J. Appl. Environ. Biol. Sci., 4(8S)160-165, 2014

ISSN: 2090-4274
Journal of Applied Environmental and Biological Sciences
www.textroad.com

The Relevancy of a Unified Data Mining Theory (UDMT) For the Big Data

Dost Muhammad Khan[1] and Nawaz Mohamudally[2]

[1] Department of Computer Science & IT, the Islamia University of Bahawalpur, PAKISTAN
[2] School of Innovative Technologies and Engineering (SITE), University of Technology, MAURITIUS,

Received: September 1, 2014
Accepted: November 13, 2014

ABSTRACT

For decades, the companies are making their business decisions based on the structured data stored in relational databases but now the trend is to mine the unstructured data such as web-logs, social media, email, images and graphics for useful information. There is a need to formulate a unified data mining theory (UDMT) to address the fundamental question of discovery of knowledge from the big data. It is a fact that data mining is not a single step process and knowledge discovery is the result of successive processes; therefore, we have to unify different data mining tasks. In the proposed UDMT the knowledge is extracted from the given data through the unification of the data mining processes; clustering, classification and visualization. In this paper we try to prove the relevancy of the UDMT for the big data.
KEYWORDS: Unified Data Mining Theory (UDMT), Big Data, MAS, UDMP

INTRODUCTION

Big data is an exponential growth, availability and use of data both in structured and unstructured way which can serve as the basis for innovation. According to IDC, it is important that the organizations and IT professionals must focus on the ever-increasing volume, variety, variability, complexity and speed of information that forms the big data. According to Gartner, "*The velocity and speed means how fast the data is being produced and how fast the data must be processed to meet the demands*". As per Gartner's assessment, the term big data is relative and it applies whenever an organization's ability to handle, store and analyze data exceeds its current capacity. According to Scott Zucker of Family Dollar, "*Small data is gone. Data is just going to get bigger and bigger and bigger, and people just have to think differently about how they manage it*". We are now in the era of big data, It is not an issue that we have large amount of data, the issue is that how to use it to make the best decisions. Now days many tools and technologies are available to collect and store big data [1-7]. The following are the usages of the big data:

i. Analyze to determine optimal prices for the profit.
ii. Mine customer data for insights that drive new strategies for customer acquisition, retention, campaign optimization and next best offers.
iii. Quickly identification of customers who matter the most.
iv. Analysis of data from social media to detect new market trends and changes in demand.
v. Determine root causes of failures, issues and defects by investigating user sessions, network logs and machine sensors [1-7].

There are many challenges to handle the big data. We enumerate some of them.

i. What if the data volume gets so large and varied?
ii. How to deal with the big data?
iii. Do we store all our data?
iv. Do we analyze it all?
v. How we can find out which data points are really important?
vi. How we can use it to get the best advantage [1-7]?

The focus of this study is on the question how we can use the big data to obtain its best advantages and to find the vital data points which play significant role in the discovery of knowledge. According to Dan Briody, *"Most businesses have made slow progress in extracting value from big data. And some companies attempt to use traditional data management practices on big data, only to learn that the old rules no longer apply"*. The following technologies can be used to process and make available the big data:

i. Faster processors.
ii. Parallel processing, clustering, virtualization and many more.
iii. Cloud computing [1-7].

After the extensive study it is concluded that these tools and technologies can easily handle the big and large datasets but are inefficient to discover the knowledge. The discovery of the knowledge from a dataset big, medium or small is a multi-step process where the output of one process must be the input of other processes; therefore, a unified theoretical framework for data mining which unifies data mining processes such as clustering, classification and visualization is required. Hence a unified data mining theory is required [9-15].

The rest of the paper is organized as follows: in section 2 we present a Unified Data Mining Theory (UDMT), section 3 is about a Unified Data Mining Server (UDM Server), section 4 is about Methodology and results are discussed in section 5 and finally the conclusion is drawn in section 6.

*** Corresponding Author:** Dost Muhammad Khan, Department of Computer Science & IT, The Islamia University of Bahawalpur, PAKISTAN. khan.dostkhan@iub.edu.pk

1. Unified Data Mining Theory (UDMT)

The proposed UDMT is given below:

Step 1: Create (appropriate) partitions of the dataset.

Step 2: Create the clusters of each partition. This step is also called the clustering. The clustering is a technique of dividing a dataset into different groups. The goal of clustering is to find groups that are very different from each other and whose members are very similar to each other.

Step 3: Construct the 'decision rules' (in the form of if-then statements) of each clusters. This step is also referred as the classification. The classification is a technique of placing an object into group based on common properties among the objects.

Step 4: Plot the 2D or 3D graphs of each rule or classifier. This step is also known as visualization. The visualization is a process of presenting data in a special and easy to understandable form. It is also a relationship within the data which is not evident from the raw data.

Finally, after the interpretation and evaluation of 2D graphs, the 'knowledge' is extracted. The foundation of the proposed UDMT is that without clustering there is no classification, without classification there is no visualization and hence without visualization there is no 'knowledge'. The 'knowledge' can only be produced through the unified process of clustering, classification and visualization [9-15].

Proposition: Our proposition is: 'Knowledge' is an intersection of clustering, classification and visualization. More precisely we can say that 'Knowledge' can only be extracted from the given dataset after clustering, classification and visualization processes of data mining as it is given in eq. (1).

$$\text{Clustering} \rightarrow \text{Classification} \rightarrow \text{Visualization} = \text{Knowledge} \quad (1)$$

Another way to write is given in eq. (2)

$$A \rightarrow B \rightarrow C = K \quad (2)$$

Where A is clustering, B is classification, C is visualization and K is knowledge.

Proof: The "Implies (→)" which is actually $A \rightarrow B = A^{-} \cup B$. The truth table is shown in Table 1.

Table 1. Truth Table

Case #	A	B	C	K
1	0	0	0	**0**
2	0	0	1	**1**
3	0	1	0	**0**
4	0	1	1	**1**
5	1	0	0	**1**
6	1	0	1	**1**
7	1	1	0	**0**
8	**1**	**1**	**1**	**1**

Explanation of Table 1: The output is high only if visualization is high in case 2, which is against the proposition. The output is high in cases 4, 5 and 6 either classification and visualization are high, only clustering is high or clustering and visualization are high, again it is against the proposition. The proposition is correct in cases 1, 3, 7 and 8. It is clear from these results that "Implies (→)" does not hold in the case of the proposed unified data mining processes. This is true in only one case when all A, B and C are 1. If the "Implies (→)" fails then it is not possible to prove the proposed UDMT using the set theory.

Therefore, we apply the mathematical functions approach for the proposed unified data mining processes. In [16] the proposed theory is discussed and is illustrated with different data mining algorithms and datasets. Furthermore, in [17-18] different unified data mining theories and their techniques are discussed.

In summary we can say that the proposed unified data mining theory is: small, simple and structured i.e. first step is clustering followed by classification and visualization respectively where the output of one step is an input of the next step. In other words, it is a single comprehensive framework which contains all the data mining tasks. It is simple, easy and the extraction of knowledge is straightforward and provides knowledge directly in the form of 2D graphs. It is flexible to add new data mining tasks.

1.2 The Limitations of the UDMT. In this section we will highlight some of the limitations of the proposed UDMT.

1. Scalability. The scalability is measured in terms of number of attributes, number classes and sample size in a dataset. The scalability of UDMT is discussed: the first step is to create the all possible partitions of the given dataset. If a dataset has large number of attributes then its number partitions will increase, it is difficult to handle these partitions. They will take a lot of space and it will also affect the processing of the system. The next step is to create the clusters of each partition. There are different approaches to optimize the number of clusters of a dataset and number of clusters will vary with each approach. Again it will increase the number of clusters of the dataset which will take a lot of space and also effect the processing of the system. Finally, as the number of clusters increases, the number of 2D graphs will also increase and they will take a lot space. This is called the scalability.

This limitation can be reduced by applying some selection and evaluation criteria on each cluster. Select the useful clusters and their respect 2D graphs and discard the remaining. In this way we can minimize the issue of scalability of the proposed UDMT. It is obvious that all created clusters are not useful for knowledge extraction.

2. Sequence of Data Mining Tasks. The sequence of the data mining tasks is: first apply the clustering then the classification and finally the visualization. The same sequence must be followed; if the order is not conformed then the theory will not work. This is a sequential way of knowledge extraction.

Remark: It is an important point that there is no limit on the sample size of the given dataset i.e. the UDMT supports the big datasets.

2.2 Technique used to Minimize the Limitations. We address the above mentioned limitations of the UDMT by using the Model Selection Criterion, which helps to test that the dataset is rightly fitted for data mining algorithm(s). If the right dataset (model) is not selected there is risk in deteriorated knowledge or no knowledge at all. The Model Selection Criterion will set a limit on number of attributes and number of clusters in a dataset, hence we avoids the above mentioned limitations.

Furthermore, the value of Model Selection Criterion can be applied to select the suitable and appropriate data mining algorithm for the given dataset by mapping its value with the logarithm value of complexities O of data mining algorithms.

We develop a unified data mining tool (UDMTool), a Multiagent System (MAS), the tool can also be assimilated to a unified data mining server (UDM Server). The UDM Server is based on a Unified Data Mining Theory (UDMT) and all the limitations are properly addressed. The next section is about the UDM Server.

3. A Unified Data Mining Server (UDM Server). A Unified Data Mining Server (UDM Server) is a new and better next generation solution. It is a unified way of architecting and building software solutions by integrating different data mining tasks. It can be deployed to find new insights and capitalize on hidden relationships and helps to analyze the big data. The architecture of the UDM Server is demonstrated in Figure 1.

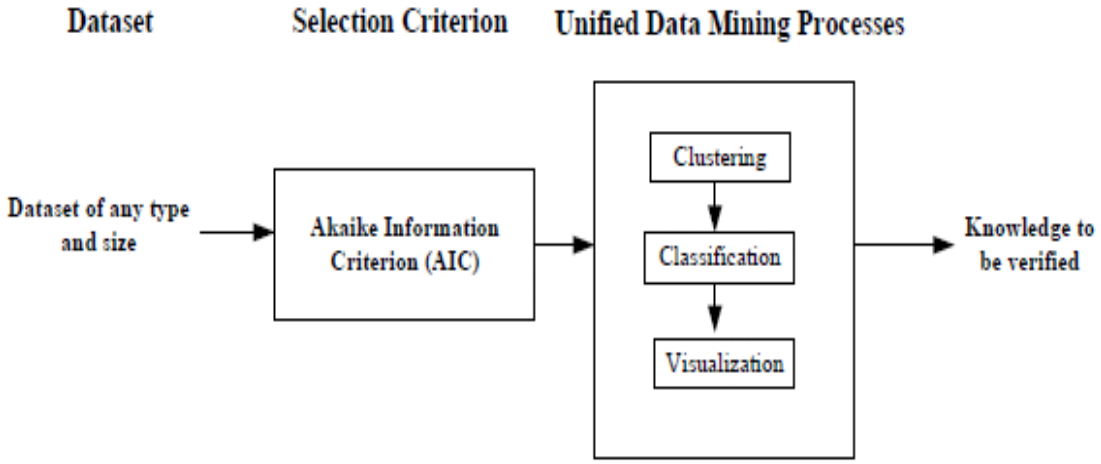

Fig. 1. The Architecture of the UDM Server

Explanation of the architecture of Figure 1: The dataset of any data type is the required input. The model selection criterion AIC is applied to determine the fitness of the dataset and the value of AIC is used to choose the appropriate data mining algorithm at each data mining process. A unified data mining processes (UDMP) unifies clustering, classification and visualization and produces the domain of knowledge. Finally, after the interpretation and evaluation the 'knowledge' is selected which is verified by the user. Dataset, Model Selection Criterion, UDMP and Knowledge are main components of UDM Server [9-15]. We illustrate them more precisely:

Let Dataset D = {Numeric, Multimedia, Text, Categorical}

The Model Selection Criterion S = AIC

The output of S = {Over-fitted, Under-fitted, Best-fitted}

The Unified Data Mining Processes U = {Clustering, Classification, Visualization} [The 'clustering' is the first step followed by the rest of the steps]. Followed by the 'interpretation' and extract the 'knowledge'.

Knowledge K={Accepted, Rejected} ['Accepted' means that the required results are according to the business goals and 'Rejected' means that the output is not within the domain of the business goals. The 'knowledge' will be verified by the user, the Server (Model) cannot play any role in this regard]

The UDM Server is tested only on numeric dataset i.e. a data file. The value of AIC the model selection criterion is computed. If the dataset is under or over-fitted then the dataset requires the cleansing. Only the best-fitted dataset is accepted as an input [19-20]. The next step is to create the partitions (vertical) of the dataset. Suppose there are 'n' attributes in a dataset and we want to create 'm' partitions. The formula is given below in eq. (3).

$$m = \frac{n-1}{2} \qquad (3)$$

Where $0 < n < m$, round the value of 'm' and the each partition contains two attributes along with the class or the target attribute, therefore one is subtracted from 'n'. The selection of attributes depends on the user. Since the clustering is a distance based process; therefore, the attributes of a partition must be selected in such a way that they have an equal

impact on the distance computation. In this way we can avoid to obtain the clusters that are dominated by the attribute of high value. These partitions become the input to UDMP which produces the domain of knowledge. Finally, the knowledge is extracted which is either accepted or rejected [9-15]. Figure 2 demonstrates the knowledge extraction process in the UDM Server.

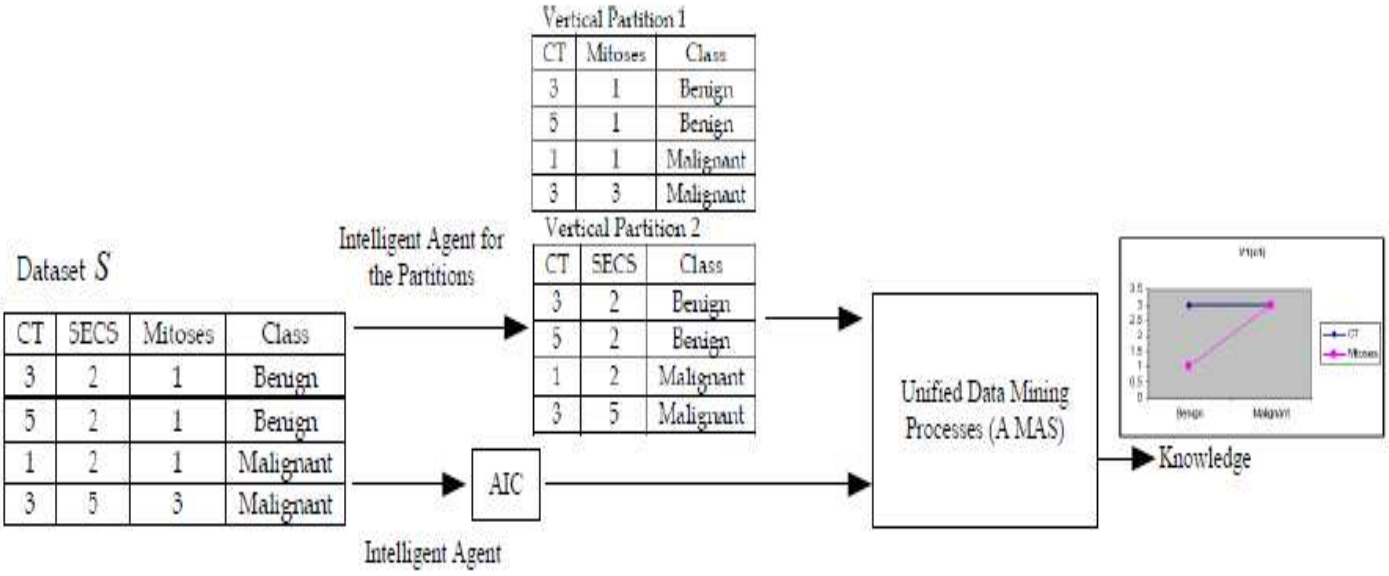

Fig. 2. The Function of the UDM Server

The UDM Server is a Multiagent System (MAS), the agent takes the dataset and computes the value of AIC, other agent creates the appropriate partitions of the dataset according to eq. (1) and the logarithm value of the complexities O of data mining algorithms is computed through an agent. Another agent is used to input the partitions of the dataset to UDMP, which itself is a MAS, where each data mining task is accomplished through the agent [21]. The appropriate algorithm is selected at each data mining task through the value of AIC of the given dataset, the process is completed by an agent which maps the value of AIC with the logarithmic value of the complexities O of data mining algorithms. The 'knowledge' is extracted after interpretation and evaluation as a final output. Due to the Model Selection Criterion AIC, the UDM Server supports up to 23 numbers of parameters and 211 numbers of attributes in a dataset and there is no limit on the sample size of the dataset, which shows that the UDM Server has a capacity to handle large and big datasets [9-15]. Table 2 summaries the agents and their functions in UDM Server.

Table 2. Agents and their Functions in UDM Server

Sr. #	The Function of Agent
1	Computes the value of AIC
2	Creates the partitions
3	Computes the complexities 'O' of Data Mining Algorithms
4	Creates the clusters
5	Creates the rules/classifiers
6	Plots the 2D graphs
7	Maps the value of AIC with Log (O)
8	Provides the partitions as input to UDMP

The future works also known as the enhancement in the UDM Server:

i. Other data types for example categorical, multimedia, text and many more can further be added in the tool. The type conversion (if required) can be implemented through an intelligent agent.
ii. The horizontal partitions can be implemented using an intelligent agent.
iii. More data mining algorithms can be added in the algorithm's bank of the tool.
iv. The business goals and the user requirements can separately be written and the extracted 'knowledge' can be compared with these scenarios.
v. An intelligent agent can be used to interpret and evaluate the 2D graph.

Furthermore, the proposed model is compared with the most commonly used data mining suites namely Oracle Data Mining (ODM) and MS SQL Server and the results reveal that the proposed model performs better [23-26].

4. Methodology. A MAS approach is used to implement the architecture of the UDM Server where intelligent agents are deployed in an autonomous manner to select the right model fitted for any domain or problem and subsequently conduct each task of the data mining processes within a Multiagent System (MAS) architecture. We identify a metric that will determine the choice of the best algorithm at each step of the data mining process. The MAS approach has proven to be useful in designing of a system where the domains require deploying the MAS, even in those systems which are not distributed. The parallelism, scalability, robustness and simple programming are the most common characteristics of a Multiagent System which are helpful to speed up the performance and operation of the system.

We also formulate the interpretation and evaluation criteria which help to select 'knowledge' from the domain of knowledge. The Evaluation criteria are:

i. Compute the population of the each cluster.
ii. Calculate the percentage of each parameter in a cluster.
iii. Determine the Minimal Description Length (MDL) value of each cluster. The following steps explain to calculate the value of MDL.

Step 1: The formula to calculate the maximum likelihood is shown in eq. (4).

$$likelihood(\theta \mid x1, x2, \ldots, xn) = p(x1, x2, \ldots, xn \mid \theta) = \prod_{i=1}^{n} p(xi \mid \theta) \quad (4)$$

Take the logarithm of this value will give the value of model accuracy which is shown in eq. (5).

$$ModelAccuracy = \log(likelihood) \quad (5)$$

Step 2: The formula to calculate the model size is $ModelSize = k(\log n)$ where k, n are the parameter and datapoints respectively.

Step 3: The MDL is computed by using eq. (6).

$$MDLScore = ModelSize - ModelAccuracy \quad (6)$$

Minimal Description Length (MDL) is also referred as BIC (Bayesian Information Criterion). For the interpretation and explanation; the decision rules are utilized which are also helpful for the future prediction. The 2D graph(s) of minimum MDL value is 'knowledge' [9-15].

5. Results and Discussion. The dataset DNA was chosen; a medical dataset contains information about DNA. There are 3 different classes (parameters) of DNA with 181 attributes and 14000 sample size [8]. Number of partitions of dataset DNA is 90 according to eq. (1) and if we create 3 clusters per partition then there will be 270 graphs. It is difficult to manage and store 270 graphs; we apply the selection criteria discussed in section 4 and select only the attributes of minimum MDL. There are different approaches to optimize number of clusters; we are optimizing number of clusters through the number of parameters in a dataset. Table 3 shows the results obtained from UDM Server for dataset DNA after applying the proposed selection criteria discussed in section 4.

Table 3. The Results of Dataset DNA

Partition#	Partition's Attributes	Cluster#	Cluster's Population	Parameters' %	Cluster's MDL Value
1	A1, A2, Class	1	Class 1=6000 Class 2=5000 Class 3=3000	Class 1=42.8 Class 2=35.7 Class 3=21.4	17.7
		2	Class 1=2000 Class 2=7000 Class 3=5000	Class 1=14.3 Class 2=50.0 Class 3=35.7	19.0
		3	Class 1=8000 Class 2=3000 Class 3=3000	Class 1=57.1 Class 2=21.4 Class 3=21.4	21.3
2	A25, A26, Class	1	Class 1=9000 Class 2=2000 Class 3=3000	Class 1=64.3 Class 2=14.3 Class 3=21.4	19.9
		2	Class 1=10000 Class 2=500 Class 3=3500	Class 1=71.4 Class 2=3.6 Class 3=25.0	17.6
		3	Class 1=9500 Class 2=1500 Class 3=3000	Class 1=67.8 Class 2=10.7 Class 3=21.4	23.5
3	A89, A90, Class	1	Class 1=8500 Class 2=3500 Class 3=2000	Class 1=60.7 Class 2=25.0 Class 3=14.3	13.9
		2	Class 1=7500 Class 2=4500 Class 3=2000	Class 1=53.6 Class 2=32.1 Class 3=14.3	15.7
		3	Class 1=6500 Class 2=5500 Class 3=2000	Class 1=46.4 Class 2=39.3 Class 3=14.3	18.2
4	A131, A132, Class	1	Class 1=5500 Class 2=4500 Class 3=4000	Class 1=39.3 Class 2=32.1 Class 3=28.6	15.3
		2	Class 1=9000 Class 2=2000 Class 3=3000	Class 1=64.3 Class 2=14.3 Class 3=21.4	14.9
		3	Class 1=2500 Class 2=7500 Class 3=4000	Class 1=17.8 Class 2=53.6 Class 3=28.6	9.2
5	A171, A172, Class	1	Class 1=4000 Class 2=7000 Class 3=3000	Class 1=28.6 Class 2=50.0 Class 3=21.4	14.1
		2	Class 1=8500 Class 2=1500 Class 3=4000	Class 1=60.7 Class 2=10.7 Class 3=28.6	16.3
		3	Class 1=4500 Class 2=3500 Class 3=6000	Class 1=32.1 Class 2=25.0 Class 3=42.8	14.5

Explanation of Table 3: The population of each partition of the dataset is same i.e. the first cluster of each partition is less populated then the second cluster. The percentage value of parameter 'Class 1' is high in the highly populated cluster of each partition. The percentage value of parameter 'Class 2' is high in the less populated cluster of each partition and finally, the percentage value of parameter 'Class 3' is less in the last cluster of each partition. The value of MDL of each cluster of the dataset varies. The cluster 1 of partition 1, cluster 2 of partition 2, cluster 1 of partition 3, cluster 4 of partition 4 and cluster 1 of partition 5 is less than the other clusters of these partitions. Therefore, the attributes A1, A2, A25, A26, A89, A90, A131, A132, A171 and A172 play important and vital role in the extraction of knowledge from the dataset DNA.

6. Conclusion. In this paper, a unified data mining theory is presented (UDMT) where the data mining processes; clustering, classification and visualization are unified and the relevancy of UDMT for the big data is discussed. The limitations of the theory are highlighted. The model selection criterion, AIC is applied to overcome these limitations which also helps to determine the fitness of the dataset and to select the appropriate data mining algorithm. The usages, challenges and techniques for the big data are also presented. Based on the UDMT, an automated, a Multiagent System (MAS) called a UDM Server is developed and tested on a variety of real life big datasets. Although the data mining tools and suites available in the market can handle the large and big datasets but the knowledge extraction process is not straightforward even for the experienced users. Of course, one cannot deny the significance of the statistical information of a dataset but the combination of the above said data mining steps is noteworthy and both are the determinant factor for knowledge. The proposed UDM Server possesses both characteristics; therefore it is an ultimate selection for the big data. We conclude that the UDM Server helps to extract knowledge from the big datasets and the process of knowledge extraction is straightforward and the proposed theory (UDMT) is relevant for big data.

Future Work

i. The architecture of the UDM Server can help to create the Decision Support Systems.
ii. The proposed Framework can further be enhanced to develop a UML like framework for data mining.

Acknowledgement

The authors are thankful to The Islamia University of Bahawalpur, Pakistan for providing financial assistance to carry out this research activity under Higher Education Commission (HEC), Pakistan project 6467/F – II.

REFERENCES

1 URL: http://www.sas.com/big-data/ visited on June 2013.

2 IDC. "Big Data Analytics: Future Architectures, Skills and Roadmaps for the CIO," September 2011.

3 American Bankers Association, March 2009.

4 URL: http://www.economist.com visited on June 2013.

5 URL: http://blog.twitter.com visited on June 2013.

6 URL: http://newsroom.fb.com/ visited on June 2013.

7 Dan Briody, "Big Data: Harnessing a Game-Changing Asset.", Economist Intelligence Unit's publication, 2011.

8 URL: www.sgi.com/tech/mlc/db/, US Census Bureau for the datasets Iris, Breastcancer and Sales, visited 2012.

9 Khan, Dost Muhammad., and Mohamudally, Nawaz., "A Multiagent System (MAS) for the Generation of Initial Centroids for K-means Clustering Data Mining Algorithm based on Actual Sample Datapoints", Journal of Next Generation Information Technology, Volume 1, Number 2, August, 2010 pp.: 85-95.

10 Khan, Dost Muhammad., Mohamudally, Nawaz. and Babajee, DKR., "Investigating the Statistical Linear Relation between the Model Selection Criterion and the Complexities of Data Mining Algorithms", Journal of Computing, vol. 4, Issue 8, pp: 14-28, 2012.

11 Pham, D.T., Dimov, S. S., and Nguyen, C.D., "Selection of K in K-means clustering", Proc. IMechE Vol. 219, Part C: J. Mechanical Engineering Science, 2005.

12 Mark Ming-Tso Chiang, and Boris Mirkin, "Intelligent Choice of the Number of Clusters in K-Means Clustering: An Experimental Study with Different Cluster Spreads", Journal of Classification 27, 2009.

13 Khan, Dost Muhammad., Mohamudally, Nawaz. and Babajee, DKR., "Towards the Formulation of a Unified Data Mining Theory, Implemented by Means of Multiagent Systems (MASs)", Advances in Data Mining Knowledge Discovery and Applications, pp. 03-42, InTech, ISBN: 978-953-51-0748-4, 2012.

14 Khan, Dost Muhammad., Mohamudally, Nawaz. and Babajee, DKR., "The Formulation of a Data Mining Theory for the Knowledge Extraction by means of a Multiagent System", Journal of Computing, vol. 4, Issue 8, pp: 29-38, ISSN 2151-9617, 2012.

15 Khan, Dost Muhammad., and Mohamudally, Nawaz., "An Agent Oriented Approach for Implementation of the Range Method of Initial Centroids in K-Means Clustering Data Mining Algorithm", IJIPM, pp. 104-113 vol. 1 No.1, 2010.