International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2017 IJSRCSEIT | Volume 2 | Issue 5 | ISSN : 2456-3307

The Efficient Clustering algorithms for Data Mining : A Review

Gurveer Singh[1], Dinesh Kumar[2]
[1]Research Scholar, Giani Zail Singh Campus College of Engineering & Technology, Bathinda, Punjab, India
[2]Associate Professor, Giani Zail Singh Campus College of Engineering & Technology, Bathinda, Punjab, India

ABSTRACT

The data mining is the technique which is used to mine the useful information from the rough data. The clustering is the technique which is used under data mining to cluster similar and dis-similar type of data. The various algorithms has been proposed in data mining to cluster similar and dissimilar type of data. In this work, various clustering algorithms are reviewed and it is been analyzed that symmetric and asymmetric type of algorithms are performed well in terms of accuracy and exaction time.

Keywords: Clustering, Symmetric, Asymmetric, K-Means

I. INTRODUCTION

Data mining, the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions [1]. Most companies already collect and refine massive quantities of data. Data mining techniques can be implemented rapidly on existing software and hardware platforms to enhance the value of existing information resources, and can be integrated with new products and systems as they are brought on-line. Data mining derives its name from the similarities between searching for valuable information in a large database and mining rocks for a vein of valuable ore. Both imply either sifting through a large amount of material or ingeniously probing the material to exactly pinpoint where the values reside [2].

Clustering is a division of data into groups of similar objects. Each group, called cluster, consists of objects that are similar between themselves and dissimilar to objects of other groups. Representing data by fewer clusters necessarily loses certain fine details (akin to lossy data compression), but achieves simplification. It represents many data objects by few clusters, and hence, it models data by its clusters. Representing data by fewer clusters necessarily loses certain fine details (akin to lossy data compression), but achieves simplification [3]. It represents many data objects by few clusters, and hence, it models data by its clusters. Data modeling puts clustering in a historical perspective rooted in mathematics, statistics, and numerical analysis [4].

There are two types of clustering techniques: Partitional and Hierarchical. In partitional Clustering given a database of *n* objects, a partitional clustering algorithm constructs *k* partitions of the data, where each cluster optimizes a clustering criterion, such as the minimization of the sum of squared distance from the mean within each cluster. Hierarchical algorithms create a hierarchical decomposition of the objects [5]. They are either agglomerative (bottom-up) or divisive (top-down). Agglomerative algorithms start with each object being a separate cluster itself, and successively merge groups according to a distance measure. Divisive algorithms follow the opposite strategy. They start with one group of all objects and successively split groups into smaller ones, until each object falls in one cluster, or as desired [6].

Apart from the two main categories of partitional and hierarchical clustering algorithms, many other methods have emerged in cluster analysis, and are mainly focused on specific problems or specific data sets available. Density-Based Clustering algorithms group

objects according to specific density objective functions. In these approaches a given cluster continues growing as long as the number of objects in the neighborhood exceeds some parameter [7]. Grid-Based Clustering has main focus on spatial data, i.e., data that model the geometric structure of objects in space, their relationships, properties and operations [8]. The objective of these algorithms is to quantize the data set into a number of cells and then work with objects belonging to these cells. Model-Based Clustering algorithms find good approximations of model parameters that best fit the data. They are closer to density-based algorithms, in that they grow particular clusters so that the preconceived model is improved. Categorical Data Clustering algorithms are specifically developed for data where Euclidean, or other numerical-oriented, distance measures cannot be applied [9].

II. Literature Review

MD. Ezaz Ahmed, et.al (2013) proposed in paper [10] that unlabeled document collections are turning out to be progressively common and mining such databases turns into a noteworthy test. It is a noteworthy issue to retrieve good websites from the larger collections of websites. As the number of available Web pages grows, it is turned out to be more troublesome for clients finding documents applicable to their interests. Clustering is the classification of a data set into subsets (clusters), so that the data in every subset share some common trait – frequently proximity as per some defined distance measure. By clustering one enhances the quality of websites by grouping comparable websites in groups. This paper addresses the applications of data mining tool Weka by applying k means clustering to discover clusters from huge data sets and discover the characteristics that represent advancement of search engines.

Mohnish Patel, et.al (2014) proposed in this paper [11] that Efficient Privacy preserving association rule mining has emerged as a most recent research issue. In this theory work, existing algorithms, Increase Support of Left and Decrease Support of Right are implemented effectively on the real data for Privacy Preserving Association Rule Mining. Keeping in mind the end goal to hide an association rule, a hybrid algorithm is proposed which is based on two previous existing algorithms ISL and DSR. KMean, Neural gas Cluster Algorithm with Number of cluster in this algorithm, first the support of right hand side of the rule in a rule is decreased where object to be hide is in right side for Experimental work. Authors have utilized a real time database of Doctor Patient Evaluation from Medical College.

Richa Sharma, et.al (2016) proposed in this paper [12] that one of the applications of data mining is disease diagnosis for this purpose one needs medical dataset to identify hidden patterns lastly extracts valuable knowledge from medical database. As of late, researchers have utilized different classification and clustering algorithms for diagnosing diseases. This paper gives overview on two different complex diseases which incorporates the heart disease and Cancer disease, paper fundamentally watched the existing writing work to discover significant knowledge in this area and summarized different approaches utilized as a part of disease diagnosing, promote examined about the tools available for processing and classification of data. This study reveals the importance of research in area of life debilitating disease diagnosis.

G. Anuradha, et.al (2014) proposed in this paper [13] that the popular expression in research is Big Data. Big Data gets described by 5 V's: Volume, Velocity, Variety, Veracity and Value of data. Volume all together of penta bytes, velocity which alludes to click stream data in various domains, variety containing heterogeneous data, veracity demonstrating the cleanliness of data and value emphasizing on the arrival on investment for companies who invest in Big Data technologies. This Big Data is better modeled not as persistent tables but rather as transient data streams which require different clustering and mining techniques to be effectively processed and managed. In this paper a few suggestions on online learning through clustering and mining of stream data are introduced. Since the volume and velocity of big data keeps expanding consistently, more propelled techniques for clustering and mining such humongous data is the need of the hour.

Edem Inang Edem, et.al (2015) proposed in this paper [14] that the proliferation of malware as of late have accounted for the increase in computer crimes and prompted for a more aggressive research into improved investigative strategies, to keep up with the menace.

Late techniques and tools that have been developed and adopted to keep up in an arms race with malware authors who have resorted to the utilization of evasive techniques to avoid examination amid investigation is an on-going concern. Exploring dynamic examination is unarguably, a positive step to supporting static evidence with malware dynamic conduct logs. In perspective of this, dissecting these huge generated reports raises concerns about speed, accuracy and performance. The implementation results of the sub-components recorded in this study demonstrated a considerable time gain in feature extraction utilizing unique feature approach furthermore yielded an improved data matrix embedding strategy which made data mining clustering of behavioral reports data got from an online sandbox relatively fast utilizing k-means with appropriate distance measure.

CHENG-FA TSAI, et.al (2014) proposed in this paper [15] that this investigation develops another data clustering method. It is another density-based clustering scheme by diagonal sampling and another technique for crease and revolution for enhancing data clustering performance. The proposed algorithm's expansion without selecting data points to increase computation cost and it might considerably lower time cost. The experimental results affirm that the displayed approach has genuinely high clustering accuracy and noise filtering rate, and is faster than numerous notable existing density-based data clustering algorithms, for example, DB SCAN, IDBSCAN, KIDBSCAN and FDBSCAN approaches. Experimental results indicate that the proposed data clustering algorithm surpasses other existing celebrated real approaches, for example, the DB SCAN, IDBSCAN, KlDBSCAN, and FDBSCAN techniques, and its high accuracy and low execution-time cost make it efficient and effective for data clustering in numerous data mining applications.

III. RESULTS AND DISCUSSION

Table 1: Table of Comparison

Name	Year	Description	Outcome
MD. Ezaz Ahmed, Preeti Bansal	2013	In this paper the authors proposed that unlabeled document collections are turning out to be progressively common and mining such databases turns into a noteworthy test.	This paper addresses the applications of data mining tool Weka by applying k means clustering to discover clusters from huge data sets and discover the characteristics that represent advancement of search engines.
Mohnish Patel, Prashant Richhariya, Anurag Shrivastava	2014	In this theory work, existing algorithms, Increase Support of Left and Decrease Support of Right are implemented effectively on the real data for Privacy Preserving Association Rule Mining.	Keeping in mind the end goal to hide an association rule, a hybrid algorithm is proposed which is based on two previous existing algorithms ISL and DSR.
Richa Sharma, Dr. Shailendra Narayan Singh, Dr. Sujata Khatri	2016	This paper gives overview on two different complex diseases which incorporates the heart disease and Cancer disease, paper fundamentally watched the existing writing work to discover significant knowledge in this area and summarized different approaches utilized as a part of disease diagnosing, promote	This study reveals the importance of research in area of life debilitating disease diagnosis.

		examined about the tools available for processing and classification of data.	
G. Anuradha, Bidisha Roy	2014	In this paper a few suggestions on online learning through clustering and mining of stream data are introduced.	Since the volume and velocity of big data keeps expanding consistently, more propelled techniques for clustering and mining such humongous data is the need of the hour.
Edem Inang Edem, Chafika Benzaidy, Ameer Al-Nemrat and Paul Watters	2015	Exploring dynamic examination is unarguably, a positive step to supporting static evidence with malware dynamic conduct logs. In perspective of this, dissecting these huge generated reports raises concerns about speed, accuracy and performance.	The implementation results of the sub-components recorded in this study demonstrated a considerable time gain in feature extraction utilizing unique feature approach.
CHENG-FA TSAI, PO-VI SHE	2014	The proposed algorithm's expansion without selecting data points to increase computation cost and it might considerably lower time cost.	The experimental results affirm that the displayed approach has genuinely high clustering accuracy and noise filtering rate, and is faster than numerous notable existing density-based data clustering algorithms.

IV. CONCLUSION

In this paper, it is been concluded that various techniques has been proposed which can cluster the similar and dissimilar type of data in the efficient manner . The basic steps for data clustering includes to calculate arithmetic mean of the input data set which will be the central point to calculate Euclidian distance to all other points. The similarity between the points are calculated by applying affinity matrix. The similar and dissimilar points are clustered together and it Is been analyzed that asymmetric clustering performs well in terms of accuracy and execution time

V. REFERENCES

[1]. Geqi Qi, Yiman Du, Jianping Wu, Ming Xu," Leveraging longitudinal driving behavior data with data mining techniques for driving style analysis", 2015, IET Intell. Transp. Syst., Vol. 9, Iss. 8, pp. 792-801

[2]. M. Guder, O. Salor, I. Çadirci, B. Ozkan and E. Altintas," Data Mining Framework for Power Quality Event Characterization of Iron and Steel Plants", 2015, IEEE, 0093-9994

[3]. Claudia Plant, Andrew Zherdin, Christian Sorg, Anke Meyer-Baese, and Afra M. Wohlschläger," Mining Interaction Patterns among Brain Regions by Clustering", 2014, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 26, NO. 9

[4]. JOHANNES GRABMEIER, ANDREAS RUDOLPH," Techniques of Cluster Algorithms in Data Mining", 2002 Kluwer Academic Publishers, 303-360

[5]. P. Berkhin," A Survey of Clustering Data Mining Techniques", 2010, Springer, 3485-34-533

[6]. G. Sreenivasulu, S. Viswanadha Raju and N. Sambasiva Rao," Review of Clustering Techniques", 2016, Springer Science+Business Media Singapore

[7]. Lamine M. Aouad, Nhien-An Le-Khac, and Tahar M. Kechadi," Lightweight Clustering Technique for Distributed Data Mining

Applications", 2007, ICDM, LNAI 4597, pp. 120-134
[8]. Murilo Coelho Naldi," Genetic Clustering for Data Mining", 2001, Springer, 35-343-578, pp-343-967
[9]. Francesco Gullo," From Patterns in Data to Knowledge Discovery: What Data Mining Can Do", 2015, Francesco Gullo / Physics Procedia 62, 18 - 22
[10]. MD. Ezaz Ahmed, Preeti Bansal," Clustering Technique on Search Engine Dataset using Data Mining Tool", 2013, IEEE, 978-0-7695-4941
[11]. Mohnish Patel, Prashant Richhariya, Anurag Shrivastava," A Novel Approach for Data Mining Clustering Technique using NeuralGas Algorithm", 2014, IEEE, 978-1-4799-4910
[12]. Richa Sharma, Dr. Shailendra Narayan Singh, Dr. Sujata Khatri," Medical Data Mining Using Different Classification and Clustering Techniques: A Critical Survey", 2016, IEEE, 978-1-5090-0210
[13]. G. Anuradha, Bidisha Roy," Suggested Techniques for Clustering and Mining of Data Streams", 2014, IEEE, 978-1-4799-2494
[14]. Edem Inang Edem, Chafika Benzaidy, Ameer Al-Nemrat and Paul Watters," Analysis of Malware Behaviour: Using data Mining Clustering Techniques to Support Forensics Investigation", 2015, IEEE, 978-1-4799-8825
[15]. CHENG-FA TSAI, PO-VI SHE," A NEW EFFICIENT DENSITY-BASED DATA CLUSTERING TECHNIQUE USING CROSS EXPANSION FOR DATA MINING", 2014, IEEE, 978-1-4799-4215