloud Computing Systems”, International Journal of Electronics and Computer Science Engineering, ISSN- 2277-1956

[9] Maneesha Sharma, Himani Bansal and Amit Kumar Sharma, “Cloud Computing: Different Approach & Security Challenge”, International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-2, Issue- 1, March 2012.

[10] Gerard Conway and Edward Curry, “Managing Cloud Computing: A Life Cycle Approach”

[11] KPMG: From hype to future: KPMG’s 2010 Cloud Computing survey.

[12] Rosado DG, Gómez R, Mellado D, Fernández-Medina E: Security analysis in the migration to cloud environments. Future Internet 2012, 4(2):469–487.

[13] Mather T, Kumaraswamy S, Latif S: Cloud Security and Privacy. Sebastopol, CA: O’Reilly Media, Inc.; 2009.

[14] Li W, Ping L: Trust model to enhance Security and interoperability of Cloud environment. In Proceedings of the 1st International conference on Cloud Computing. Beijing, China: Springer Berlin Heidelberg; 2009:69–79.

[15] Rittinghouse JW, Ransome JF: Security in the Cloud. In Cloud Computing. Implementation, Management, and Security, CRC Press; 2009.

[16] Kitchenham B: Procedures for perfoming systematic review, software engineering group. Australia: Department of Computer Scinece Keele University, United Kingdom and Empirical Software Engineering, National ICT Australia Ltd; 2004. TR/SE-0401 TR/SE-0401

[17] Kitchenham B, Charters S: Guidelines for performing systematic literature reviews in software engineering. Version 2.3 University of keele (software engineering group, school of computer science and mathematics) and Durham. UK: Department of Conputer Science; 2007.

[18] Brereton P, Kitchenham BA, Budgen D, Turner M, Khalil M: Lessons from applying the systematic literature review process within the software engineering domain. J Syst Softw 2007, 80(4):571–583. 10.1016/j.jss.2006.07.009

[19] Cloud Security Alliance: Top Threats to Cloud Computing V1.0. 2010. Available:https://cloudsecurityalliance.org/research/top-threats

[20] ENISA: Cloud Computing: benefits, risks and recommendations for information Security. 2009. Available:http://www.enisa.europa.eu/activities/risk-management/files/deliverables/cloud-computing-risk-assessment

[21] Dahbur K, Mohammad B, Tarakji AB: A survey of risks, threats and vulnerabilities in Cloud Computing. In Proceedings of the 2011 International conference on intelligent semantic Web-services and applications. Jordan: Amman; 2011:1–6