International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2017 IJSRCSEIT | Volume 2 | Issue 6 | ISSN : 2456-3307

Compressive Study on Data Mining Methods in Cloud Computing

Achi Sandeep*, K Gurnadha Guptha
Assistant Professor, CSE Department, Sri Indu College of Engineering and Technology, JNTU Hyderabad, Hyderabad, India

ABSTRACT

Data mining is the technique to find out previously unknown relationship and new patterns in big data, which even can predict for future decisions by the use of some helpful algorithms and techniques like clustering, classification, association, regression etc. Cloud computing is delivering software, platform and infrastructure as a service over the internet and customer can access these from a remote area. Rapid development of internet and increase in commerce allows the firms to safe guard, store, retrieve and analysis of their data through cloud services with data mining. Data mining in cloud computing is the process of finding structured information from web data sources which might be unstructured or semi-structured.

Keywords: Data Mining; Cloud Computing; Data Mining Techniques; Clustering; BC-PDM, PD miner, ESOM-Maps

I. INTRODUCTION

Today‘s era uses these buzzword cloud computing and data mining to extract out hidden possibilities in their data pools, which are located somewhere on remotely accessed storage servers or cloud. A great amount of data coming from marketing, detection in surveillance fraud, geological department, issues related to human factor, scientific discovery, data in the medical field, geographical information system, ecosystem etc. requires deep analysis & proper decision making which is achieved through data mining. Data mining is a method where database analysis is used to discover hidden relationship and useful patterns helps companies to focuses on their essential information in their warehouses. Various Enterprises, Colleges, Hospitals, geologists etc. uses data mining to analyze their data warehouses for some previously hidden unknown data and can even make future prediction of their data. Data mining has developed uses in various fields of activities today. Data mining is achieved through various data mining techniques like clustering, association, sequence and path analysis, anomaly detection, neural networking, genetic algorithm, Forecasting etc. Cloud computing is a paradigm where data is permanently stored in servers and can be accessed whenever needed via internet on desktops, laptops, tablets etc. cloud have no limitations and information can be located anywhere in the world. Cloud computing provides its users to access power at the level of super computer due to its characteristics such as high level of elasticity that it offers in networking, storage and processing. Use of data mining techniques in cloud computing enables users to extend their power to analyze their data that is situated on cloud storage.

II. DATA MINING

1. Data Mining

A computational process extracts potentially useful patterns or relationships from raw data, is a new technology with huge potential to help companies to derive benefits from resources valuable in their data warehouses. It is used for enhancement and replacement of human intelligence by scanning through massive data warehouses to discover meaningful patterns, new correlations, and trends, by using pattern recognition technologies and advanced statistics [8]. The overall motive of data mining is to fetch data from an existing dataset and transform it into an understandable pattern for further use. This process is often known as Knowledge Discovery in database (KDD).

B. Data mining technique

a. Clustering : Useful for finding natural groupings and exploring data. Members of cluster are more similar to each apart from the members belonging to other clusters. Some example includes life science discovery and finding a new customer segments.
b. Classification : Used mostly in predicting a specific result like response high/medium/low or like to buy/not to buy.
c. Association: Arrive at the set of rules related to items that co-occur frequently and are made use in cross-sell, market basket analysis, and analysis of root cause.
d. Regression Technique : It is for predicting continuous numerical results such a customer process yield rates.

- Attribute Importance : According to effectiveness of relationship with target attribute, attributes are ranked. Use cases include searching factors most linked with customers who respond to an offer.
- Anomaly Detection: It identifies suspicious cases or unusual based on deviation from the norm. Common examples include expense report fraud, health care fraud, and tax compliance.
- Feature Extraction: It produces new attributes as linear combination of previously existing attributes. Applicable for projection, text data decomposition and pattern recognition.

There are various applications of data mining in real world as, web mining, space organization, mathematics, biometrics, forecasting, government, telecommunication, airline reservation, hospital, student management, etc.

C. Clustering

Clustering is an advance unsupervised learning technique, by which records similar to each other are grouped. Generally, this is done to provide the user a high-level view of what is going on in the database [5] [12]. Clustering is also referred to as segmentation - which most business people will take it helpful in obtaining a hawk eye view of the business. It is utilized mainly for the purpose of consolidation of information into a high-level view and general clustering of records into like behaviors. Data Space is set as default n-dimensional space, or is set by the user, or is a predefined space driven by past driven experience (unsupervised learning).

There are two main types of clustering techniques: Non-hierarchical clusters and hierarchy of clusters. The hierarchy of clusters is defined as a tree where the smallest clusters fuses together to form the next higher level of clusters and those at that level fuse together to form next higher level of clusters. This hierarchy of clusters is achieved through the algorithms that create clusters.

Two main types of hierarchical clustering algorithms:

1. Divisive - Divisive clustering technique approach is opposite to agglomerative technique‘s approach. These techniques initiate with all records in one cluster and then try to disintegrate that cluster into smaller clusters and then in turn to try to disintegrate those smaller pieces. [12]

Algorithm

1. Start considering each of the inputs as a separate cluster and each successive step combine the nearest pair of clusters.
2. C representative points are stored to calculate the distance between a pair of cluster.
3. These are determined by first selecting C well-scattered points within the cluster and then diminishing them towards center of the cluster by a fraction α.
4. Representative points of cluster are used to calculate its distance from other clusters.

E. BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies)

BIRCH algorithm is an agglomerative type hierarchical clustering algorithm. It is used for very large databases because it reduces the number of input/output operations. BIRCH works by using tree structure for partitioning objects hierarchically and then other clustering algorithm used to refine clusters. BIRCH dynamically and incrementally clusters incoming multi-dimensional metric data points to attempt to yield the highest quality clustering with the resources that are available like time constraints and available memory. To produce best quality clusters this process takes four phases [20].

F. K-means Clustering Algorithm

It is one of the easiest unsupervised learning algorithms that are quite efficient in solving complex clustering problems. The procedure employs an easy and simple way to categories a given data set into a certain number of clusters. [9]

K-means clustering algorithm groups/clusters the various observations related to each other without the any idea of the relationships existing among them. Some feature vectors in an n-dimensional space can be used to represent the objects, where n means the total number of the features that are being used for

description of the clusters. After this is done, the algorithm chooses k-points in the vector space randomly. These points act as initial centers of the cluster. Then all objects are assigned to the center points, which are at a least distance from them. Going this way a separate and new center is generated through the mean of the feature vectors of all the objects to which they are assigned.

The process of assigning objects and recreating the centers is done repeatedly until the converging of the process.

Algorithm

1. Choose k number of items in a random way and declare them as initial centroids.
2. Find the nearest centroid for each of the points and then assign the point to the cluster associated with the centroid positioned at a least distance from the point.
3. Update the centroid of each cluster based on the items present in that cluster. Generally, the new updated centroid will be the mean of all points in the cluster.
4. Repeats steps 2 and 3, until no point switches cluster.

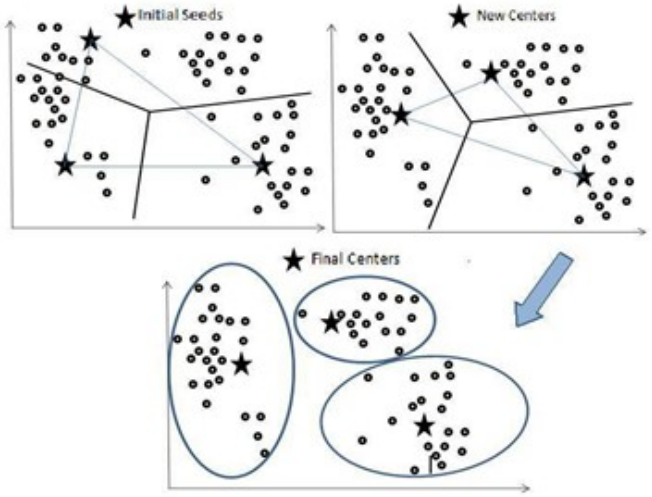

Figure 2 : The working of K-means clustering algorithm

III. CLOUD COMPUTING

Cloud computing in a concept of providing computing resources like servers, networks application software in a way that is all pervasive, readily available to customers on demand, can be made use by multiple users one after the other with minimized efforts required in managing the resources.

Cloud computing model consists of three service models, four deployment models and five required characteristics. Service models of cloud computing is classified as Platform as a Service (PaaS), Infrastructure as a Service (IaaS) and Software as a Service (SaaS). The deployment models of cloud computing are hybrid cloud, public cloud, community cloud, and private cloud. Essential characteristics of cloud computing are: resource pooling, measured service, broad network access on-demand self-service and rapid elasticity.

IaaS: Provides computer infrastructure as a utility service, commonly in a virtualized environment.
Provides great scope for scalability and extensibility.

PaaS: Provides a platform on a cloud infrastructure. It is positioned over underlying IaaS architecture and is merged with development and middleware capabilities along with database, queuing and messaging functions.

SaaS: Provides application software as a service over the networks– both internet and intranet via a cloud Infrastructure Built over IaaS and PaaS Layer.

Cloud computing represents all existing resources on the Internet, offering unlimited computing power [3].

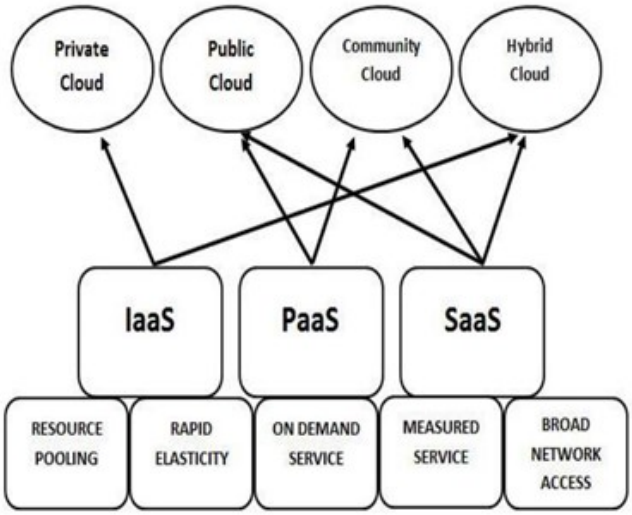

Figure 3 : Cloud Deployment Model

IV. DATA MINING IN CLOUD COMPUTING

Data mining in cloud computing is the process of finding structured information from web data sources-both structured as well as non-structured ones. Data mining in Cloud Computing allows users to centralize the data storage and management of software, with assured services in terms of efficiency, reliability and security. Since the cloud, computing is a technique that

delivers software, hardware and infrastructure as services over the Internet, data mining software is provided in this way as in:

a)SaaS: Developers can use ready to use data mining tools and well defined algorithms which can be accessed directly through web browser.

b)PaaS: Developers can use supporting platforms for their data analytics and storage like Hadoop, Apache Cassandra.

c)IaaS: Developers can make use of virtualized resources belonging to computing infrastructure for the purpose of implementing data analytics and to use data mining tools.

Some platforms are developed to provide data mining in cloud services. Talia etal [11] [13] summarize data mining in cloud computing in four levels. Single KDD steps: the underlying composition data mining algorithms.

Single data mining tasks: a separate class of services meant for data mining like classification, clustering, etc.

Distributed data mining patterns: distributed data mining models like aggregation, parallel classification, and machine learning. Data mining applications or KDD processes: Complete data mining application that is based on the elements belonging to all above. Some of the data mining tools being delivered on cloud are:BC-PDM, It provides SaaS service and minimize the investment of enterprises over IT systems. It is based on the Map reduction implementation on cloud computing. BC-PDM [11] [15] is a set Consisting of mining system and mass data processing analysis. It has high performance low cost scalability characteristics.

PD miner [11] developed by institute of computing technology is a parallel distributed data mining platform that is based on Hadoop. It features open architecture, which makes it possible for its user to back a loaded algorithm components into system through a simple configuration

ESOM-Maps are a data-mining tool used for clustering; visualization and classification can be used as SaaS via cloud.

Workflow-based data mining frameworks that operate over cloud platforms and make use of an approach that is service oriented and that offers a very flexible model of programming, distributed task interoperability, and last but not the least, execution scalability that drastically reduces data analytics completion time.

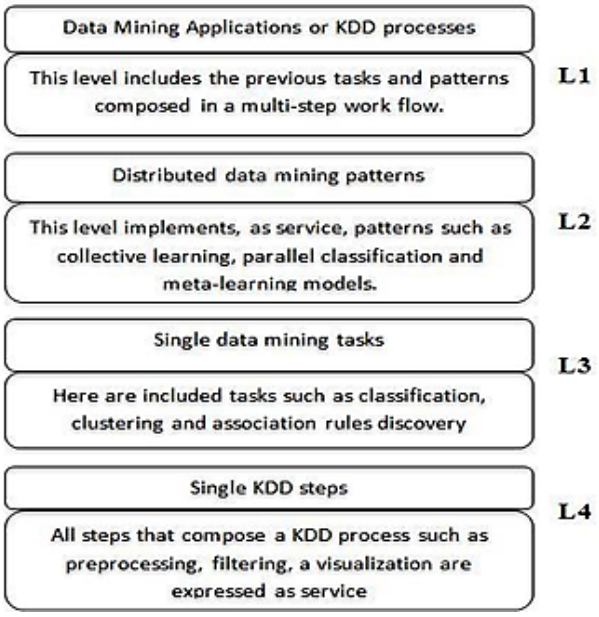

Figure 4: Four levels of data mining services [11]

Data analysis tasks, scientific methods of computation, and complex techniques of simulation can be designed by application developers as workflows in a way that they integrate single Web services and get them concurrently executed on virtual machines in the cloud.

A browser – this implies that he has to only pay the costs that are generated by using Cloud computing.

In addition, major companies in the area of Business Intelligence provide data mining services that are business-oriented large-scale, like micro-strategy, IBM and many other companies that own cloud computing platform based data mining services.[11]

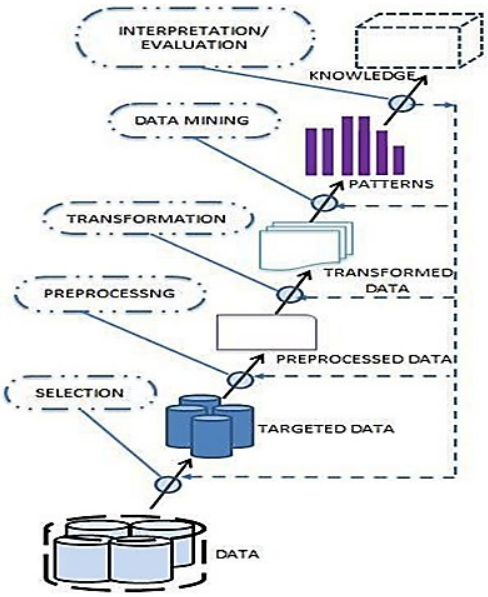

Figure 5: Knowledge Discovery in Data Mining

The Effects of data mining tools being delivered on Cloud are [3]: the customer only pay for the data mining tools that he/she uses – that lowers his costs as he is not required to pay for complex data mining suites that is not being used by him/her. The customers do not have to look after a hardware infrastructure, as they through can apply data mining

V. CONCLUSION

After going through many papers what we can conclude and infer is that data mining is an advanced technique that is quite useful in the fields wherever one needs to search, analyze and discover unknown or hidden patterns in a set of raw data. We also have concluded that clustering serves as one of the best algorithms used for data mining processes owing to its less complexity and ability be readily implemented without any hindrance. Cloud computing does the demand of all the industries exist today that are required to utilize computing resources over the internet. In order to cope up with the requirements and demands of industries - both those existing at present and those that would come up in the future, that require data analysis, a large number of tools can be used for the purpose of gathering information from data warehouses that are situated over cloud or virtual servers. Some of the very useful tools used in cloud computing for data mining are BC-PDM, PD miner, ESOM-Maps.

VI. REFERENCES

[1]. Prof. Mr. A. Srinivas, M. Kalyan Srinivas, A.V.R.K.Harsha Vardhan Varma:A Study On Cloud Computing Data Mining,International Journal of Innovative Research in Computer and Communication Engineering, July 2013

[2]. Ritu Chauhan,Harleen Kaur, M.Afshar Alam:Data Clustering Method for Discovering Clusters in Spatial Cancer Databases,International Journal of Computer Applications (0975 – 8887), November 2010

[3]. Ruxandra-Ştefania PETRE,Data mining in Cloud Computing :Database Systems Journal vol. III, no. 3/2012

[4]. Bhagyashree Ambulkar,Vaishali Borkar:Data mining in Cloud Computing,MPGI National Multi Conference 2012 (MPGINMC-2012),7-8 April 2012

[5]. Astha Pareek, Manish Gupta:Review of Data Mining Techniques in Cloud Computing Database,International Journal of Advanced Computer Research,Volume-2 Number-2 Issue-4 June-2012

[6]. Prof. V. B. Nikam, Viki Patil:Study of Data Mining algorithm in cloud computing using MapReduce Framework Journal of Engineering, Computers & Applied Sciences (JEC&AS) Vol ume 2, No.7, July 2013

[7]. Pavel Berkhin, Accrue Software, 1045 Forest Knoll Dr., San Jose, CA, 95129:Survey of Clustering Data Mining Technique

[8]. Jeffrey Voas and Jia Zhang, ―Cloud Computing: New Wine or Just a NewBottle?‖, Database Systems Journal vol. III, no. 3/2012 71 IEEE Internet Computing Magazine, 200

[9]. Yudho Giri Sucahyo, Ph.D, CISA:Introduction to Data Mining and Business Intellegence

[10]. Tapas Kanungo,Nathan S. Netanyahu, Angela Y. Wu:An Efficient k-Means Clustering Algorithm:Analysis and Implementation, ieee transactions on pattern analysis and machine intelligence, vol. 24, no. 7, july 2002

[11]. Xia Geng1,a,Zhi Yang2,b:Data Mining in Cloud Computing, International Conference on Information Science and Computer Applications (ISCA 2013)

[12]. Peter Mell, Timothy Grance:The NIST Definition of Cloud Computing,U.S. Department of Commerce, Special Publication 800-145

[13]. D. Talia and P. Trunfio, How distributed data mining tasks can thrive as knowledge services Communications of the ACM. 53(2010) 132-137

[14]. Boniface, M.; et al. (2010), ―Platform-as-a-Service Architecture for Real-Time Quality of Service Management in Clouds‖, 5th InternationalConference on Internet and Web Applications and Services (ICIW), Barcelona, Spain: IEEE, pp. 155–160.

[15]. L. Yu, J. Zheng, W. C. Shen, B. Wu, B. Wang, L. Qian, B. R. Zhang, BC-PDM: data mining, social network analysis and text mining system based on cloud computing, Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining. (2012) 1496-1499

[16]. Jain, A.K., Murty M.N., and Flynn P.J. (1999): Data Clustering: A Review

[17]. L. Kaufinan, and P.J. Rousseeuw, Finding Groups in Data: an Introduction to Cluster Analysis, John Wiley & Sons1990.
[18]. M. S. Chen, J. Han, and P. S. Yu. Data mining: an overview from database perspective. IEEE Trans. On Knowledge and Data Engineering, 5(1):866—883, Dec.1996
[19]. Muhammad Husnain Zafar, and Muhammad Ilyas: A Clustering Based Study of Classification Algorithms, International Journal of Database Theory and Application, Vol.8, No.1 (2015), pp.11-22
[20]. Yogita Rani, Dr. Harish Rohil: A Study of Hierarchical Clustering Algorithm, International Journal of Information and Computation Technology, Volume 3, Number 11 (2013), pp. 1325-123