2013 International Conference on Advances in ICT for Emerging Regions (ICTer): 228 - 233

Dynamic Partitional Clustering Using Multi-Agent Technology

D.M.M.B. Dehideniya, A.S. Karunananda
Department of Computational Mathematics,
University of Moratuwa
Sri Lanka

mahasen_d@yahoo.com, asoka@itfac.mrt.ac.lk

Abstract*— Most of the well established clustering algorithms assume that the underlying clustering structure of dataset does not change over the time. Hence, those algorithms fail to identify underlying cluster structures in currently available large scale dynamic data sources in an efficient manner. This paper presents a Multi Agent based approach to identify partitional clusters in a dynamic data source. Set of partitional clusters in a dynamic data source is identified by interactions and negotiations among the agents who represent data records in the data source. After identification of potential clusters for data records that are assigned to what are called cluster agents. By interactions and negotiations between cluster agents and data record agents, the identified cluster configuration is continuously improved according to the internal cluster evaluation measures. The proposed method is evaluated by synthetic data sets with different number of clusters in 2D and 3D spaces. Results indicate that the proposed method successfully identifies the clusters in those datasets with minimal human intervention.

Keywords — **Partitional Clustering , Dynamic Clustering , Multi-agent Technology**

I. INTRODUCTION

Data Clustering is a main unsupervised machine learning technique which identifies meaningful groups of objects that have common characteristics in a given dataset. It is also employed as an important starting step which identifies the natural structure of data in many fields such as Statistics, Machine Learning, Patter Recognition, Information Retrieval, and Bioinformatics. There are two types of clustering algorithms: hierarchical clustering and partitional clustering. Today, most of scientific and industrial applications continuously generate data in massive scale and extracting valuable information from those dynamic data sources is a prominent task in the today's knowledge intensive world. Nevertheless, most of the well-established clustering algorithms assume that the underlying cluster structure of the dataset does not change with time. In order to fulfil the new requirement of real-time information extraction, dynamic data clustering has emerged as a research interest attached to the field of dynamic data mining.

Inherent capability of dealing with dynamic environments in Multi Agent technology can be used to address the need of identification of clusters in a dynamic data source. The proposed solution is based on the ability of modelling dynamic and complex systems by Multi Agent technology. The proposed solution consists of two main types of agents, namely, data record agent representing a single data record in the dataset and cluster agent representing the emerging clusters through communication and negotiation among data record agents. The cluster formation over dynamic data source is emerged as an outcome of interactions and negotiations among agents in the proposed system.

The rest of the paper is organized as follows; section 2 gives an overview of current trends in dynamic data clustering, the section 3 presents the introduction of multi agent technology and its role in data clustering. The proposed method is presented in Section 4 and subsequent section provides details on the implementation of the proposed method. Section 6 presents the results of evaluation of the proposed method with artificially created datasets. The paper concludes with the discussion of the outcomes of the study and suggestions for further works.

II. STATE OF THE ART : DYNAMIC DATA CLUSTERING

Most of the classical clustering algorithms are assumed that the dataset and criteria for assigning data to clusters remain fixed during the execution of the algorithm. Hence, those algorithms fail to identify the cluster structure of today's data sources which change with time. As a result, the identification of clusters in a dynamic data set has become a new challenge in both data mining and machine learning research fields.

In the classical clustering, both data records and clusters are considered as static. However, in the context of dynamic clustering, there are two possibilities that make the dataset dynamic: (1) arrival of new data into a dataset and the deletion of existing data in the dataset, (2) fixed number of data records have trajectory of feature values instead of a single feature vector. In both cases the underlying cluster structure of the dataset can be changed over the time in terms of number of clusters, shapes and positions of clusters. The customer base of a supermarket is an example of first type dynamic data source. Clustering monthly household electricity usage of a given area can be considered as a dynamic data clustering task with dynamic data records, since monthly electricity usage of a household should be represented as a trajectory of feature values instead of a feature vector.

Incremental K-Means [1] and dynamic fuzzy c-means [2] clustering can be considered as an extension to static data clustering for dynamic data clustering. Adjusts the model that has been built, according to the new data whenever they arrive to the system and periodically updates the model to capture

978-1-4799-1276-6/13/$31.00 ©2013 IEEE

the changes in the dataset through learning cycles are the main approaches in those extensions. Fuzzy logic and rough set are used in dynamic data clustering to capture uncertainties in dynamic environments [3]. Apart from that novel methods also have been reported in the literature: ClusTree[4], Dynamic hierarchical compact and dynamic hierarchical star algorithms[5] .

Chakraborty and Nagwani [1] presented an incremental version of K-mean clustering which can be applied to a dynamic database where the data may be frequently updated. The proposed algorithm, obtains the new cluster centres by combing the new data with the means of the existing clusters instead of rerunning the K-means algorithm from the scratch. Experimental results indicate that this algorithm is outperformed when the number of clusters increased, the length of the cluster radius decreased while new data objects are inserted into the existing database. Thus, the removal of exiting data records has not been taken into an account in this study. Because it may be due to the authors mainly concentrated on incremental learning paradigm. The results of the algorithm at the later stages are still based on the initial cluster result of this algorithm. This is another limitation of this algorithm.

Prior knowledge on the number of clusters in the dataset is a primary question in static data clustering and also in dynamic data clustering where the number of clusters changes with time. Tasoulis and Vrahatis[6] extended the static k-window cluster to identify clusters in dynamic databases. The proposed dynamic version of the unsupervised k-windows algorithm utilizes the Bkd-tree structure. It is assumed that the static unsupervised k-windows algorithm has been applied on the initial database, producing a set of windows that describe the clustering result. Then the dynamic algorithm periodically yields a clustering model that adapts to the changes in the database: insertion and deletion of records.

Gil-García and Pons-Porrata [5] proposed two clustering algorithms namely dynamic hierarchical compact and dynamic hierarchical star. Those are aimed to construct a cluster hierarchy for the cluster structure of dynamic data sets. Both of algorithms are mainly focused on text document clustering and the first one creates disjoint hierarchies of clusters, while the second obtains overlapped hierarchies of clusters. Wai-chiu and Wai-chee Fu [7] presented an incremental hierarchical clustering algorithm for web document clustering. It employs a new tree structure called Document Clustering tree (DC-tree) to update the cluster hierarchy as a new document arrives.

Kranen and et.al [4] proposed a parameter-free hierarchical cluster algorithm called ClusTree that automatically adapts to the speed of the data stream. This algorithm is mainly based on the micro-cluster, a popular technique used in the data stream clustering. In this study, the data distribution is represented as micro-clusters and those micro-clusters are stored in the proposed index structure. The proposed index structure is capable of adapting its size according to the speed of the data stream. Hence this algorithm is capable of capturing the data structure of the data stream without dropping any data point in the stream.

Mary and Kumar [8] proposed a density based dynamic data clustering algorithm for clustering incremental datasets and they addressed the problems of clustering a dynamic dataset in which the dataset increases in size over the time by adding more and more data. The comparison of the proposed Incremental DBSCAN with Chameleon and DBSCAN clustering indicates that it performed significantly well in terms of clustering accuracy as well as speed [8].

Fuzzy set and rough set have been successfully employed in dynamic clustering to handle uncertainty associated with the dynamic data source[9]. Both Dynamic Fuzzy c-means [2] and Dynamic Rough k-means [10] use a common updating cycle called "Dynamic Clustering cycle" to capture the changes in the dynamic dataset. The Dynamic Clustering cycle consists of four steps : update or initially set the parameters of cluster algorithm, performs the cluster algorithm, classify new data and according to the results of classification check the structural changes.

In the Dynamic Fuzzy c-mean [2] algorithm proposed by Crespo and Weber, the adaptation of the existing cluster model to the new data is handled by three strategies: create a new cluster, move the existing clusters and delete clusters. The length of the period (update cycle) depends on the particular application. The experimental results show that this can be successfully applied in customer segmentation and traffic management. Nevertheless, the results in the later phases of the algorithm depend on the initial parameters. It can be seen as a limitation of this method.

Dynamic Rough k-means [10] algorithm, proposed by Peters and Weber used the same structure of the update cycle used in Dynamic Fuzzy c-mean. However, it keeps the uncertainty of the cluster configuration stable instead of moving clusters. That is the proportion of the objects with sure memberships (members of lower approximations) and the ones with unclear memberships (objects in boundary regions) [3].

Apart from the above approaches, use of Multi agent technology is another possible approach to identify clusters in dynamic data sources. The next section discusses the role of Multi agent technology in data clustering and some of the reported works in the literature.

III. Multi Agent Technology In Data Clustering

Multi Agent technology is another newly developing field of Artificial Intelligence mainly focused on developing group of small and intelligent software component work together to solve problems in complex and dynamic systems. Multi Agent technology has evolved from the field of Distributed AI [11]. Autonomous behaviour of distributed agents and the capability of communication with each other are the main potentials behind the success of the many applications of Multi Agent technology.

There are several advantages in using Multi Agent technology for dynamic data clustering. Reactive and proactive behaviour of agents can be used to decrease the response time : time taken by the algorithm to detect changes in the data source and updates the cluster configuration according to the changes. Capability of distributed problem solving in the Multi Agent system can be used to increase the efficiency of the particular method by distributing computation among the agents. In traditional parallel and

distributed data mining techniques need a centralized master process to control processes that direct data mining tasks. By using Multi Agent technology, this issue can be avoided [12].

Extracting valuable information from dynamic data sources is another potential application of Multi Agent technology. Hence in the last decade Multi Agent Data Mining (MADM) also known as Agent-based Data Mining was emerged as a main branch of Agent-Mining interaction and integration (AMII) research field [13]. Consequently, in the past decade Multi Agent Data Mining (MADM) has become a new research interest in both Data mining and Multi Agent research communities, providing potential solutions for dynamic data mining. In MADM, group of agents performs mining tasks in a cooperative manner in order to extract information from data sources. Originally, MADM aimed at extracting information from large scaled data which were originally distributed, using distributed and parallel processing power inherent in Multi Agent technology. In late 1990s, PADMA [14] and PAPYRUS [15] are two reported Multi Agent clustering systems which were developed to integrate the knowledge discovered at different sites, with a minimum amount of network communication and a maximum amount of local computation.

Moreover, several studies have been conducted in a way such that a group of agents mines information using different clustering algorithms and exchange the mined knowledge among them. In study [16], Chaimontree, Atkinson, and Coenen developed Multi Agent based clustering system which selects the best set of clusters for a given dataset based on result extracted knowledge of agents using three different clustering algorithms: K-means, K-Nearest Neighbour (KNN) and DBSCAN.

Chaimontree, Atkinson, Coenen proposed a Multi Agent based clustering mechanism where each individual cluster is represented by an agent [12]. The proposed Clustering mechanism consists of two phases, Bidding phase and Refinement phase. In the Bidding Phasé, an each clustering agent bid for records where the Data Agent: the "owners" of data sources, acts as an auctioneer. Then in the Refinement phase, Clustering agents act as local auctioneers and try to sell unwanted records to other clustering agents based on bids placed by them.

In literature, there were many attempts to develop agent based clustering algorithms but there is lesser number of reported works on an agent based clustering systems which represent data records as agents and enhance cluster configurations through intra-agent interactions and negotiations. Rzevski et.al proposed a method [17] which data records and clusters are represented as agents. In this method, agents represent data records; pro-actively search for suitable clusters while clusters change their members (data records) in real time whenever a new data element arrives.

Kiselev and Alhajj [18] modelled the task of online unsupervised learning as a dynamic distributed resource allocation problem. They proposed market-based negotiation between different self-interested agents in order to satisfy a distributed constraint network and achieve an implicit global solution to the problem. This work addressed the issue of online agglomerative hierarchical clustering of streaming data in complex dynamic environments under conditions of uncertainty. This proposed solution is appropriate for various scenarios where the data processing is time critical: run-time detection of previously unknown dispatching rules and effective scheduling policies in transportation scheduling.

IV. Multi Agent Approach for Dynamic Partitional Clustering

According to the literature, only few attempts were made to develop agent based solutions for dynamic partitional clustering. Partitional clustering is a more complex task than hierarchical clustering due to various reasons. For instance, the number of possible clusters is a crucial matter for this purpose, as it can change over the time in a dynamically changing data sources.

The proposed Multi Agent approach (Agent Based Dynamic Clustering - ABDC) in this paper identifies clusters in dynamic data source, by dividing the data into partitions. It is based on the hypothesis that the underlying cluster structure in a dynamic data source can be identified by interactions and negotiation between agents who represent data records and identified clusters in that data source. In addition to that, reactive and proactive behaviours of agents are used to identify the changes in the underlying cluster structure of the dynamic data source in an effective and efficient manner.

A. Proposed Multi Agent System

The proposed solution mainly consists of two parts: the Multi Agent environment and data retrieval component. The Multi Agent environment provides the platform to agents to live and interact with each other to discover the cluster structure of the dataset they represent. To retrieve the data from the data sources, a separate agent is assigned to each data source as shown in the Figure. 1. Data Source agent reads data records from the data source and creates a Data Record agent in the Multi Agent environment, for each data record it reads.

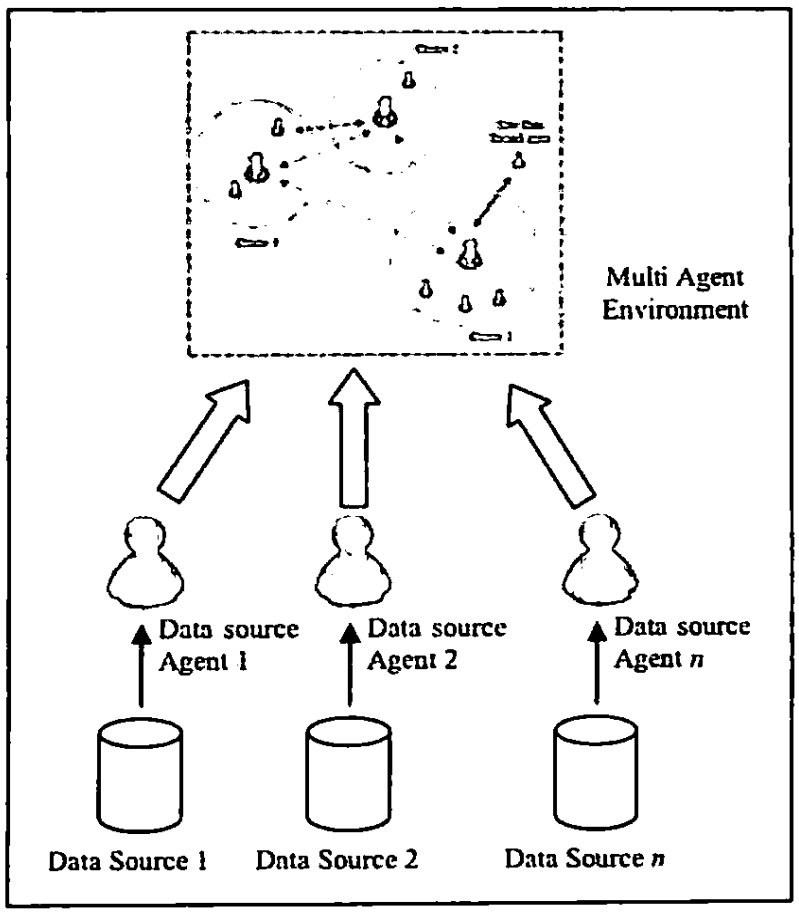

Figure 1: Top Level Architecture of the Proposed System

B. Proposed Multi Agent Environment

The Proposed Multi Agent environment in this study consists of two main types of agents, namely Data Record agent and Cluster agent. During the process of cluster identification, the Data Source agent can add new Data Record agents and exiting Data Record agents are allowed to leave the system. Hence, this method is capable of identifying the partitional clusters in a dynamic data source which new records are added and existing records are deleted over the time.

As illustrated in Figure. 2, three types of interactions are proposed for : (1) newly created Data record agent to get an initial membership from its nearest cluster among the existing clusters (2) existing data record agent to move to the nearest cluster (3) Inter-cluster merging negotiation between two neighbouring clusters.

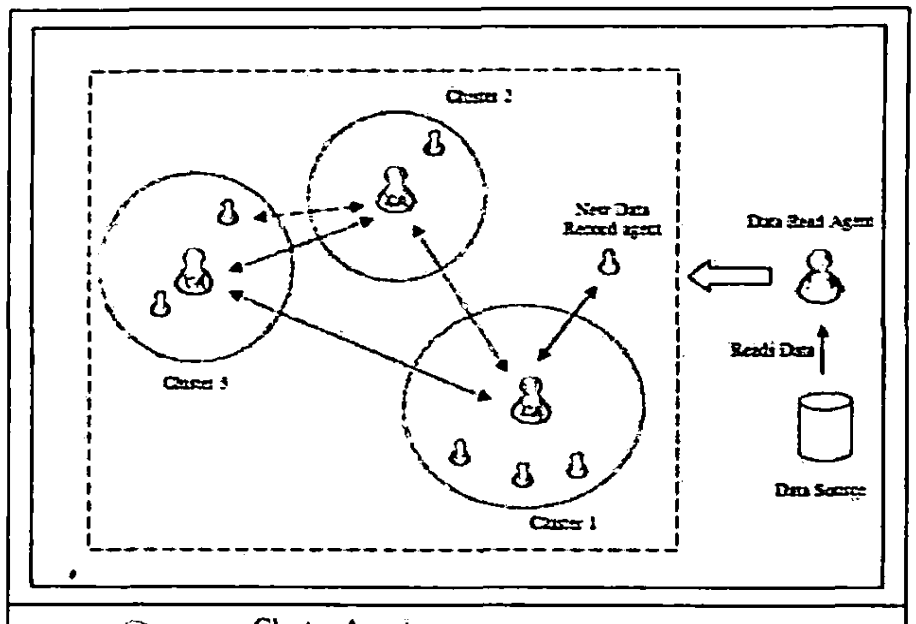

Figure 2 : Proposed Multi Agent Environment

C. Data Record Agent

The data record agent represents a single Data Record in the dataset and its main objective is getting membership of its nearest cluster. In order to achieve this goal, the Data Record agent executes two plans : (1) getting the initial membership (2) moving to its nearest Cluster. Execution details of those plans are explained in the following section: Proposed Cluster Identification

D. Cluster Agent

The cluster agent represents emerging clusters through communication and negotiation between data record agents. Cluster agents always try to minimise their cluster with-in variance. Hence, the cluster agent does not offer the membership to data records who increase the cluster with-in variance of the cluster. In addition to this, cluster agents merge with their neighbouring cluster to improve the overall cluster configuration.` the next section describers role of each agent to identify the best cluster configuration for a given data set.

E. Proposed Cluster Identification Process

Initially, Data record agents try to join with existing clusters based on the distance to mean of each cluster. A newly created data record agent sends initial membership request to the nearest cluster agent. Cluster Agent accepts the membership request of the Data record agent only if it does not increase cluster with-in variance (Sum of squared error). In the absence of a suitable cluster, the Data record agent is allowed to create a new cluster including itself.

Each Data Record agent continuously updates its beliefs (known facts) about the environment. Based on those updated beliefs, it identifies the emergence or the existence of the more suitable cluster than the cluster it currently belongs. The data record agent sends membership request to any cluster if its meaning is closer than the mean of the cluster it belongs. In this case, the nearest cluster agent accepts the request and offers the membership to that data record agent without considering any selection criteria. By moving to the nearest clusters, Data record agents continuously improve the quality of the existing cluster configuration.

By adding newly created Data Record agents or accepting membership request from existing Data Record agents in other clusters, Cluster agent can expand its cluster. However, there can be two clusters which are too close together and which can be merged to produce a better cluster configuration. Hence, Cluster agents are allowed to merge with their neighbouring cluster to improve the overall cluster configuration.

Cluster agents perceive their environment to update their beliefs on the emergence or the existence of neighbouring clusters. Using those updated facts Cluster agents interact with their nearest cluster to evaluate the possibility of improving the overall cluster configuration by merging each other. In this negotiation process, internal cluster validation measures are used to make the decision of merging. In this method Silhouette coefficient is used as the internal cluster validation measure.

Silhouette coefficient of i^{th} data point is given by

$$s_i = \frac{b_i - a_i}{\max(a_i, b_i)}$$

where , a_i is the average intra-cluster distance of the i^{th} data point to all other data points within its cluster, and b_i is the minimum of the average inter-cluster distances of the i^{th} point to all data points in each other cluster.

As a system, this enhancement process continues until the internal cluster validity measure on both coherence and separation of clusters stabilizes at a maximum value.

V. Implementation

The agent based Dynamic Clustering system has been implemented using Jadex [19] agent development framework. Jade is a Java based agent development framework with a Belief-Desire-Intention reasoning engine which supports to

develop rational software agents. The Jadex BDI reasoning engine allows the development of rational agents using mental notions in the form of beliefs, goals and plans. Beliefs capture informational attitudes, desires motivational attitudes and intentions deliberative attitudes of agents [20].

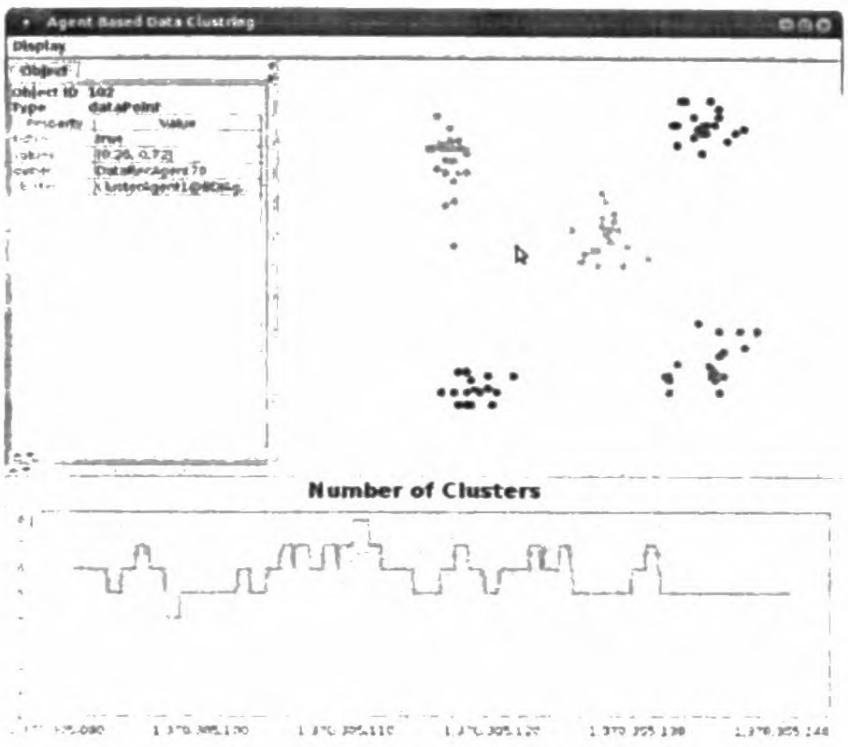

Figure 3 : Visualization of the cluster result

The visualization of the Multi Agent environment provided by the Jadex agent platform is used to visualize the result of the clustering process. Visualization of a final cluster result of an artificial dataset is shown in the Figure 3.

VI. Evaluation

Artificially generated data from Multinomial distribution were used to evaluate the proposed method. According to the results of the evaluation, the proposed method successfully identified moderately separated clusters in 2-dimensional space. In 3-dimensional space, the proposed method successfully identifies non-spherical clusters. Cluster result of an artificial 3D data set is shown in the Figure 4. As shown in the figure the proposed method successfully identifies clusters in different shapes which are closely located with each other.

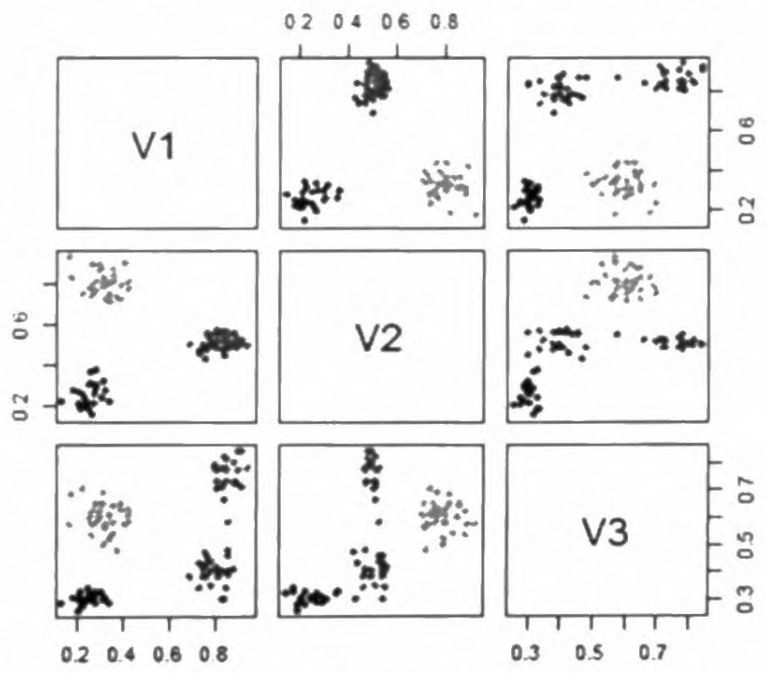

Figure 4 : Scatter plot matrix of an artificial 3D data set

The maximum Silhouette coefficient obtained by the proposed method and the K-mean algorithm is compared in the Table 1.

Dimension	Number of Data points	Number of clusters	Silhouette Coefficient	
			ABDC	K-means
2	65	3	0.8013	0.8013
	90	4	0.7903	0.7903
	100	5	0.7839	0.7839
3	70	3	0.7835	0.7835
	105	4	0.7620	0.7620
	125	5	0.7243	0.7243

Table 1: Comparison cluster results of the proposed method and k-means algorithm

According the result of the evaluation of the proposed method with artificial data, it performs as k-mean algorithm in both 2 and 3 dimensional spaces without any initial parameter. In some cases, particularly in 2 dimensional spaces the proposed method is trapped in local maxima of the internal evaluation measures. Hence it fails to identify the actual number of clusters in the dataset. However, it produces considerably good cluster configuration as the clustering result produced by the K-mean algorithm with the number of clusters suggested by the proposed method. The Figure 5. shows example of such a situation.

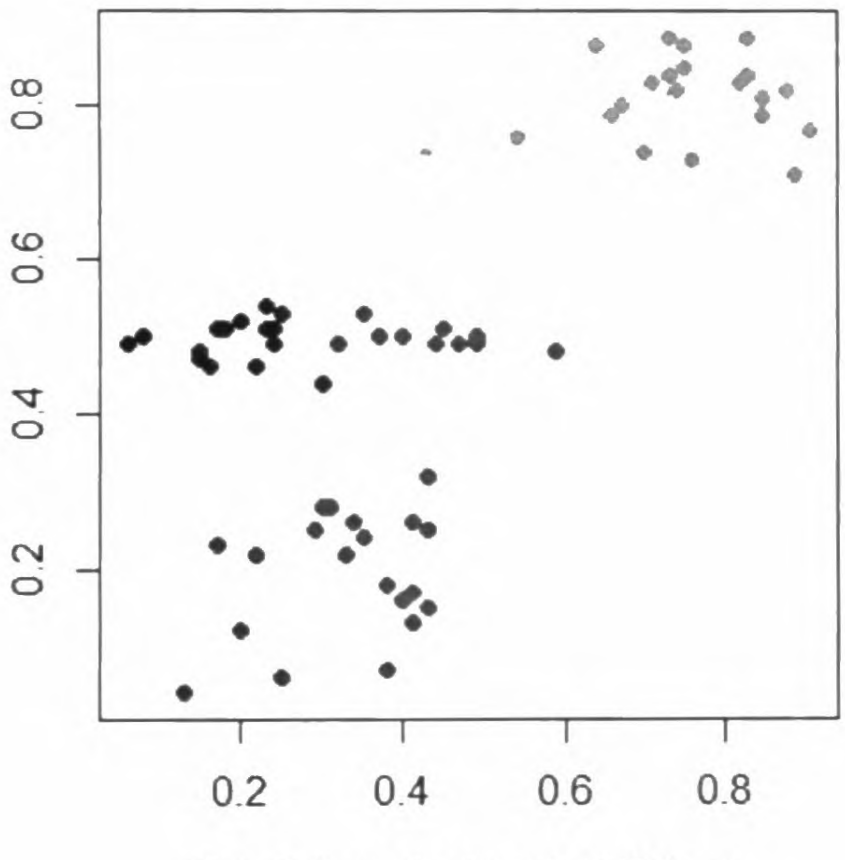

Figure 5 : Scatter plot of an artificial 2D data set

VII. Discussion and Further work

According to the result of this study, Multi Agent technology can be used to address the need of identification of partitional clusters in a dynamic data source. In addition to that, the proposed method has three advantages over k-mean algorithm. Firstly, the number of clusters is not required as an initial parameter. This addresses a classical problem with static data clustering. The second feature can be stated as the

ability to restructure the cluster configuration upon the arrival of new data from the data source. This restructuring could be seen as deleting an existing cluster, forming a new cluster, and/or merging two clusters. As the third advantage, the use of multi agent systems technology for dynamic data clustering drastically reduces the human intervention in the clustering process.

The internal evaluation measure: Silhouette Coefficient used in this method is computationally expensive. The scalability of the proposed method can be improved by developing less complex internal evaluation measure to use in the cluster merging process. Since, the proposed method allows to Data record agents to enter and leave the system during the cluster identification process, this method can be further improved to identify the partitional clusters in a data stream.

REFERENCES

[1] S. Chakraborty and N. K. Nagwani, "Analysis and Study of Incremental K-Means Clustering Algorithm," in High Performance Architecture and Grid Computing. vol. 169, A. Mantri, S. Nandi, G. Kumar, and S. Kumar, Eds., ed: Springer Berlin Heidelberg, 2011, pp. 338-341.

[2] F. Crespo and R. Weber, "A methodology for dynamic data mining based on fuzzy clustering." Fuzzy Sets and Systems, vol. 150, pp. 267-284, 3/1/ 2005.

[3] G. Peters, R. Weber, and F. Crespo, "Uncertainty modeling in dynamic clustering - A soft computing perspective," in *Proc.* IEEE International Conference on Fuzzy Systems (FUZZ) ,pp. 1-6 ,2010.

[4] P. Kranen, I. Assent, C. Baldauf, and T. Seidl, "The ClusTree: indexing micro-clusters for anytime stream mining," Knowledge and Information Systems, vol. 29, pp. 249-272, 2011.

[5] R. Gil-Garc and A. Pons-Porrata, "Dynamic hierarchical algorithms for document clustering," Pattern Recogn. Lett., vol. 31, pp. 469-477, 2010.

[6] D. K. Tasoulis and M. N. Vrahatis, "Unsupervised clustering on dynamic databases," Pattern Recogn. Lett., vol. 26, pp. 2116-2127, 2005.

[7] W. Wong and A. Fu, "Incremental Document Clustering for Web Page Classification," 2000,

[8] S. A. L. Mary and K. R. S. Kumar, "A Density Based Dynamic Data Clustering Algorithm based on Incremental Dataset," Journal of Computer Science, vol. 8, pp. 656-664.

[9] G. Peters and R. Weber, "Dynamic clustering with soft computing," WIREs Data Mining and Knowledge Discovery vol. 2, pp. 226-236, May/June 2012

[10] G. Peters and R. Weber, "Intelligent cluster algorithms for changing data structures," International Journal of Intelligent Defence Support Systems, vol. 2, pp. 105-119, 2009.

[11] Gerhard Weiss, Multiagent systems: a modern approach to distributed artificial intelligence, MIT Press, Cambridge, 1999

[12] S. Chaimontree, K. Atkinson, and F. Coenen, "A Multi-agent Based Approach to Clustering: Harnessing the Power of Agents," in Agents and Data Mining Interaction. vol. 7103, L. Cao, A. C. Bazzan, A. Symeonidis, V. Gorodetsky, G. Weiss, and P. Yu, Eds., ed: Springer Berlin Heidelberg, 2012, pp. 16-29.

[13] C. Ralha, "Towards the Integration of Multiagent Applications and Data Mining," in Data Mining and Multi-agent Integration, L. Cao, Ed., ed: Springer US, 2009, pp. 37-46.

[14] H. Kargupta, I. Hamzaoglu, and B. Stafford, "Scalable, distributed data mining using an agent based architecture," in *Proc.* Third International Conference on the Knowledge Discovery and Data Mining, 1997, pp. 211–214.

[15] S. Bailey, R. Grossman, H. Sivakumar, and A. Turinsky, "Papyrus: a system for data mining over local and wide area clusters and super-clusters," in *Proc.* 1999 ACM/IEEE conference on Supercomputing (CDROM), Portland, Oregon, USA, 1999.

[16] S. Chaimontree, K. Atkinson, and F. Coenen, "Multi-agent based clustering: towards generic multi-agent data mining," in *Proc.* Industrial conference on Advances in data mining: applications and theoretical aspects, Berlin, Germany, 2010.

[17] G. Rzevski, P. Skobelev, I. Minakov, and S. Volman, "Dynamic pattern discovery using multi-agent technology," in *Proc.* of the 6th WSEAS Int., Dallas, Texas, 2007.

[18] I. Kiselev and R. Alhajj, "A Self-organizing Multi-agent System for Adaptive Continuous Unsupervised Learning in Complex Uncertain Environments," in *Proc.* Twenty-Third AAAI Conference on Artificial Intelligence, 2008, pp. 1808-1809.

[19] L. Braubach, A. Pokahr, and W. Lamersdorf, "Jadex: A Short Overview," in Main Conference Net.ObjectDays 2004, 2004, pp. 195-207.

[20] A. S. Rao and M. P. Georgeff, "BDI-agents: from theory to practice," in *Proc.* the First Intl. Conference on Multiagent Systems,