Chapter 1

Towards the Formulation of a Unified Data Mining Theory, Implemented by Means of Multiagent Systems (MASs)

Dost Muhammad Khan, Nawaz Mohamudally and D. K. R. Babajee

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/47984

1. Introduction

Data mining techniques and algorithms encompass a variety of datasets like medical, geographical, web logs, agricultural data and many more. For each category of data or information, one has to apply the best suited algorithm to obtain the optimal results with highest accuracy. This is still a problem for many data mining tools as no unified theory has been adopted. The scientific community is very much conscious about this problematical issue and faced multiple challenges in establishing consensus over a unified data mining theory. The researchers have attempted to model the best fit algorithm for specific domain areas, for instance, formal analysis into the fascinating question of how overfitting can happen and estimating how well an algorithm will perform on future data that is solely based on its training set error (Moore Andrew W., 2001, Grossman. Robert, Kasif. Simon, et al, 1998, Yang. Qlang, et al, 2006 & Wu. Xindong, et al 2008).

Another problem in trying to lay down some kind of formalism behind a unified theory is that the current data mining algorithms and techniques are designed to solve individual consecutive tasks, such as classification or clustering. Most of the existing data mining tools are efficient only to specific problems, thus the tool is limited to a particular set of data for a specific application. These tools depend again on the correct choice of algorithms to apply and how to analyze the output, because most of them are generic and there is no context specific logic that is attached to the application. A theoretical framework that unifies different data mining tasks including clustering, classification, interpretation and association rules will allow developer and researchers in their quest for the most efficient and effective tool commonly called a unified data mining engine (UDME), (Singh. Shivanshu K., Eranti. Vijay Kumer. & Fayad. M.E., 2010, Das. Somenath 2007 & Khan. Dost Muhammad, Mohamudally. Nawaz, 2010).

A Multi Agent System (MAS) approach has proven to be useful in designing of a system where the domains require the MAS, even in those systems which are not distributed. The multiagent system speeds up the performance and operation of the system by providing a method for parallel computation i.e. a domain that is easily broken into components, several independent tasks that can be handled by separate agents, could benefit from the MAS. Furthermore, a MAS approach provides the scalability, since they are inherently modular; it is easy to add new agents in a multiagent system and robustness is another benefit of multiagent system (Peter Stone & Manuela Veloso, 1997 & Khan. Dost Muhammad, Mohamudally. Nawaz, 2010).

A first tentative to integrate different data mining algorithms using a MAS was implemented in the application of a Unified Medical Data Miner (UMDM) for prediction, classification, interpretation and visualization on medical datasets: the diabetes dataset case and the integration of K-means clustering and decision tree data mining algorithms. The study is conducted on the development and validation of a MAS coupling the K-means and C4.5 algorithms. This approach had been successful with different datasets namely Iris, a flower dataset, BreastCancer and Diabetes medical datasets (US Census Bureau., 2009). The results produced were highly satisfactory, encouraging and acceptable. The interpretation and visualization of individual clusters of a dataset and the whole partitioned clustered dataset had also been dealt with (Mohamudally. Nawaz, Khan. Dost Muhammad 2011, Khan, Dost Muhammad. & Mohamudally, Nawaz. 2011).

However, the choice of the algorithms was empirical and intuitive. Some researchers have evaluated the VC (Vapnik-Chervonenkis)-dimension of different algorithms in order to map the appropriate algorithm with a particular dataset. There have been also similar research conducted using the CV (Cross-validation), AIC (Akaike Information Criterion), BIC (Bayesian Information Criterion), (SRMVC) Structural Risk Minimize with VC dimension models. The main aim of this book chapter is to investigate into the roadmap towards using the intelligent agents in an autonomous manner to select the right model fitted for any domain or problem and subsequently conduct each task of the data mining process within the MAS architecture.

The rest of the book chapter is organized as follows; section 2 is about the problematical issues in data mining and section 3 is about Unified Data Mining Theory (UDMT). In section 4 the Mathematical Formulation of Unified Data Mining Theory is discussed, section 5 deals with the Unified Data Mining Tool (UDMTool) and finally the conclusion is drawn in section 6.

2. Problematical issues in data mining

Data mining has achieved tremendous success and many problems have been solved by using data mining techniques. But still there are some challenges in the field of data mining research which should be addressed. These problems are: Unified Data mining Processes, Scalability, Mining Unbalanced, Complex and Multiagent Data, Data mining in Distributed and Network setting and Issues of Security, Privacy and Data Integrity in data mining

(Yang. Qlang, et al, 2006). The focus of this book chapter is on unified data mining processes such as clustering, classification, visualization followed by interpretation and proposes a unified data mining theory.

2.1. Unified data mining processes

There are many data mining algorithms and techniques which are designed for individual problems, such as classification or clustering. A theoretical framework is required that unifies different data mining tasks including clustering, classification, interpretation and association rules which would help the field of data mining and provide a basis for future research (Yang. Qlang, et al, 2006 & Wu. Xindong, et al 2006).

The following two figures 1 and 2 explain the process of uniformality.

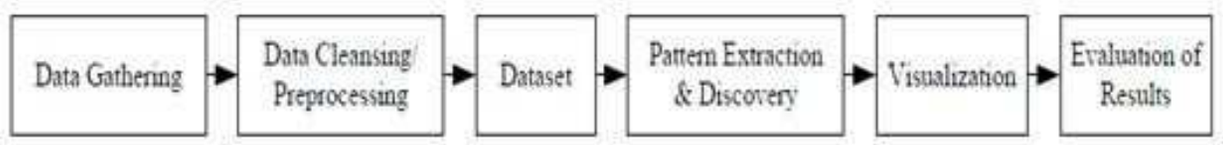

Figure 1. Conventional Life Cycle of Data Mining

Data mining is an iterative process and different stages or steps or processes are required to complete the data mining process. The figure 1 is the conventional life cycle of data mining. The first step is data gathering, then data cleansing and then preparing a dataset, the next stage is pattern extraction & discovery [The pattern extraction and discovery from large dataset is a two steps process. In the first step, the clusters of the dataset are created and the second step is to construct the 'decision rules' (if-then statements) with valid pattern pairs]. The choice of the algorithm depends on the intended use of extracted knowledge. The other stages are visualization and evaluation of results. The user has to select a data mining algorithm on each step in this life cycle, i.e. one algorithm for clustering, one for classification, one for interpretation and one for visualization. Every process is individually carried out in this life cycle.

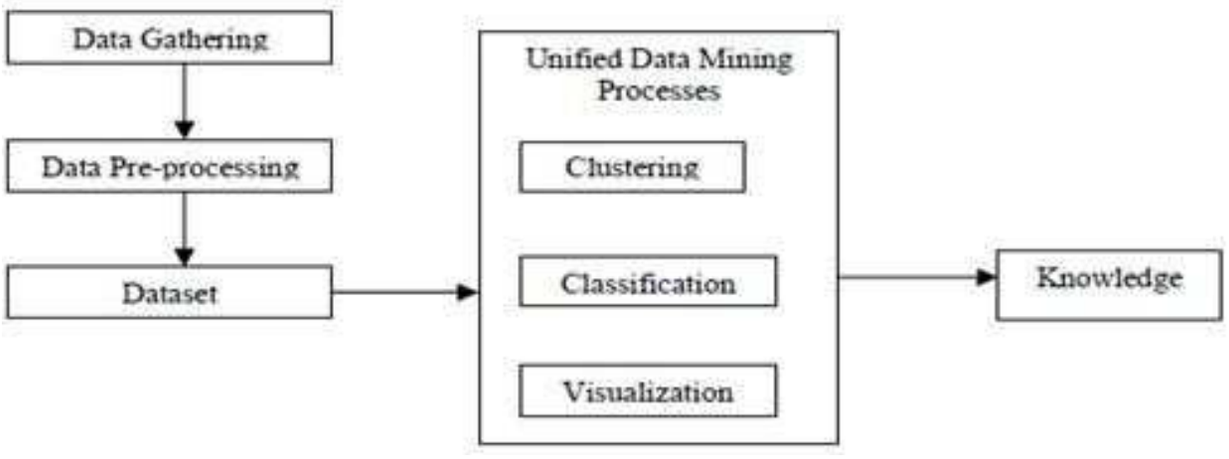

Figure 2. The Unified Data Mining Life Cycle

The figure 2 is the proposed unified data mining life cycle. The first three processes of unified data mining processes are the same as the cycle of data mining processes, i.e. data gathering, data cleansing followed by the preparing a dataset. The next process unifies the clustering, classification and visualization processes of data mining, called unified data mining processes (UDMP) followed by the output which is 'knowledge'. The user evaluates and interprets the 'knowledge' according to his business rules. The dataset is the only required input; the 'knowledge' is produced as final output from the proposed cycle of unified data mining processes. There is no need to select any data mining algorithm at any stage in the unified cycle of data mining processes. Multiagent systems (MASs) approach is used to unify clustering, classification and visualization processes of data mining.

3. Unified Data Mining Theory (UDMT)

The theories either explain little or much. Generally, science prefers those that explain much. The best theory would surely be the one that explains everything in all scientific disciplines. However, such a theory has proven hard to find. Instead the separate disciplines have been working towards unifying the sub-theories present in their respective subjects. In physics, for instance, there is an agreed upon taxonomy of the explanatory capacity of theories that illustrate strive of the unification. Physicists believe that there are four forces, or interactions, that affect matter in the universe: the strong force, the electromagnetic force, the weak force, and the gravitational force. A theory which has the set of requirements, comprehensiveness, preciseness, consistency and correctness is called a unified theory (Johnson, P., Ekstedt, M., 2007).

In 1687, Sir Isaac Newton proposed a unified theory of mechanics in order to explain and predict all earthly and heavenly motion (Newton, I., 1687). John Dalton 1803 revived a unified theory of matter, the atomic theory, explaining the nature of all physical substance (Dalton, J., 1808). In 1839, Matthias Schleiden and Theodor Schwann developed the theory of the cell, explaining the fundamental structure of living organisms (Schwann, T., 1839). In 1859, Charles Darwin proposed a unified theory of evolution by natural selection, in order to explain all variability of the living (Darwin, C., 1859). In 1869, Dmitri Ivanovich Mendeleev presented the periodic system, a unified theory explaining properties of all the chemical elements (Mendeleev, D., 1869). In mid 1800, James Clerk Maxwell, in his unified theory of electromagnetism, explained the interrelation of electric and magnetic fields. In 1884, Hertz, demonstrated that radio waves and light were both electromagnetic waves, as predicted by Maxwell's theory. In early 20th century, Albert Einstein, in general theory of relativity which deals with gravitation, became the second unified theory (Brog, Xavier., 2011) .

The term unified field theory was first coined by Einstein, while attempting to prove that electromagnetism and gravity were different manifestations of a single fundamental field and failed in this ultimate goal. When quantum theory entered the picture, the puzzle became more complex. Einstein spent much of his later life trying to develop a Unified Theory (UT) that would combine gravitation and electromagnetism. The theory resulting from the combination of the strong, weak and electromagnetic forces has been called the

Grand Unified Theory (GUT). At the end of the rainbow lies the dream of unifying all four forces. The resulting theory is often referred to as the Theory of Everything (TOE). As recently as 1990, Allen Newell proposed a unified theory of cognition in order to explain and predict all human problem-solving (Newell, A., 1990). The history of science is written in terms of its unified theories, because arguably, unified theories are the most powerful conceptual vehicles of scientific thought. Each force has a theory that explains how it works (Johnson, P., Ekstedt, M., 2007).

There is no accepted unified field theory, and thus remains an open line of research. The term was coined by Einstein, who attempted to unify the general theory of relativity with electromagnetism, hoping to recover an approximation for quantum theory. A "theory of everything" is closely related to unified field theory, but differs by not requiring the basis of nature to be fields, and also attempts to explain all physical constants of nature (Brog, Xavier., 2011, Johnson, P., Ekstedt, M., 2007, Mielikäinen, Taneli., 2004, Sarabjot S. Anand, David A. Bell, John G. Hughes, 1996).

3.1. Prerequisites for a unified theory

A good unified theory is beneficial for the discipline of data mining. In order to know whether the theory is good or not, the requirements for a unified theory need to be outlined. The important attributes of a unified theory are divided into four main categories (King, G., R. Keohane, and S. Verba, 1994, Popper K., 1980).

i. Comprehensiveness – that the theory covers all relevant phenomena
ii. Preciseness – that the theory generates applicable explanations and predictions
iii. Consistency – that the theory does not contradict itself
iv. Correctness – that the theory represents the real world (Johnson, P., Ekstedt, M., 2007).

3.1.1. Comprehensiveness

The comprehensiveness of a theory measures the scope of the theory; how much the theory can explain. All theories have delimitations, bounds outside of which they to not aspire to be applicable. Comprehensiveness is an important, not to say defining, characteristic of a unified theory. A unified theory of data mining should thus be able to express, explain and predict those issues which the users are facing nowadays (Johnson, P., Ekstedt, M., 2007).

3.1.2. Preciseness

The preciseness of a theory measures to what extent the theory is able to produce specific explanations and predictions. Theories with this capability are more useful than those that lack it, partly because precise results are generally more useful, but also because it is more or less impossible to tell whether imprecise theories are true of false. The philosopher of science Karl Popper called this property of a theory for falsifiability (Kuhn, T., 1962). A theory that is not falsifiable lacks all explanatory and predictive capacity. Ambiguous terminology ensures imprecision (Johnson, P., Ekstedt, M., 2007).

3.1.3. Consistency

A consistent theory does not contradict itself. This is a fairly obvious requirement on a theory; an inconsistent proposition is of no explanatory or predictive value. Consistency is, however, difficult to achieve. It is oftentimes difficult to once and for all verify that a theory is free of inconsistencies. However, three properties of theories greatly facilitate the identification of any self-contradictions. Firstly, small, simple, parsimonious theories are easier to grasp than large ones. It is easier to relate a small set of propositions to each other than a large set. Secondly, formalized, or structured, theories are easier to manage than poorly structured ones. Mathematically formulated theories are therefore easier to check for consistency than theories presented in natural language prose. Finally, imprecise theories may be interpreted differently. Some of these interpretations may be consistent while others are not. Precision thus also facilitates consistency checking (Johnson, P., Ekstedt, M., 2007).

3.1.4. Correctness

The arguably most important of all qualities of a theory is that it is correct, i.e. that it generates true predictions and explanations of the part of the world that it represents. It is the goal of all science to propose theories and then to test them. This is oftentimes a lengthy process, where many different aspects of a theory are tested against observations of the real world. A problem with the property of correctness is that, according to Popper, the correctness of a theory can never be demonstrated beyond doubt (unless it is a tautology), but the theory is corroborated by passing various empirical tests. Popper argues that once evidence has been discovered that falsifies a theory, it should be abandoned. Compared to established sciences such as physics, medicine and macro economics, little research effort is currently directed towards theory corroboration in data mining (Johnson, P., Ekstedt, M., 2007). According to Kuhn, this is not surprising given the scientific maturity of the discipline (Kuhn, T., 1962).

3.2. The advantages of a unified theory

A unified theory provides not only greater explanatory and predictive power, but also unifies the people involved in the discipline. Research and practice with a unified theory can become more coordinated and focused than with a set of disparate micro-theories. A unified theory provides a common vocabulary for expressing and communicating about the world, it defines common problems within that world, and it delimits the acceptable methods of reaching scientific conclusions regarding those problems. The important advantages and benefits of unified theory are, it provides better explanatory and predictive capacity, and it provides a common world view for the researchers and practitioners in the field (Johnson, P., Ekstedt, M., 2007). The following are the advantages of a unified theory:

3.2.1. Explanatory and predictive capacity

This is a question of organization of work, which also reappears as an underlying explanatory theory. Summarizing, a unified theory is liable to have greater explanatory and predictive power than a set of micro-theories (Johnson, P., Ekstedt, M., 2007).

3.2.2. Common terminologies

A unified theory of data mining implies a common terminology, by providing the facility to the users to understand each other. There are astounding numbers of definitions of many important concepts of data mining. Some of the concepts are unclear; there is no generally agreed upon definition of the term data mining (Johnson, P., Ekstedt, M., 2007).

3.2.3. Common problems

Without a unified theory of data mining the researchers will have difficulties seeing the big picture, i.e. the two disciplines at a time may have very much in common and the results of one discipline may be very fruitful for the other. Theories not only answer research questions, but they also pose new ones. They define what good research questions are. With a unified theory, researchers and practitioners in the field will see the same big picture and if the picture is the same, then the relevant problems with it are more likely to be the same (Johnson, P., Ekstedt, M., 2007).

3.2.4. Common conclusions

Without a unified theory, there is no common base from which an argument may be made for or against some proposition about the state of the world. It is equally acceptable to present an argument based on sociological research on group dynamics as one based on findings within solid state physics as one based on results in fractal geometry. This is problematic, because these scientific disciplines are not consistent with each other and may very well lead to opposing conclusions, and there is really no systematic way of drawing an aggregated conclusion (Johnson, P., Ekstedt, M., 2007).

In a good unified theory, arguments will have the same foundation. It may seem strange that truths from one discipline cannot be invoked at will in another discipline, but this is the case in all mature sciences; it is a kind of separation of concerns. In Kuhn's unified normal scientific disciplines, such as electromagnetism, there are no inter-theoretical discussions (Kuhn, T., 1962). This is of course very satisfying, because it provides a strong base for reaching the same conclusions and few discussions need to end up in stalemates (at least in principle). In pre-scientific disciplines, the debaters are at least aware of when they move from the intra-theoretical arena to the inter-theoretical one, thus maintaining a separation of concerns (Johnson, P., Ekstedt, M., 2007).

In data mining and other fragmented disciplines, the lack of unified theories makes such a separation impossible and there are no real rules for distinguishing between sound arguments and unsound ones. It is consequently quite acceptable to pick any one of the thousands of micro-theories available to support an argument and simply disregard the thousands remaining. We can thus explain and predict everything and nothing (Johnson, P., Ekstedt, M., 2007).

The advantage of a unified theory over many fragmented theories is that a unified theory often offers a more elegant explanation of data, and may point towards future areas of study

as well as predict the laws of nature (Brog, Xavier., 2011, Unified Theory or Theory of Everything (TOE), 2001).

4. Mathematical formulation of a Unified Data Mining Theory (UDMT)

In this section we formulate a unified data mining theory using the mathematical functions. We start with the definition of an algorithm. An algorithm is a description of a mechanical set of steps for performing a task. The algorithms are considered a powerful tool in the field of computer science. When thinking about inputs and outputs, we can treat algorithms as functions; input a number into the algorithm, follow the prescribed steps, and get an output. A function is a relationship between an input variable and an output variable in which there is exactly one output for each input. To be a function, there are two requirements:

1. An algorithm to be consistent i.e. every time give the input, get the output.
2. Each input produces one possible output.

It is not necessary that all functions have to work on numbers and not all functions need to follow a computational algorithm (Dries. Lou van den. 2007, Lovász. L. , Pelikán. J., Vesztergombi. K. 2003 and MathWorksheetsGo, 2011). The data mining algorithms used in unified data mining processes satisfy the conditions of a function; therefore these algorithms are applied as functions.

Let S = {Set of n attributes}
A = {Set of n clusters of n attributes}
B = {Set of n rules of n clusters of n attributes}
C = {Set of n 2D graphs of n rules of n clusters of n attributes}

We can describe this in a more specific way that:

Let $S = \{s_1, s_2, ..., s_n\}$ where $s_1, s_2, ..., s_n$ are the partitions of datasets containing at least any of the two attributes and only one 'class' attribute of the original dataset.

$A = \{C_1(c_1, c_2, ..., c_n), C_2(c_1, c_2, ..., c_n), ..., C_n(c_1, c_2, ..., c_n)\}$ where $C_1(c_1, c_2, ..., c_n)$ is set of clusters of partitions 's_1', $C_2(c_1, c_2, ..., c_n)$ is set of clusters of partitions 's_2',…., and $C_n(c_1, c_2, ..., c_n)$ is set of clusters of partition 's_n'.

$B = \{R_1(r_1, r_2, ..., r_n), R_2(r_1, r_2, ..., r_n), ..., R_n(r_1, r_2, ..., r_n)\}$ where $R_1(r_1, r_2, ..., r_n)$ is set of rule of cluster 'C_1', $R_2(r_1, r_2, ..., r_n)$ is the set of rule of cluster 'C_2', …, and $R_n(r_1, r_2, ..., r_n)$ is set of rule of cluster 'C_n'.

$C = \{V_1(v_1, v_2, ..., v_n), V_2(v_1, v_2, ..., v_n), ..., V_n(v_1, v_2, ..., v_n)\}$ where $V_1(v_1, v_2, ..., v_n)$ is set of 2D graphs of 'R_1', $V_2(v_1, v_2, ..., v_n)$ is the set of 2D graphs of 'R_2', …, and $V_n(v_1, v_2, ..., v_n)$ is the set of 2D graphs of 'R_n'.

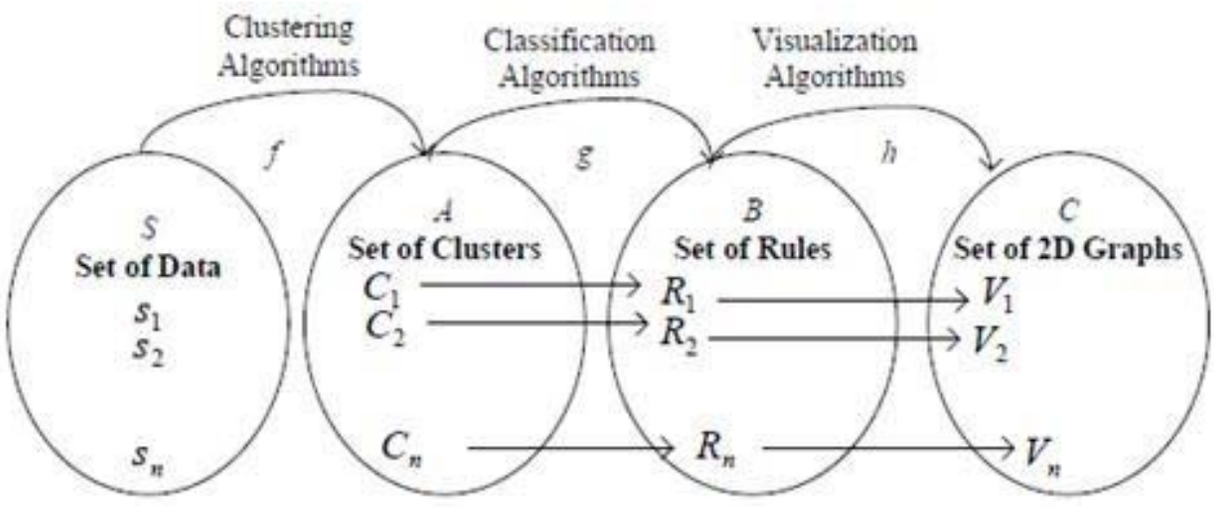

Figure 3. The Composition of the Functions

1. $f : S \to A$

$$f(s_j) = C_k, \text{ where } j, k = 1, 2, \ldots, n.$$

2. $g : A \to B$

$$g(C_l) = R_m, \text{ where } l, m = 1, 2, \ldots, n.$$

3. $h : B \to C$

$$h(R_o) = V_p, \text{ where } o, p = 1, 2, \ldots, n.$$

Lemma 1: *f, g & h are* functions

Proof:

1. *f* is a function

Suppose that *f* is not a function then s_1 can go to cluster 1 and cluster 2. This is impossible because clustering requires that s_1 goes to set of cluster 1. This leads to a contradiction. Therefore, *f* is a function.

2. *g* is a function

We define rule, depending on the clusters. Cluster 1 has Rule 1, Cluster 2 has Rule 2. We cannot define two rules to a cluster otherwise we would not be able to classify the attributes.

3. *h* is a function

The 2D graph will depend on the rule. A rule cannot produce two different 2D graphs.

The domain and co-domain of the functions are given by

$$dom(f) = S \qquad codom(f) = A$$

$$dom(g) = A \qquad codom(g) = B$$

$$dom(h) = B \qquad codom(h) = C$$

The function *f* is a mapping from the set of dataset to the set of clusters, the function *g* is a mapping from set of clusters to the set of rules and the function *h* is a mapping from set of rules to set of 2D graphs. The function *f* takes the set of data like $s_1, s_2, ..., s_n$ as input, apply clustering algorithm (K-means) and produces the clusters like $C_1(c_1, c_2, ..., c_n), C_2(c_1, c_2, ..., c_n), ..., C_n(c_1, c_2, ..., c_n)$ as the output. The function *g* takes the clusters like $C_1(c_1, c_2, ..., c_n), C_2(c_1, c_2, ..., c_n), ..., C_n(c_1, c_2, ..., c_n)$ as input, apply the classification algorithm (C4.5 Decision Tree) and produces the rules like $R_1(r_1, r_2, ..., r_n), R_2(r_1, r_2, ..., r_n), ..., R_n(r_1, r_2, ..., r_n)$ as the output. The function *h* takes the rules like $R_1(r_1, r_2, ..., r_n), R_2(r_1, r_2, ..., r_n), ..., R_n(r_1, r_2, ..., r_n)$ as input, apply data visualization algorithm and produces 2D graphs like $V_1(v_1, v_2, ..., v_n), V_2(v_1, v_2, ..., v_n), ..., V_n(v_1, v_2, ..., v_n)$ as the output. From the 2D graphs we can interpret and evaluate the results to get the knowledge, which is accepted or rejected by the user.

The domain and co-domain of the functions are given by

$$dom(g \circ f) = S \qquad codom(g \circ f) = B$$

$$dom(h \circ (g \circ f)) = S \qquad codom(h \circ (g \circ f)) = C$$

$g \circ f$ is a mapping from the set of data to the set of rules and $h \circ (g \circ f)$ is a mapping from the set of data to the set of 2D graphs. The knowledge is derived from the interpretation and evaluation of the 2D graphs which are obtained through the composition of the functions discussed above.

For $V_p \in C$

$$\because V_p = h(R_m)$$
$$= h(g(C_k))$$
$$= h(g(f(s_j)))$$
$$= (h \circ g \circ f)(s_j) \text{ [Composition of the functions]}$$

$$\therefore (h \circ g \circ f) : S \to C \tag{1}$$

We first pass the partitions of the dataset through the composition of clustering, classification and visualization and then the results obtained are interpreted and evaluated to extract the knowledge. Therefore, we attempt to unify the processes of data mining life cycle.

The order of the functions must be the same as given in equation (1). That is, first apply the clustering algorithms, then classification algorithms, then visualization algorithms and finally the interpretation of the results of visualization. The selection of the appropriate and right result(s) will give the knowledge. This is the proof of the mathematical formulation of the unified process of data mining life cycle.

Illustration:

Let S be a dataset with two vertical partitions s_1 and s_2 , the set A is a set of clusters of each partition, the set B is the set of rules of each clusters and the set C is the set of 2D graphs of each rules. The mathematical notation of these sets is given below.

$$S = \{s_1, s_2\}$$
$$A = \{C_1(c_1, c_2), C_2(c_1, c_2)\}$$
$$B = \{R_1(r_1, r_2), R_2(r_1, r_2)\}$$
$$C = \{V_1(v_1, v_2), V_2(v_1, v_2)\}$$

Dataset S

CT	SECS	Mitoses	Class
3	2	1	Benign
5	2	1	Benign
1	2	1	Malignant
3	5	3	Malignant

s_1

CT	Mitoses	Class
3	1	Benign
5	1	Benign
1	1	Malignant
3	3	Malignant

s_2

CT	SECS	Class
3	2	Benign
5	2	Benign
1	2	Malignant
3	5	Malignant

1. $f(K-means): S \rightarrow A$

K-means clustering algorithm takes three inputs, *'k'* number of clusters,, *'n'*, number of iteration and *'s_p'* dataset and will produce k clusters of the given dataset. The function *f* is illustrated in equation (2).

$$f(k,n,s_p) = C_k(c_k) \tag{2}$$

where k and n are positive nonzero integers.

Suppose we want to create two clusters of the datasets s_1 and s_2 and number of iteration is set to 10 i.e. $k = 2$ and $n = 10$ then the function $f(2,10,s_1) = C_1(c_1,c_2)$ is for the dataset s_1.

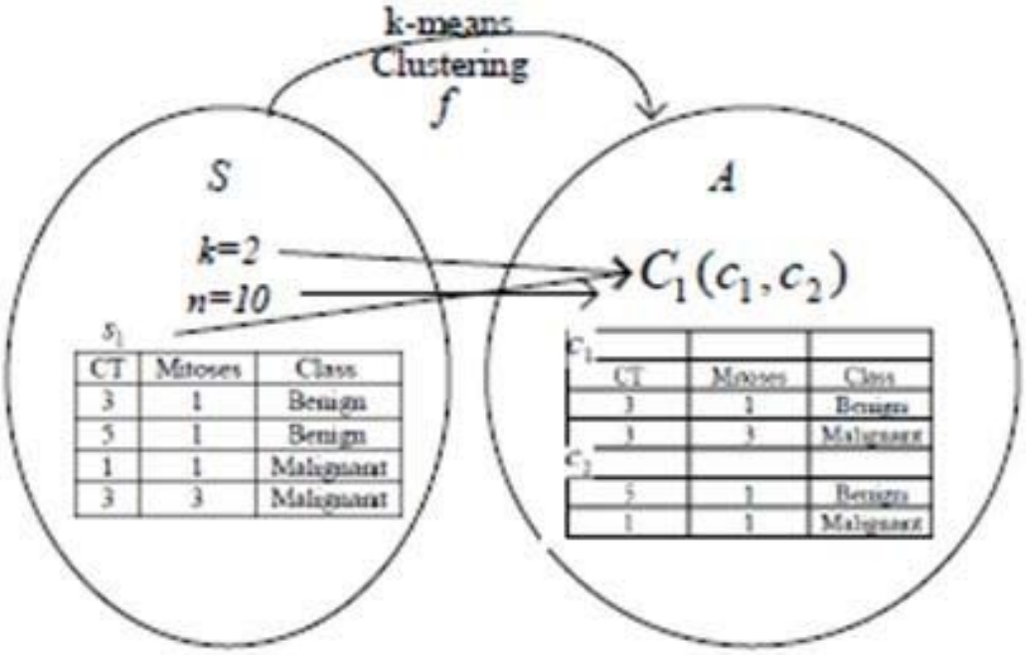

Figure 4. The mapping of function f for s_1

Figure 4 shows the mapping between the set S and the set A using K-means clustering algorithm as a function f. The values of other required parameters of this algorithm are number of clusters k = 2, number of iterations n = 10 and dataset s_1. This is m:1 (3:1) mapping because all the members of set S are mapped with the single member of set A.

Similarly, the function $f(2,10,s_2) = C_2(c_1,c_2)$ is for the dataset s_2.

Figure 5 shows the mapping between the set S and the set A using K-means clustering algorithm as a function f. The values of other required parameters of this algorithm are number of clusters k = 2, number of iterations n = 10 and dataset s_2. This is m:1 (3:1) mapping because all the members of the set S are mapped with the single member of the set A.

The figures 4 and 5 show that function f is m:1 function. Hence f is a function because it satisfies the conditions of the function; therefore the K-means clustering algorithm is a function. If we optimize the values of *'k'* and *'n'* then the K-means clustering algorithm will be 1:1 otherwise it is m:1. The process of creating clusters of the given dataset does not split the dataset into small datasets, the dataset remains the same and only datapoints are shifted within the dataset. For our own convenience we are illustrating different clusters of the dataset. This process does not create new datasets from the given dataset.

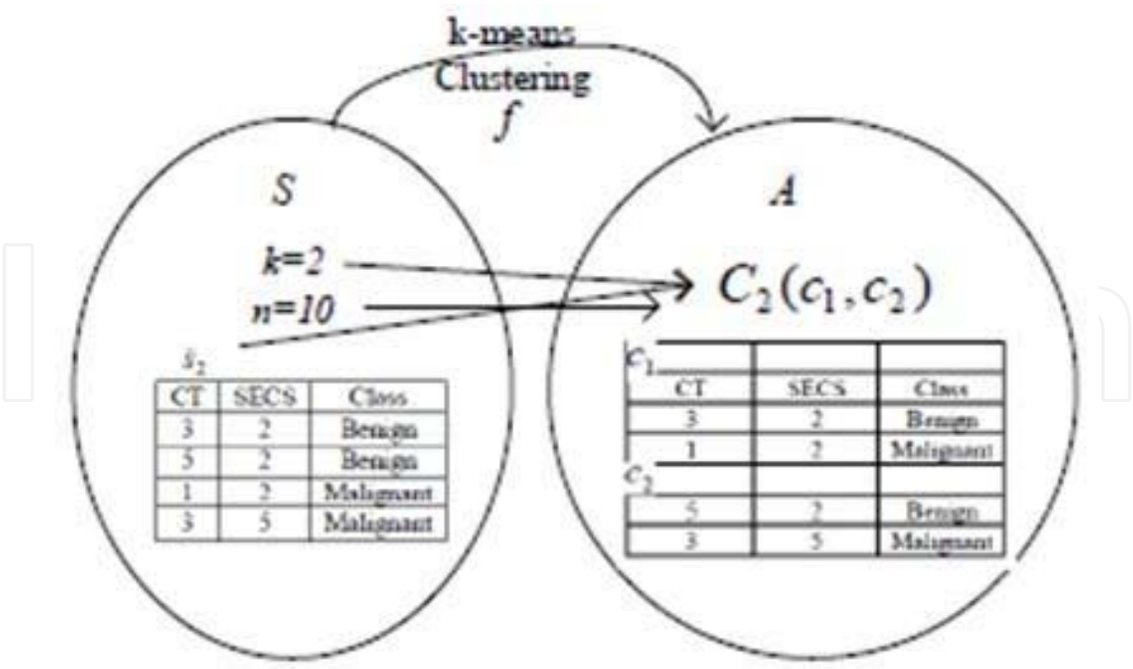

Figure 5. The mapping of function f s_2

2. $g(C4.5) : A \rightarrow B$

The C4.5 algorithm takes the clusters of the dataset as input and produces the rule as output. The function g is illustrated in equation (3).

$$g(C_k) = R_k \tag{3}$$

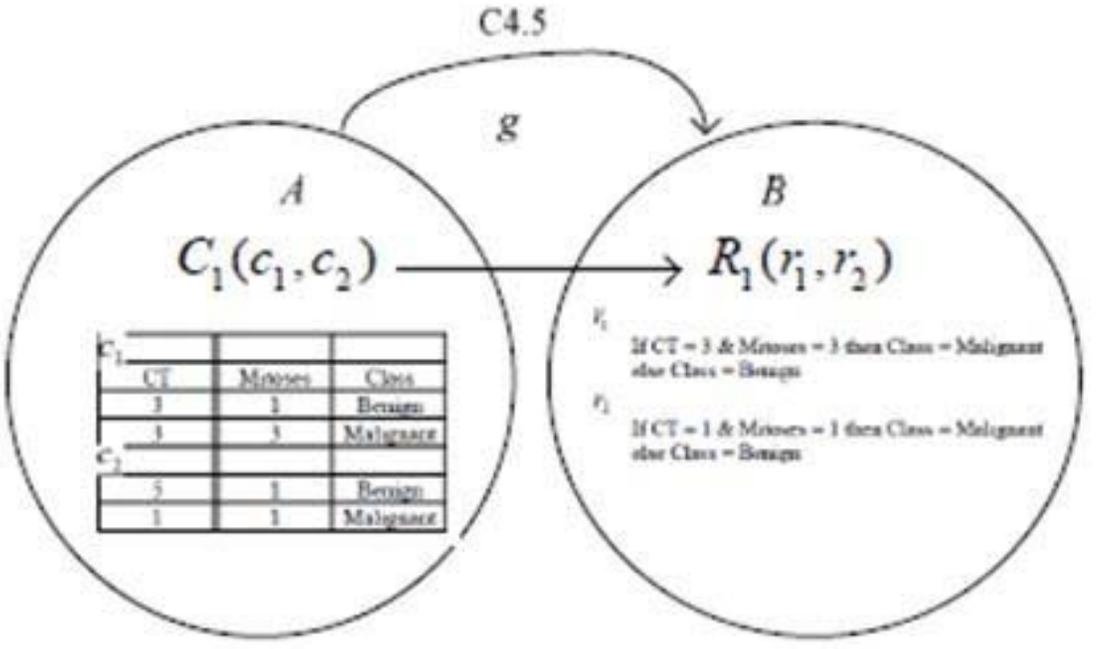

Figure 6. The mapping of function g s_1

Figure 6 shows the mapping between set *A* and set *B* using C4.5 (decision tree) algorithm as a function *g*. The algorithm takes the set of clusters $C_1(c_1, c_2)$ of dataset s_1 and produces the set of rules $R_1(r_1, r_2)$. This is 1:1 mapping because the members of the set *A* are mapped with the corresponding members of the set *B*.

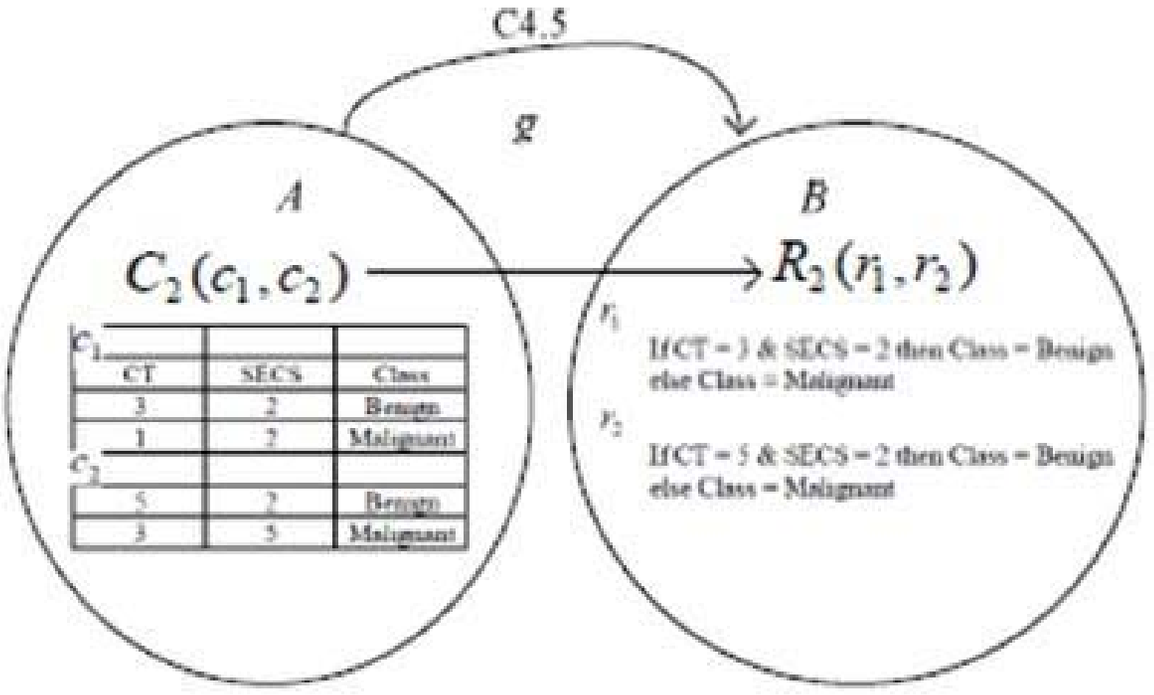

Figure 7. The mapping of function g s_2

Figure 7 shows the mapping between the set *A* and the set *B* using C4.5 (Decision Tree) algorithm as a function *g*. The algorithm takes the set of clusters $C_2(c_1, c_2)$ of dataset s_2 and produces the set of rules $R_2(r_1, r_2)$. This is 1:1 mapping because the members of the set *A* are mapped with the corresponding members of the set *B*.

The figures 6 and 7 show that function *g* is 1:1 function. Hence *g* is a function because it satisfies the conditions of the function; therefore C4.5 algorithm is a function. For our own convenience we are putting if-then-else in the rules created by the algorithm C4.5. The process of creating rules does not place if-then-else in any of the rules.

3. $h(DataVisualization) : B \rightarrow C$

The algorithm Data Visualization takes the rules as input and produces 2D graphs as output. The function *h* is illustrated in equation (4).

$$h(R_k) = V_k \tag{4}$$

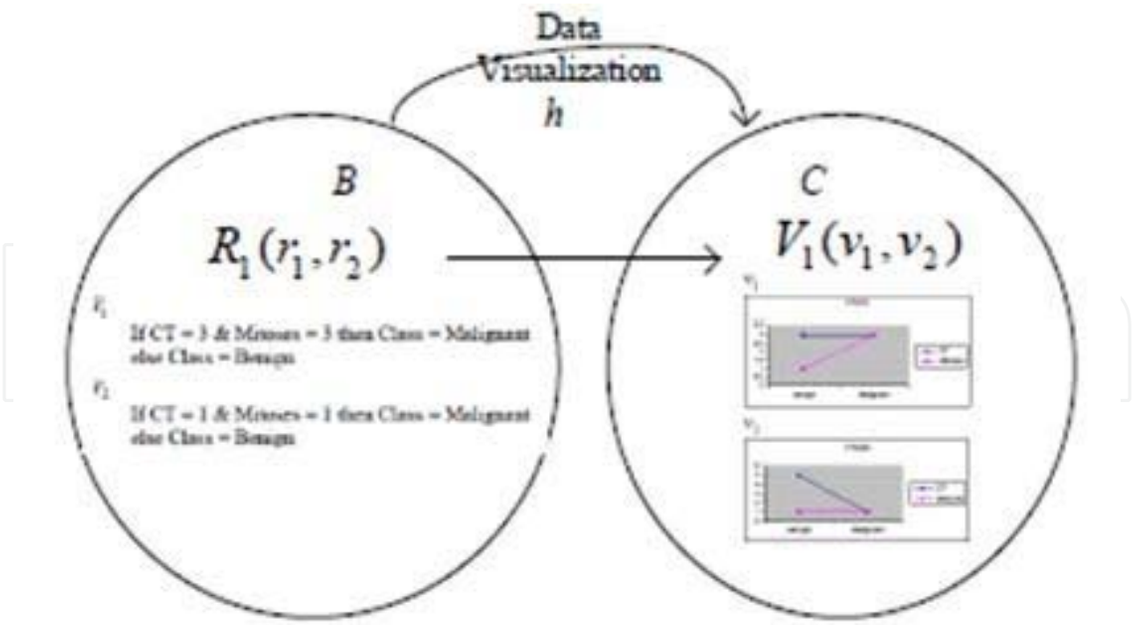

Figure 8. The mapping of function h s_1

Figure 8 shows the mapping between the set B and the set C using Data Visualization algorithm as a function h. The algorithm takes the set of rules $R_1(r_1, r_2)$ of set of cluster $C_1(c_1, c_2)$ of dataset s_1 and produces the set of 2D graphs $V_1(v_1, v_2)$. This is 1:1 mapping because the members of the set B are mapped with the corresponding members of the set C.

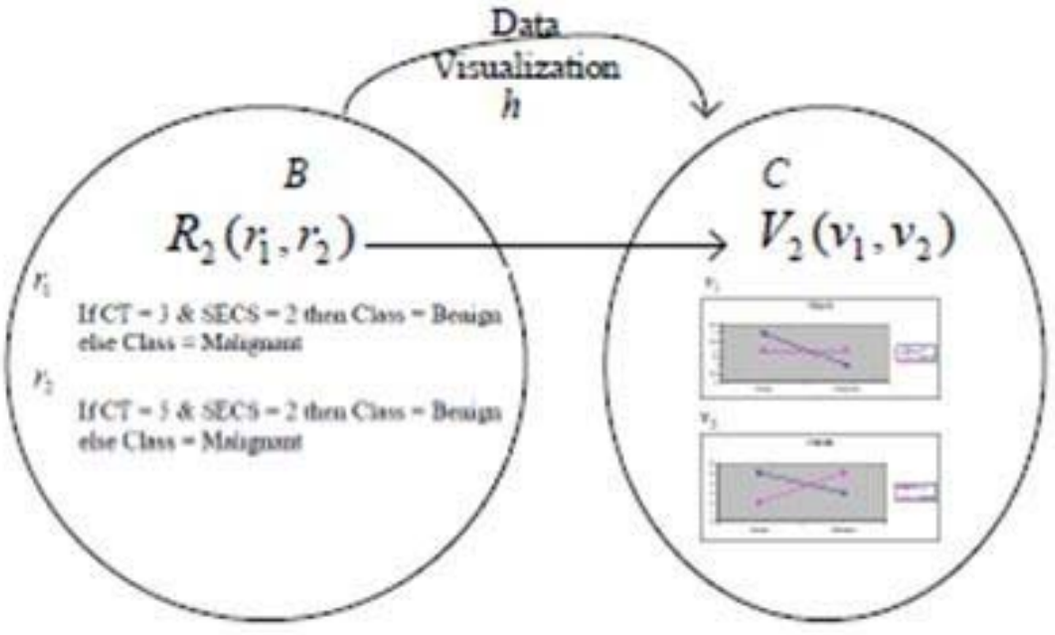

Figure 9. The mapping of function h s_2

Figure 9 shows the mapping between the set B and the set C using Data Visualization algorithm as a function h. The algorithm takes the set of rules $R_2(r_1, r_2)$ of set of cluster

$C_2(c_1, c_2)$ of dataset s_2 and produces the set of 2D graphs $V_2(v_1, v_2)$. This is 1:1 mapping because the members of the set *B* are mapped with the corresponding members of the set *C*.

The figures 8 and 9 show that function *h* is 1:1 function. Hence *h* is a function because it satisfies the conditions of the function, therefore Data Visualization (2D graphs) is a function. The purpose of 2D graph is to identify the type of relationship if any between the attributes. The graph is used when a variable exists which is being tested and in this case the attribute or variable 'class' is a test attribute. We further demonstrate the proposed unified data mining theory through the following two cases:

Case 1: BreastCancer, a medical dataset of four attributes is chosen as example to explain the theory discussed above. We create two vertical partitions, these partitions are the actual inputs of clustering algorithm and we want to create two clusters of each partition. Similarly, we will produce rules and 2D graphs of each partition and finally we will take one as output called knowledge after evaluating and interpreting all the obtained results. The whole process is discussed below:

$$S = \{s_1, s_2\}$$
$$A = \{C_1(c_1, c_2), C_2(c_1, c_2)\}$$
$$B = \{R_1(r_1, r_2), R_2(r_1, r_2)\}$$
$$C = \{V_1(v_1, v_2), V_2(v_1, v_2)\}$$

Dataset S

CT	SECS	Mitoses	Class
3	2	1	Benign
5	2	1	Benign
1	2	1	Malignant
3	5	3	Malignant

The set of data of dataset S is shown in the tables below:

CT	Mitoses	Class
3	1	Benign
5	1	Benign
1	1	Malignant
3	3	Malignant

Table 1. s_1

CT	SECS	Class
3	2	Benign
5	2	Benign
1	2	Malignant
3	5	Malignant

Table 2. s_2

The set of clusters of set of data are shown in the following tables:

CT	Mitoses	Class
3	1	Benign
3	3	Malignant

Table 3. $C_1(c_1)$

CT	Mitoses	Class
5	1	Benign
1	1	Malignant

Table 4. $C_1(c_2)$

CT	SECS	Class
3	2	Benign
1	2	Malignant

Table 5. $C_2(c_1)$

CT	SECS	Class
5	2	Benign
3	5	Malignant

Table 6. $C_2(c_2)$

The set of rules of set of clusters are described below:

$R_1(r_1)$

If CT = 3 & Mitoses = 3 then Class = Malignant
else Class = Benign

$R_1(r_2)$

If CT = 1 & Mitoses = 1 then Class = Malignant
else Class = Benign

$R_2(r_1)$

If CT = 3 & SECS = 2 then Class = Benign
else Class = Malignant

$R_2(r_2)$

If CT = 5 & SECS = 2 then Class = Benign
else Class = Malignant

The following figures show the set of 2D graphs of set of rules:

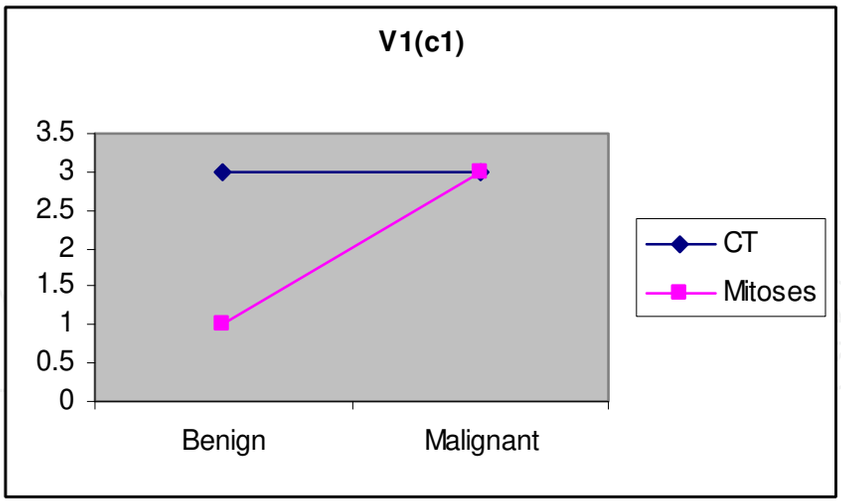

Figure 10. $V_1(v_1)$

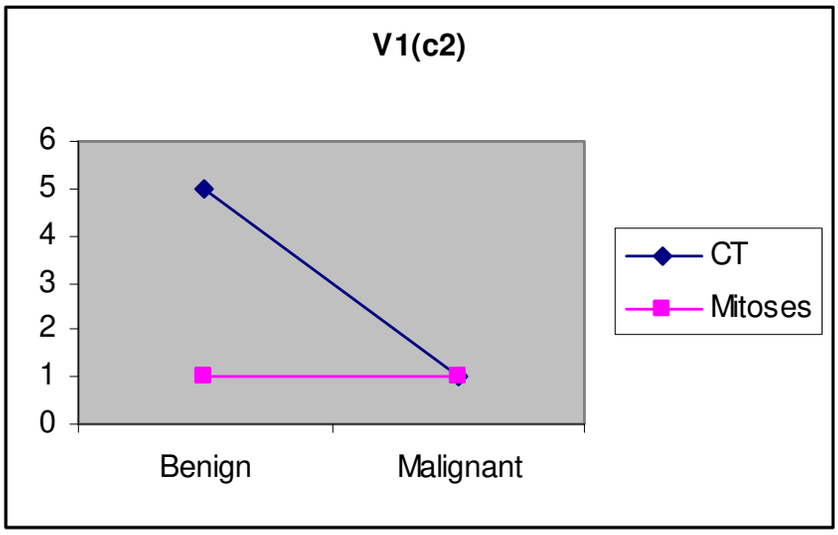

Figure 11. $V_1(v_2)$

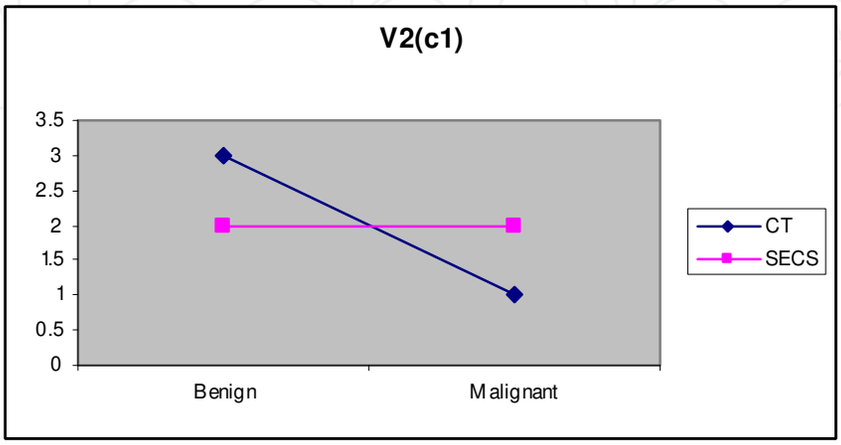

Figure 12. $V_2(v_1)$

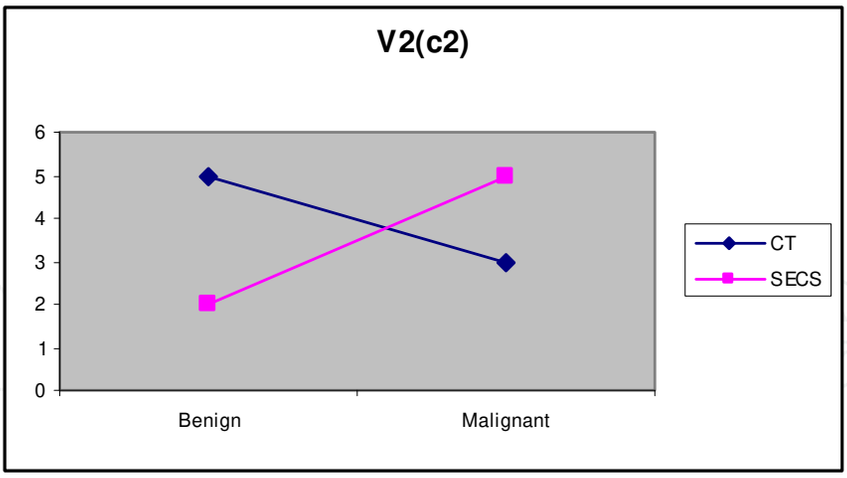

Figure 13. $V_2(v_2)$

$V_2(v_2)$ We interpret and evaluate the results obtained from the set $V_1(v_1, v_2)$. The structure of 2D graphs in figure 10 and 11 is identical. The result obtained from these graphs is that if attributes 'CT' and 'Mitoses' have the same values then the patient has 'Malignant' class of BreastCancer otherwise 'Benign' class of BreastCancer. The interpretation and evaluation of the set $V_2(v_1, v_2)$ show that the 2D graphs in figure 12 and 13 are similar. The result achieved from these graphs is if attributes 'CT' and 'SECS' have variable values then the patient has 'Benign' class of BreastCancer otherwise 'Malignant' class of Breast Cancer.

Table 7 summarizes the steps involved in the theory of unified data mining processes through the composition of functions for case 1 discussed above.

S	f	$g \circ f$	$h \circ g \circ f$
s_1	$C_1(c_1, c_2)$	$R_1(r_1, r_2)$	$V_1(v_1, v_2)$
s_2	$C_2(c_1, c_2)$	$R_2(r_1, r_2)$	$V_2(v_1, v_2)$

Table 7. Composition of the Functions

In this way the knowledge can be extracted from the given dataset S through the unified process of the composition of clustering, classification and visualization followed by interpretation and evaluation of the results which depend on the user's selection according to his business requirements.

Case 2: Diabetes, a medical dataset of nine attributes is selected. We create four vertical partitions, these partitions are actual inputs of clustering algorithm and three clusters for each partition are created. Similarly, we will produce rules and 2D graphs of each partition and finally we will select one partition as output called knowledge after evaluation and interpreting all the results. The whole process is discussed below:

$$S = \{s_1, s_2, s_3, s_4\}$$
$$A = \{C_1(c_1, c_2, c_3), C_2(c_1, c_2, c_3), C_3(c_1, c_2, c_3), C_4(c_1, c_2, c_3)\}$$
$$B = \{R_1(r_1, r_2, r_3), R_2(r_1, r_2, r_3), R_3(r_1, r_2, r_3), R_4(r_1, r_2, r_3)\}$$
$$C = \{V_1(v_1, v_2, v_3), V_2(v_1, v_2, v_3), V_3(v_1, v_2, v_3), V_4(v_1, v_2, v_3)\}$$

Dataset S

NTP	PGC	DBP	TSFT	HSI	BMI	DPF	Age	Class
4	148	72	35	0	33.6	0.627	50	Cat 1
2	85	66	29	0	26.6	0.351	31	Cat 2
2	183	64	0	0	23.3	0.672	32	Cat 1
1	89	66	23	94	28.1	0.167	21	Cat 2
2	137	40	35	168	43.1	2.288	33	Cat 2

The set of data of the dataset S is shown in tables below:

NTP	PGC	Class
4	148	Cat 1
2	85	Cat 2
2	183	Cat 1
1	89	Cat 2
2	137	Cat 2

Table 8. s_1

DBP	TSFT	Class
72	35	Cat 1
66	29	Cat 2
64	0	Cat 1
66	23	Cat 2
40	35	Cat 2

Table 9. s_2

HSI	BMI	Class
94	28.1	Cat 2
168	43.1	Cat 2
88	31	Cat 2
543	30.5	Cat 1
200	25.5	Cat 1

Table 10. s_3

DPF	Age	Class
0.627	50	Cat 1
0.351	31	Cat 2
0.672	32	Cat 1
0.167	21	Cat 2
0.2	39	Cat 1

Table 11. s_4

The set of clusters of set of data are shown in the following tables:

NTP	PGC	Class
4	148	Cat 1
2	85	Cat 2

Table 12. $C_1(c_1)$

NTP	PGC	Class
2	183	Cat 1
1	89	Cat 2

Table 13. $C_1(c_2)$

NTP	PGC	Class
2	137	Cat 2

Table 14. $C_1(c_3)$

DBP	TSFT	Class
72	35	Cat 1

Table 15. $C_2(c_1)$

DBP	TSFT	Class
66	29	Cat 2
64	0	Cat 1

Table 16. $C_2(c_2)$

DBP	TSFT	Class
66	23	Cat 2
40	35	Cat 2

Table 17. $C_2(c_3)$

HSI	BMI	Class
94	28.1	Cat 2
168	43.1	Cat 2

Table 18. $C_3(c_1)$

HSI	BMI	Class
88	31	Cat 2

Table 19. $C_3(c_2)$

HSI	BMI	Class
543	30.5	Cat 1
200	25.5	Cat 1

Table 20. $C_3(c_3)$

DPF	Age	Class
0.167	21	Cat 2
0.2	39	Cat 1

Table 21. $C_4(c_1)$

DPF	Age	Class
0.627	50	Cat 1
0.351	31	Cat 2

Table 22. $C_4(c_2)$

DPF	Age	Class
0.672	32	Cat 1

Table 23. $C_4(c_3)$

The set of rules of set of clusters are described below:

$R_1(r_1)$

If NPT = 4 and PGC = 85 then Class = Cat 2

else Class = Cat 1

$R_1(r_2)$

If NPT = 2 and PGC = 183 then Class = Cat 1
else Class = Cat 2

$R_1(r_3)$

If NPT = 2 and PGC = 137 then Class = Cat 2

$R_2(r_1)$

If DBP = 32 and TSFT =35 then Class = Cat 1

$R_2(r_2)$

If DBP = 66 and TSFT = 29 then Class = Cat 2
else Class = Cat 1

$R_2(r_3)$

If DBP = 40 and TSFT = 23 then Class = Cat 2

$R_3(r_1)$

If HSI = 94 and BMI = 23.1 then Class = Cat 2

$R_3(r_2)$

If HSI = 88 and BMI = 31 then Class = Cat 2

$R_3(r_3)$

If HSI = 200 and BMI = 30.5 then Class = Cat 1

$R_4(r_1)$

If DPF = 0.2 and Age = 21 then Class = Cat 2
else Class = Cat 1

$R_4(r_2)$

If DPF = 0.637 and Age = 50 then Class = Cat 1
else Class = Cat 2

$R_4(r_3)$

If DPF = 0.672 and Age = 32 then Class = Cat 2

The following figures show the set of 2D graphs of set of rules:

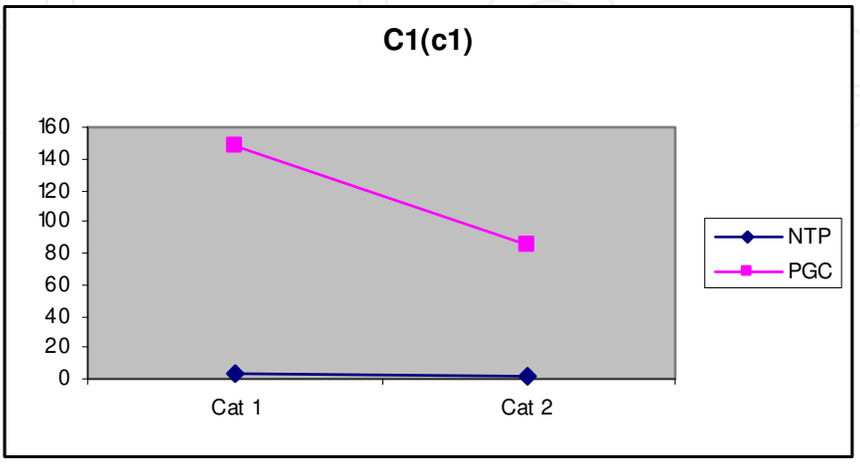

Figure 14. $V_1(v_1)$

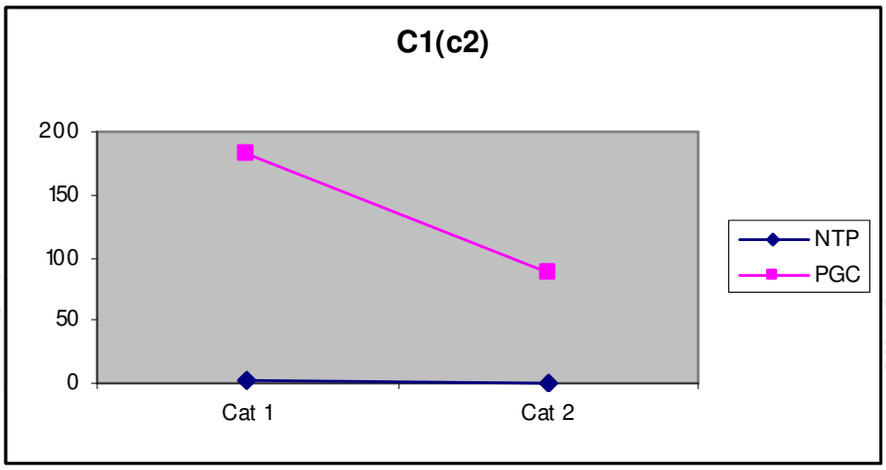

Figure 15. $V_1(v_2)$

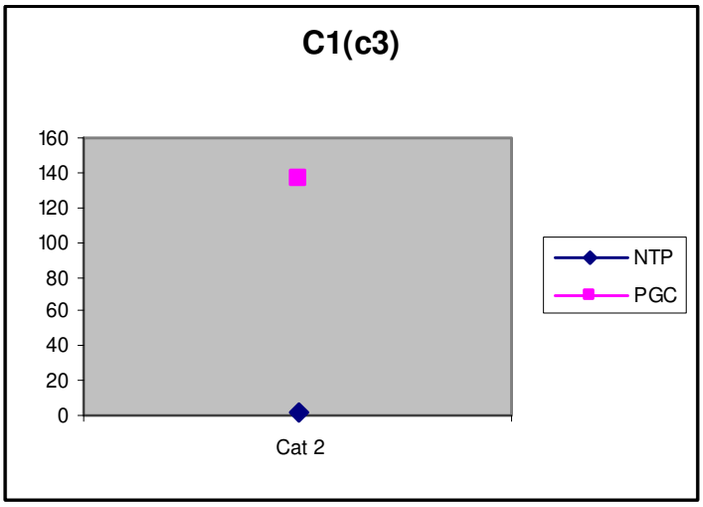

Figure 16. $V_1(v_3)$

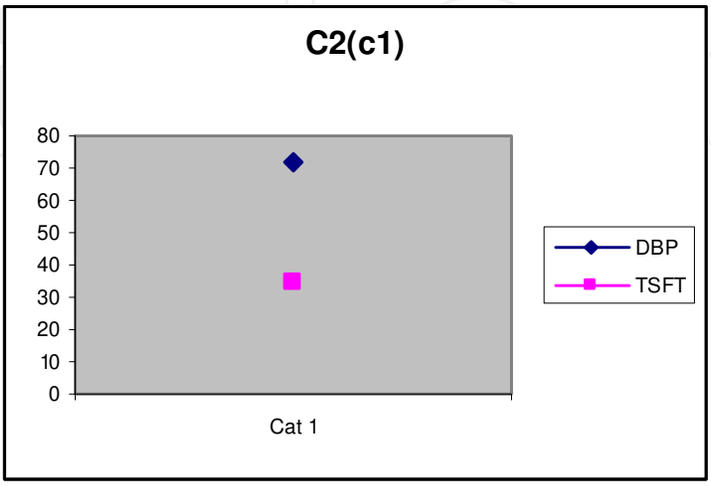

Figure 17. $V_2(v_1)$

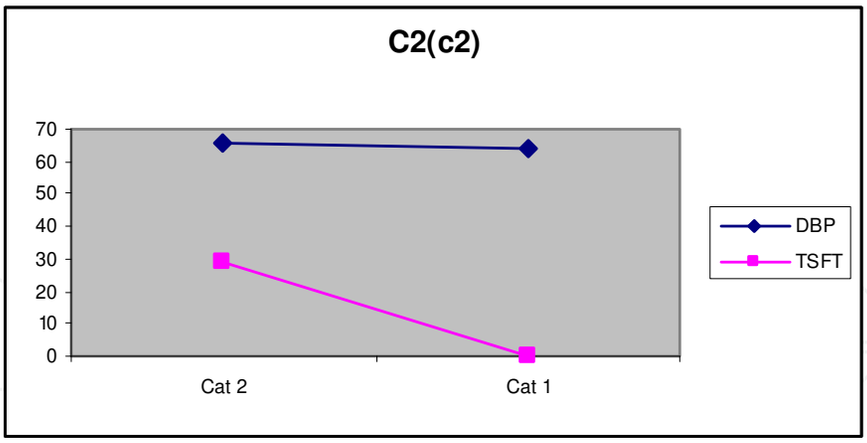

Figure 18. $V_2(v_2)$

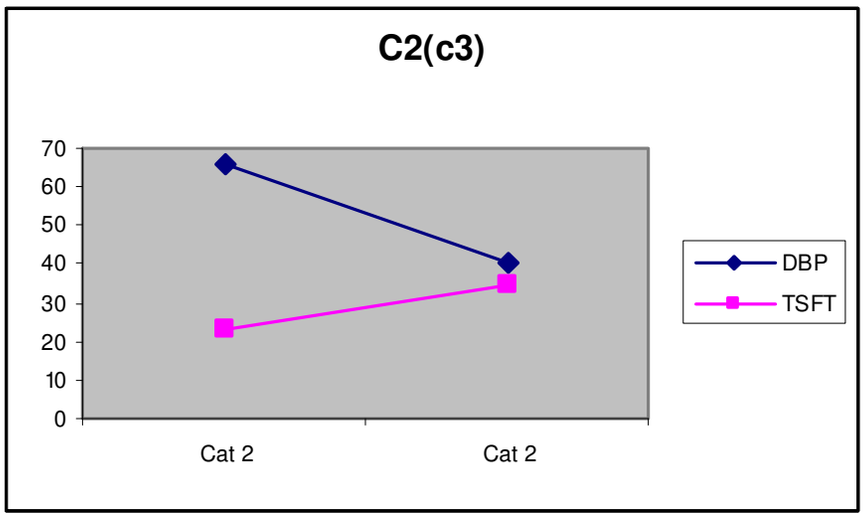

Figure 19. $V_2(v_3)$

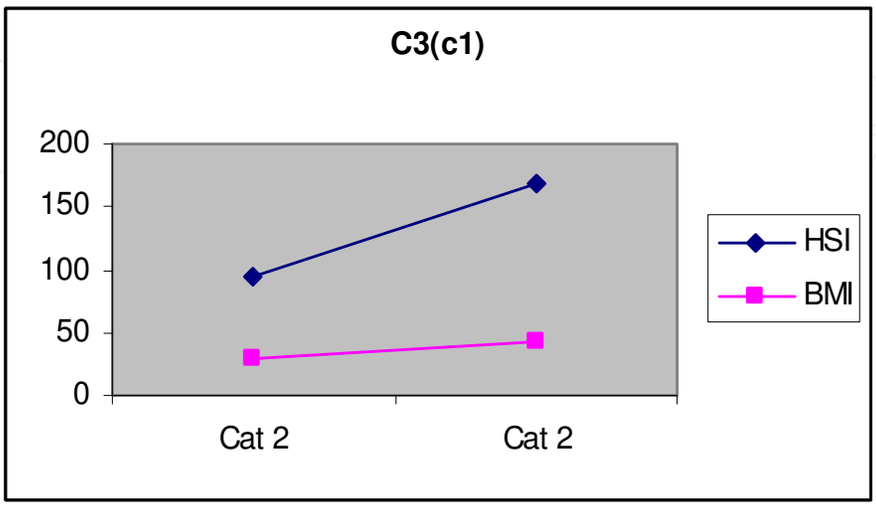

Figure 20. $V_3(v_1)$

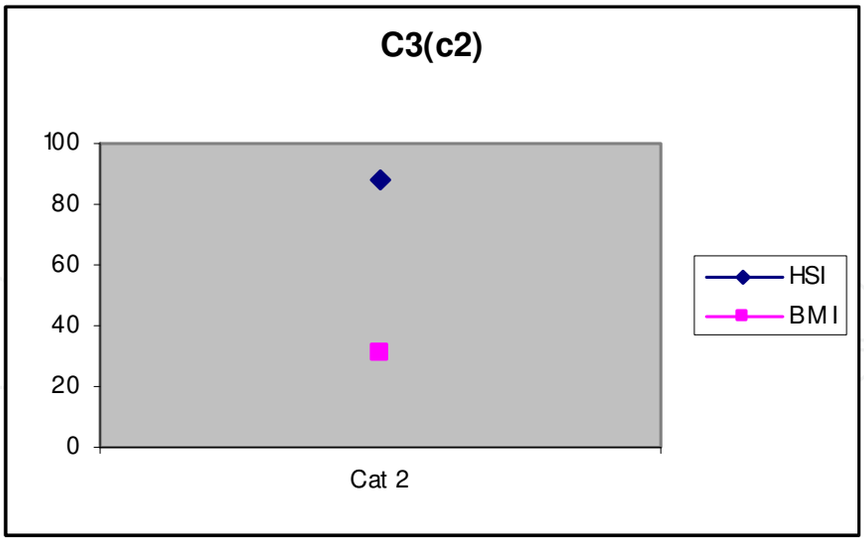

Figure 21. $V_3(v_2)$

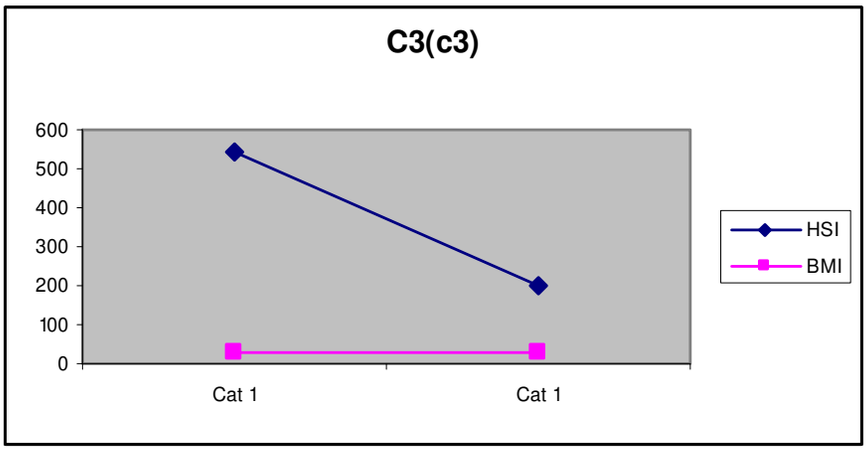

Figure 22. $V_3(v_3)$

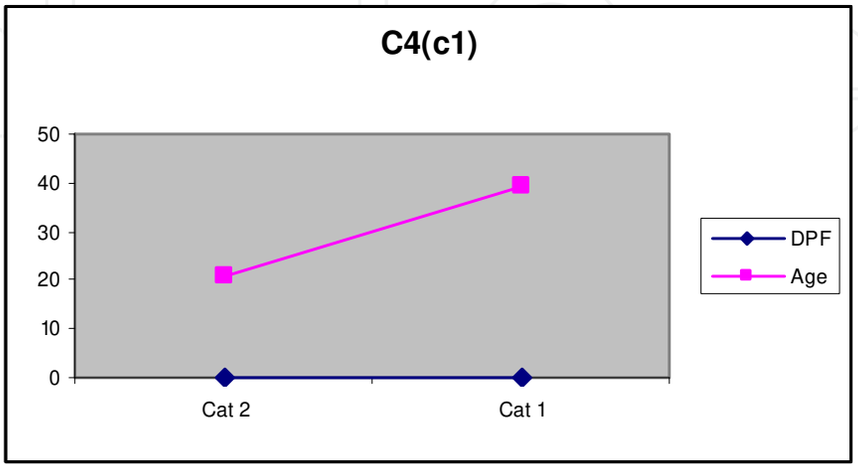

Figure 23. $V_4(v_1)$

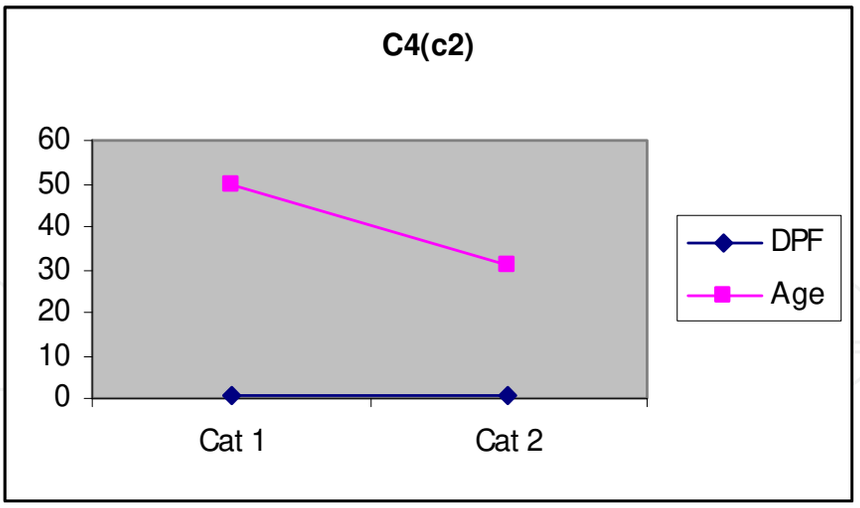

Figure 24. $V_4(v_2)$

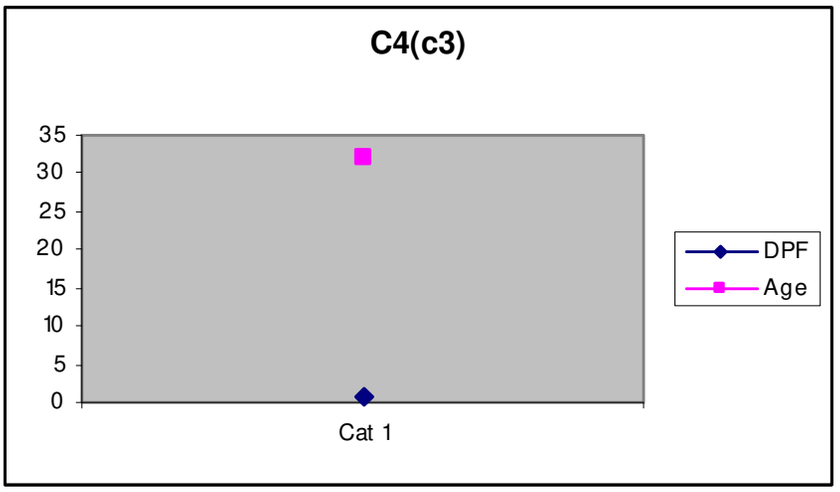

Figure 25. $V_4(v_3)$

The results of set $V_1(v_1,v_2,v_3)$ are interpreted and evaluated which show that the structure of 2D graphs in figure 14 and 15 is identical. The result obtained from these graphs is if attributes 'NPT' and 'PGC' have the variable values then the patient has either diabetes of 'Cat 1' or 'Cat 2'. But the 2D graph in figure 16 shows that if the attributes are of variable values then the patient has 'Cat 2' diabetes. The set $V_2(v_1,v_2,v_3)$ shows that the structure of 2D graphs in figure 18 and 19 is almost the same as 2D graphs in figure 14 and 15. The result obtained from these graphs is if attributes 'DBP' and 'TSFT' have the variable values then the patient has either diabetes of 'Cat 1' or 'Cat 2'. The 2D graph in figure 17 is similar to 2D graph in figure 16, which shows that if the attributes are of variable values then the patient has 'Cat 2' diabetes. The set $V_3(v_1,v_2,v_3)$ illustrates that the structure of 2D graphs in figure 20 and 22 is the same as 2D graphs in figures 14, 15, 18 and 19. The result obtained from these graphs is if attributes 'HSI' and 'BMI' have the variable values then the patient has either diabetes of 'Cat 1' or 'Cat 2'. The 2D graph in figure 21 is similar to 2D graph in

figures 16 and 17, which shows that if the attributes are of variable values then the patient has 'Cat 2' diabetes. Finally, the set $V_4(v_1, v_2, v_3)$ demonstrates that the structure of 2D graphs in figure 23 and 24 is the same as 2D graphs in figures 14, 15, 18, 19, 20 and 22. The result obtained from these graphs is if attributes 'DPF' and 'Age' have the variable values then the patient has either diabetes of 'Cat 1' or 'Cat 2'. The 2D graph in figure 25 is similar to 2D graph in figures 16, 17 and 21, which shows that if the attributes are of variable values then the patient has 'Cat 2' diabetes.

Table 24 summarizes the steps to unify the data mining process through the composition of functions for case 2 discussed above.

S	f	$g \circ f$	$h \circ g \circ f$
s_1	$C_1(c_1, c_2, c_3)$	$R_1(r_1, r_2, r_3)$	$V_1(v_1, v_2, v_3)$
s_2	$C_2(c_1, c_2, c_3)$	$R_2(r_1, r_2, r_3)$	$V_2(v_1, v_2, v_3)$
s_3	$C_3(c_1, c_2, c_3)$	$R_3(r_1, r_2, r_3)$	$V_3(v_1, v_2, v_3)$
s_4	$C_4(c_1, c_2, c_3)$	$R_4(r_1, r_2, r_3)$	$V_4(v_1, v_2, v_3)$

Table 24. Composition of the Functions

Thus the knowledge can be extracted from the given dataset S through the unified process of the composition of clustering, classification and visualization followed by interpretation and evaluation of the results which depend on the user's selection according to his business requirements. It also shows that the proposed unified data mining theory is comprehensiveness, precise, consistent and correct which are the prerequisites for a unified theory.

5. Unified Data Mining Tool (UDMTool)

The Unified Data Mining Tool (UDMTool) is a new and better next generation solution which is a unified way of architecting and building software solutions by integrating different data mining algorithms. The figure 26 depicts the architecture of UDMTool.

The figure 26 is an architecture of UDTMTool based on proposed UDMT. The dataset is inputed, there are many types of datasets like, numeric, categorical, multimedia, text and many more, the selection crterion will take the whole dataset, analayises the data and produces 'under-fitted' or 'over-fitted' values of the dataset, on the bases of these values the appropriate data mining algorithms are selected and the data is passed to the unified data mining processes for 'knowledge' as output. The UDMTool has four parts:

i. Datasets
ii. The Selection Criterion (Data Analayiser)
iii. The Unified Data Mining Processes and
iv. Finally, 'the knowledge', as an output.

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

Selection criterion S = {VC Dimension, AIC, BIC} The output of S = Over-fitted, Under-fitted}

Unified Process U = {Clustering, Classification, Visualization, Interpretation} [Assume that 'clustering' will be the first step followed by the rest of the steps]

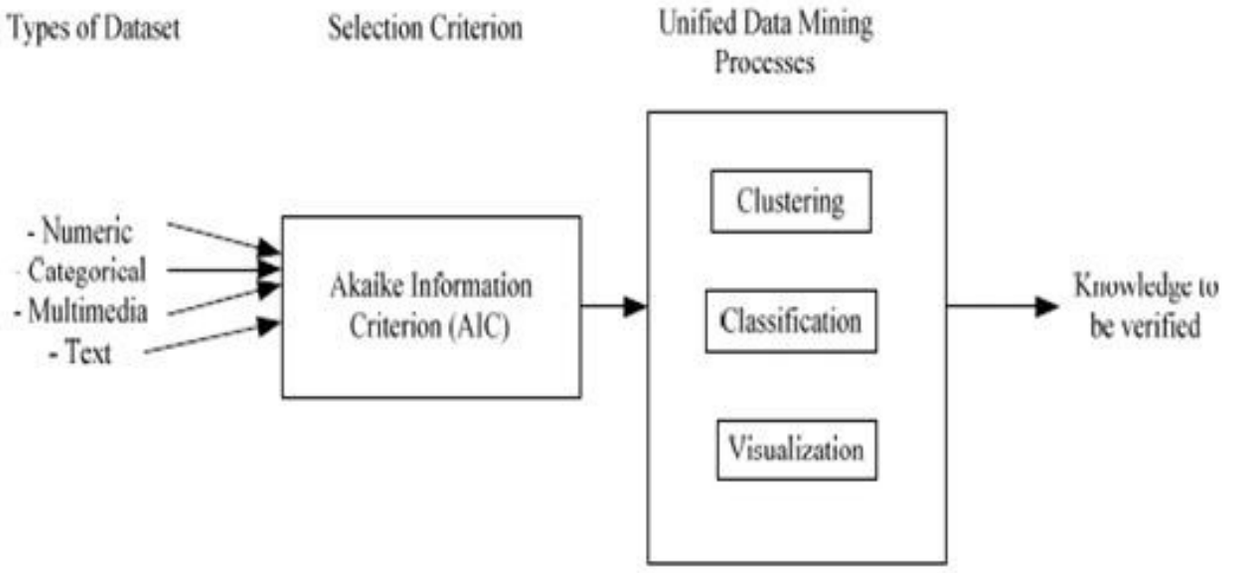

Figure 26. The Architecture of the UDMTool

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 Model cannot play any role in this regard]

Step 1. Input D
If D # Numeric Stop/Exit
else go to step 2

Step 2. S takes the dataset D and generates the values for S
If over-fitted and under-fitted then again pre-process the inputted dataset D
else create the appropriate vertical partitions of the dataset D and go to step 3

Step 3. One by one take these vertical partitioned datasets
Up = First, create clusters, then classify each cluster, then draw 2D graphs of each cluster, then evaluate and interpret these results and finally produce the 'knowledge' K as output.

Step 4. K = {Accepted, Rejected}
Rejected: Exit/Stop and select another dataset
Accepted: Ok/Stop. The process is successfully completed and the output meets the business goals.

These steps are further illustrated in a flow chart in figure 27.

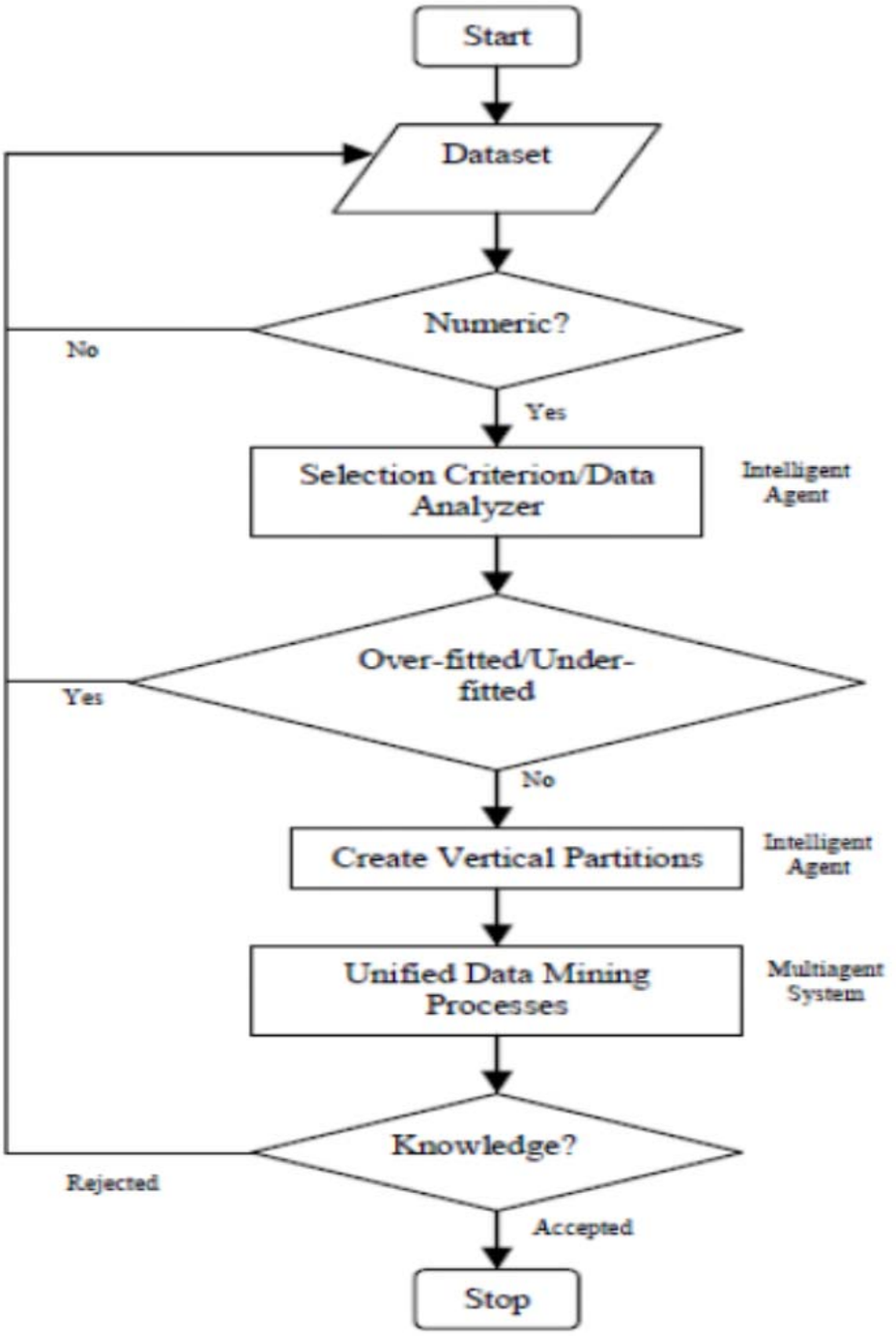

Figure 27. The Flow Chart of the UDMTool

The explanation of the flow chart is: A numeric dataset of 'n' attributes i.e. a data file is the starting point of this flowchart. An intelligent agent takes the dataset and analyzes the data using AIC selection criterion. An agent is for the selection criterion. The under-fitted or over-fitted values of the dataset show the errors in the data. This requires the proper cleansing of the data. The second agent creates the 'm' vertical partitions of the dataset where $m = \frac{n-1}{2}$, where *'n'* and *'m'* both are non-zero positive integers. Finally, these partitions are taken by a MAS, which unifies the data mining processes such as clustering, classification, visualization and interpretation and evaluation and the 'knowledge' is extracted as output. The knowledge is either accepted or rejected. Thus, multiagent systems (MASs) for unified theory of data mining processes, one agent is for selection criterion and the other agent is for vertical partition of the dataset.

Figure 28 explains the framework of multiagent systems (MASs) used in UDMTool.

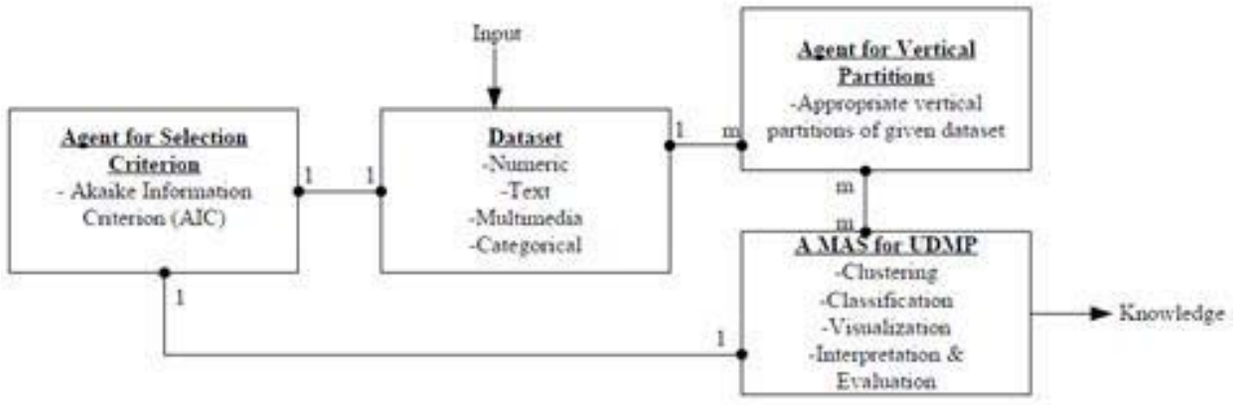

Figure 28. The Framework of MASs of the UDMTool

A well-prepared dataset is a starting input of this framework. First, intelligent agent compute the value model of selection AIC, which is used to select appropriate data mining algorithm and the second intelligent agent creates the vertical partitions, which are the inputs of UDMP. Finally, the knowledge is extracted, which is either accepted or rejected. The framework of UDMTool, a proposed model is shown in figure 29 below. The design is based on *Gaia methodology*, which is used to build models in the analysis and design phase (Wooldridge, M, Jennings, N. R., Kinny, D., 2000). The relationship between dataset and selection criterion is 1:1 i.e. one dataset and one value for model selection and between dataset and vertical partitions is 1:m i.e. more then one partitions are created for one dataset. The relationship between selection criterion and UDMP is 1:1 i.e. one value of selection model will give one data mining algorithm and finally the relationship between vertical partitions and UDMP is m:m i.e. many partitioned datasets are inputs for UDMP and only one result is produced as knowledge.

The function of the UDMTool is demonstrated in figure 29.

The dataset is a starting input of UDMTool, first intelligent agent compute the value of AIC, which is used to pick the right algorithm from unified data mining process, a MAS, for the given dataset. The second algorithm generates the required and appropriate vertical partitions of the given dataset, which are the inputs of MAS. The knowledge is produced in the forms of 2D graph(s) as the final output, which is verified by the user according to his business rules.

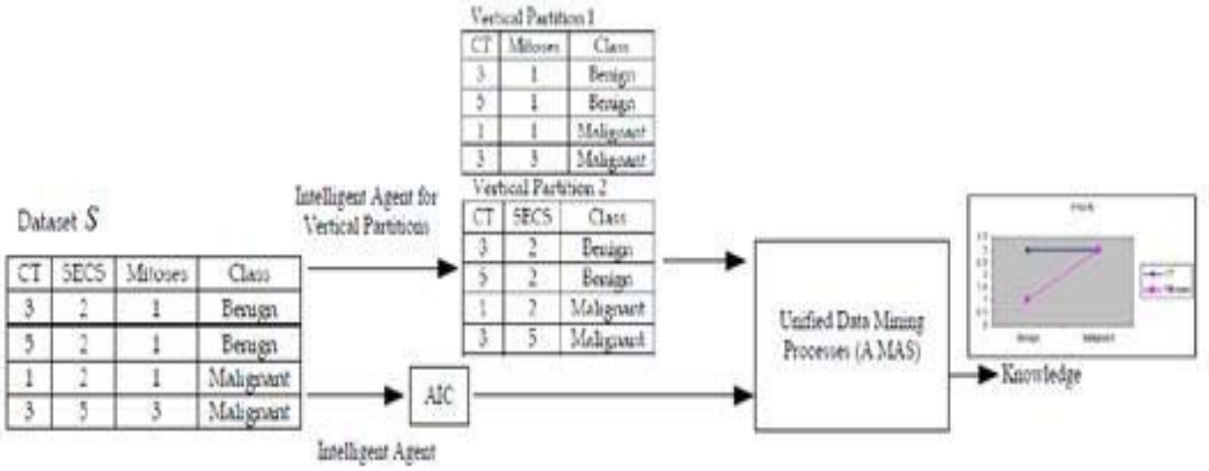

Figure 29. The Function of the UDMTool

The UDMTool is tested on variety of datasets such as 'Diabetes' and 'Breast Cancer', two medical datasets, 'Iris', an agriculture dataset and 'Sales', an account dataset. In this book chapter, we present the results of 'Iris' and 'Sales' datasets. The following figures (2D graphs) are the results and the extracted knowledge from 'Iris', an agriculture dataset using UDMTool:

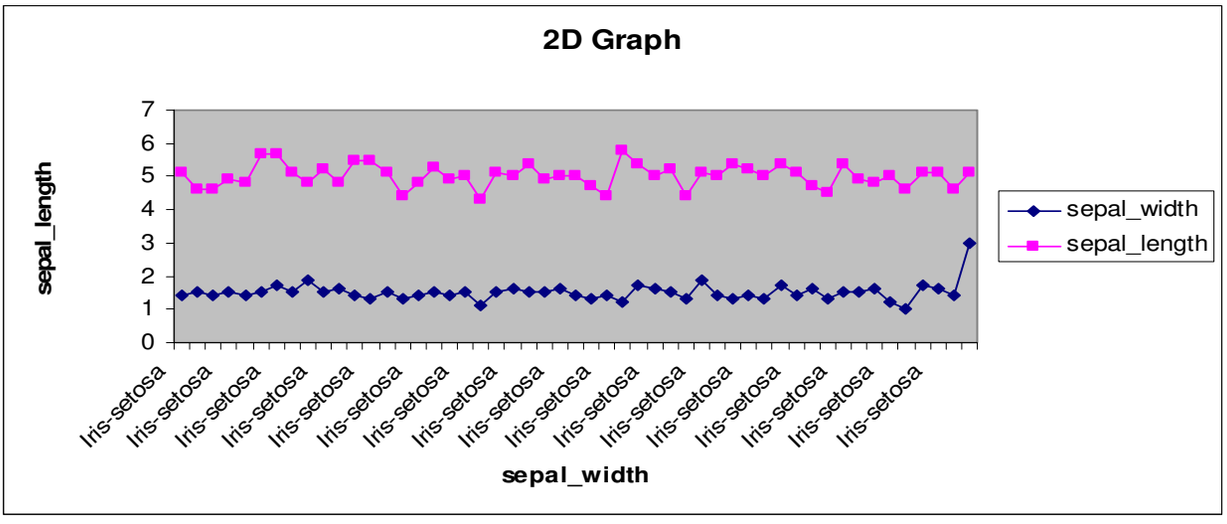

Figure 30. The Graph between 'sepal_length' and 'sepal_width' of 'Iris' dataset

The graph in figure 30 shows that there is no relationship between the attributes 'sepal_width' and 'sepal_length' from the beginning to the end. Both attributes have the distinct values which show the value of attribute 'class' is 'Irissetosa'. Therefore, the result derived from this 2D graph is if these two attributes have no relationship then the class is 'Irissetosa'.

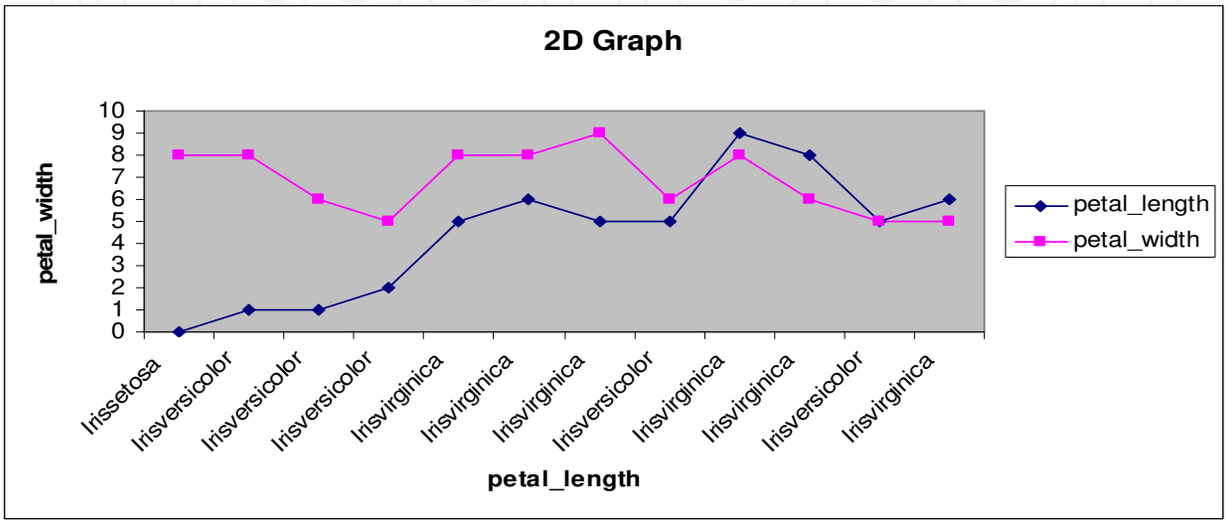

Figure 31. Graph between 'petal_length' and 'petal_width' of 'Iris' dataset

The value of the attributes 'petal_length' and 'petal_width' is almost constant at the beginning and then at the end there is some relationship between the attributes in this graph of figure 31. The graph can be divided into two main regions; the value of the attributes 'petal_length' and 'petal_width' is constant i.e. there is no relationship between the attributes and the attribute 'class' also have the distinct values. In the second region there exists a relationship between the attributes which gives almost the unique value of the attribute 'class' 'Irisvirginica'. The outcome of this graph is that if the value of the attributes is variable then the 'class' is also variable otherwise the value of 'class' is constant which is 'Irisvirginicia'.

The structure of the graph in figure 32 is complex. In the beginning of the graph there is no relationship between the attributes 'petal_length' and 'petal_width', then there exists a relationship between these attributes, again the attributes have distinct values and at the end there is relationship between the attributes. The outcome of this graph is that if there is no relationship between the attributes, the value of attribute 'class' is 'Irisvirginicia' and if there exists a relationship between the attributes then the value of attribute 'class' is 'Irisversicolor'.

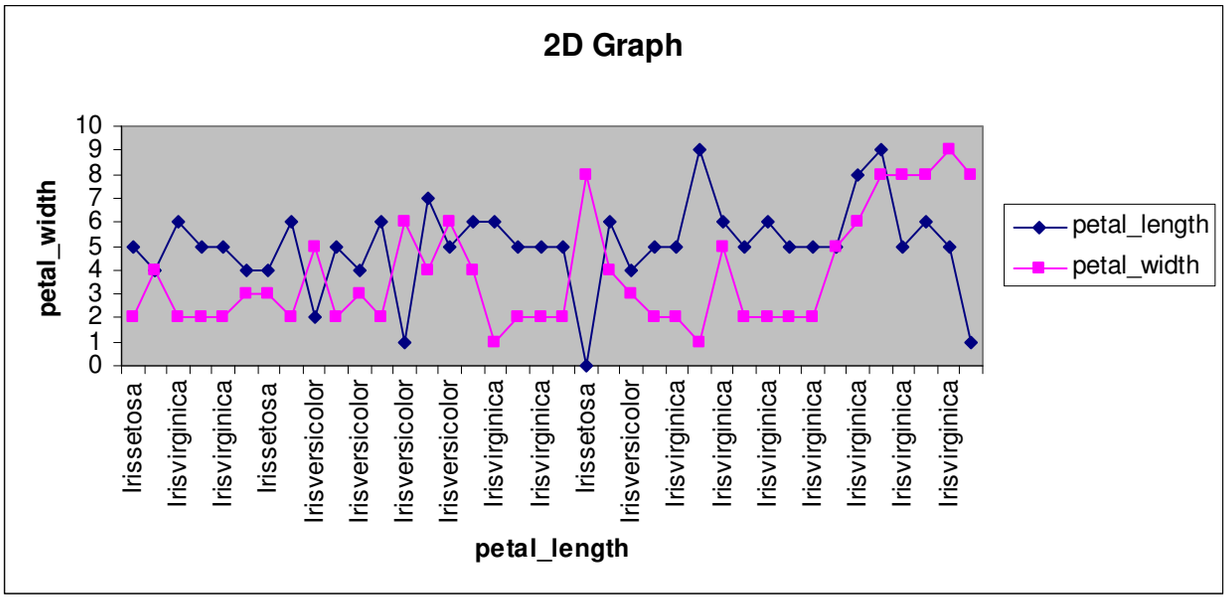

Figure 32. Graph between 'petal_length' and 'petal_width' of 'Iris' dataset

The following figures (2D graphs) are the results and the extracted knowledge from 'Sales', an account dataset using UDMTool:

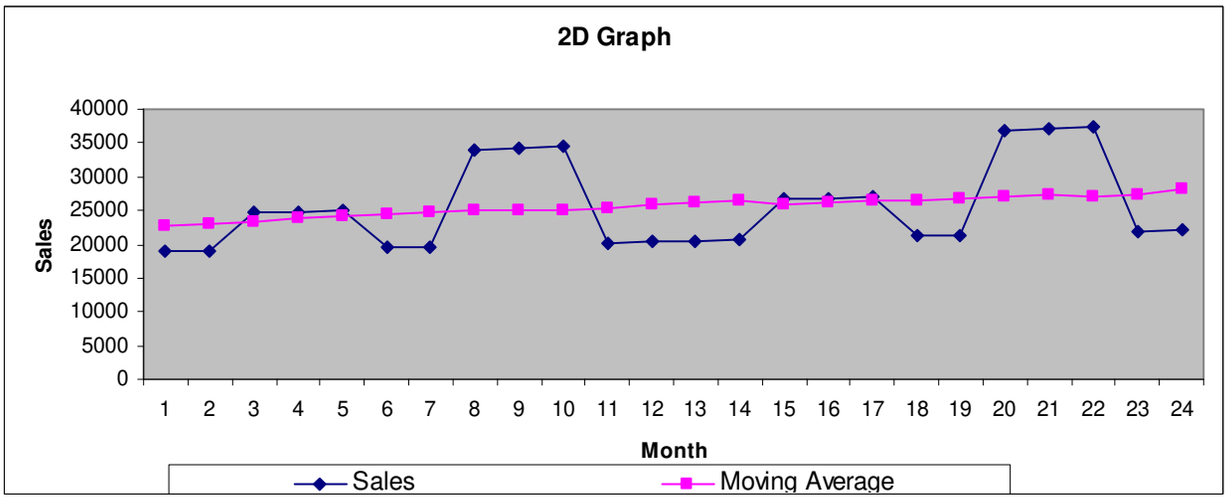

Figure 33. Graph between 'monthly sales' and 'moving average' of 'Sales' dataset

The graph in figure 33 shows a relationship between 'monthly sales' and the 'moving average sales' of the dataset 'Sales'. The graph shows that in the first two months the 'monthly sales' is below is the 'average sales' and then in the next couple of months the 'monthly sales' is almost equal to the 'average sales'. The value of 'monthly sales' is

higher then the 'average sales' during 8, 9, 10, 20, 21 and 22 months, the rest of the period shows that either the sales is low or equal to the 'average sales'. The outcome of this graph is that in these twenty four 'months sales' of the company, twelve months the sales is below the expected 'average sales' and in six months the sales is higher then the expected 'average sales' and in the remain six months the sales is equal to the expected 'average sales'.

The graph in figure 34 shows a relationship between 'average sales' and the 'forecast sales' of the dataset 'Sales'. The graph shows that in the beginning the 'average sales' and the 'forecast sales' are equal, in the middle of the graph the 'average sales' is either higher or equal then the 'forecast sales' values. The graph also shows that the gap between two values is quite significant during 20, 21 and 22 months i.e. the value of 'average sales' is much higher then the 'forecast sales' but at the end of the graph both values are again equal. The outcome of this graph is that during this period 'average sales' is higher then the 'forecast sales' values.

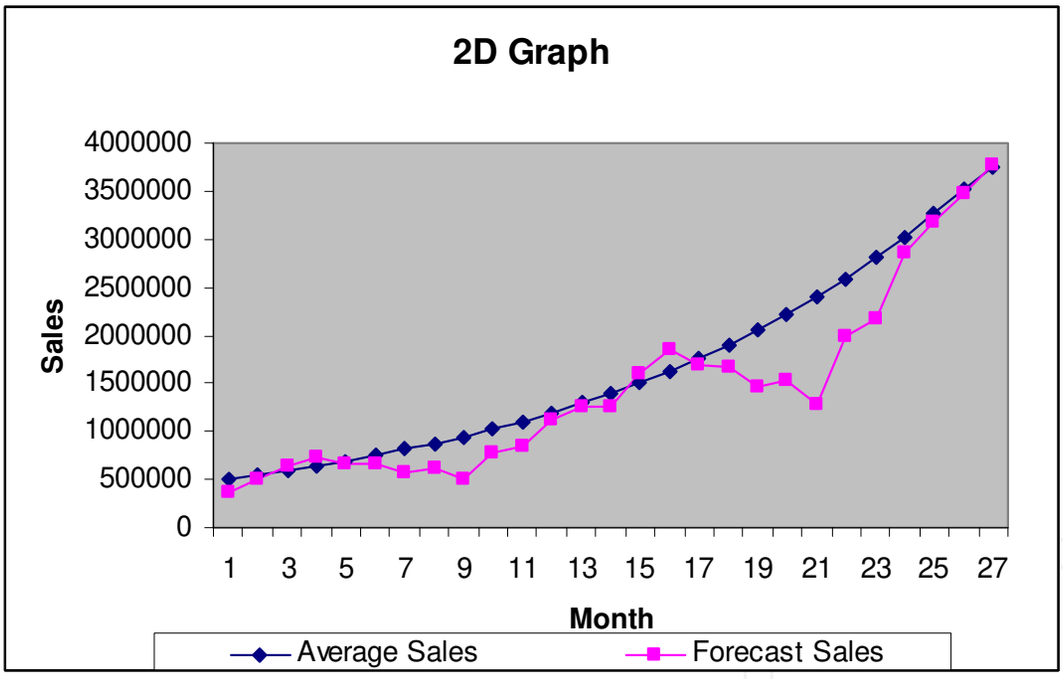

Figure 34. Graph between 'average sales' and 'forecast sales' of 'Sales' dataset

The graph in figure 35 shows a relationship between 'calculated seasonal index' and 'average index' of the dataset 'Sales'. The graph shows that at the beginning the 'calculated seasonal index' is either below or above the 'average index' and the trend remains the same up to the middle of the graph, after this both values are same up to the end of the graph. The 'knowledge' extracted from this graph is that the forecast values of sales are equal to the average sales which shows the trend of profit.

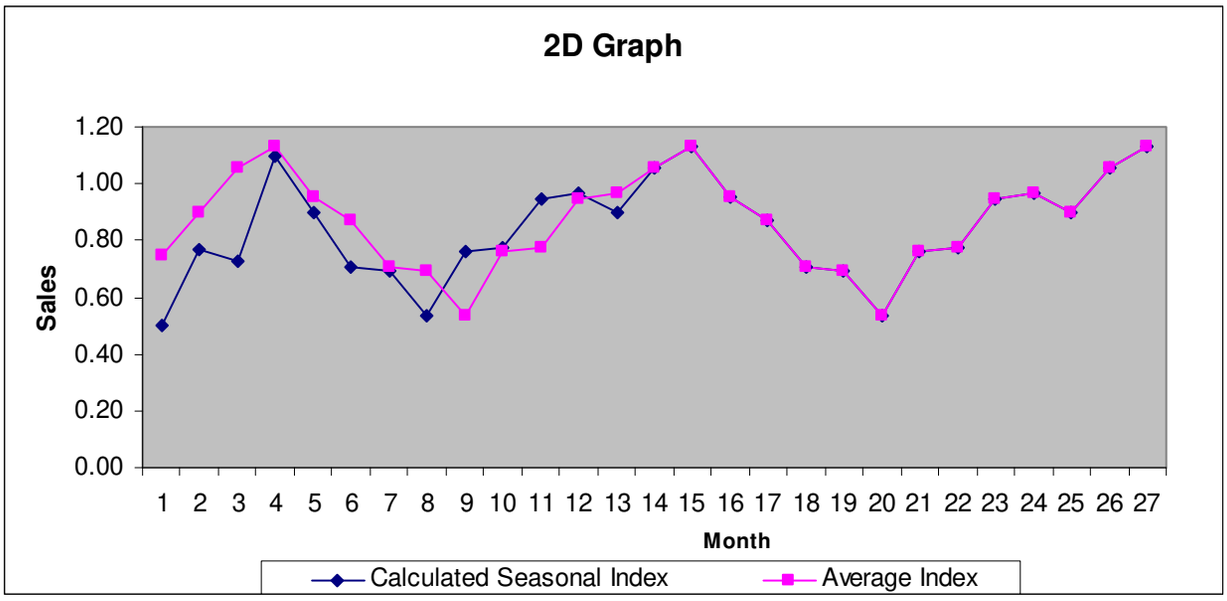

Figure 35. Graph between 'calculated seasonal index' and 'average index' of 'Sales' dataset

6. Conclusion

The book chapter presents a unified data mining theory and then the mathematical formulation of the unified data mining theory (UDMT). The data mining processes; clustering, classification and visualization are unified by means of mathematical functions. The chosen data mining algorithms are proved to be as functions i.e. the algorithms are used as functions. The first function takes the set of data as input, applies K-means clustering data mining algorithm and produces the set of clusters as output. The second function takes the set of clusters as input, applies the C4.5 (Decision Tree) data mining algorithm and produces the set of rules as output and finally the third function takes the set of rules as input, applies data visualization data mining algorithm and produces the set of 2D graphs as output. The functions are the mappings from the set of data to the set of clusters to the set of rules to the set of 2D graphs. From the set of 2D graphs one can interpret and evaluate the results to discover the knowledge i.e. the knowledge is extracted from the given dataset through the unified process of the composition of clustering, classification and visualization followed by interpretation and evaluation of the results. Thus, we attempt to unify the processes of data mining life cycle. In summary we can say, first pass the set of data through the composition of clustering, classification and visualization and then the obtained results are interpreted and evaluated to extract the knowledge. The proposed unified data mining theory covers all the pertinent essentials of the unified theory i.e. generates truly applicable explanations and predictions of the real world problems and does not contradict itself, therefore, we can say that it is comprehensiveness, precise, consistent and correct. The UDMT is tested on variety of datasets and the knowledge is discovered through the unified processes of data mining. On the basis of UDMT, a tool namely; UDMTool, multiagent systems, is developed, which takes dataset as input and

produces the knowledge as output. At this stage we can conclude that the results obtained through the UDMT are satisfactory and consistent. For future consideration, more tests can be conducted to validate the proposed UDMT.

Author details

Dost Muhammad Khan
Department of Computer Science & IT, The Islamia University of Bahawalpur, Pakistan;
School of Innovative Technologies & Engineering, University of Technology Mauritius (UTM), Mauritius

Nawaz Mohamudally
Consultancy & Technology Transfer Centre, Manager, University of Technology, Mauritius (UTM), Mauritius

D. K. R. Babajee
Department of Applied Mathematical Sciences, SITE, University of Technology, Mauritius (UTM), Mauritius

Acknowledgement

The first author is thankful to The Islamia University of Bahawalpur, Pakistan for providing financial assistance to carry out this research activity under HEC project 6467/F – II.

7. References

Brog, Xavier., (2011). Unified Theory Foundations, Blaze Labs URL http://blazelabs.com/f-u-intro.asp

Dalton, J., (1808). *A New System of Chemical Philosophy*

Darwin, C., (1859). *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life*

Das, Somenath, (2007). Unified data mining engine as a system of patterns, *Master's Theses*. Paper3440.http://scholarworks.sjsu.edu/etd_theses/3440

Davidson, Ian, (2002). Understanding K-Means Non-hierarchical Clustering, SUNY Albany– Technical Report

Dhillon IS, Guan Y, Kulis B, (2004). Kernel k-means: spectral clustering and normalized cuts, KDD 2004, pp 551–556

Dries. Lou van den. (2007). Mathematical Logic, Math 570, Lecture Notes, Fall Semester 2007

Gray RM, Neuhoff DL, (1998). Quantization, IEEE Trans Inform Theory 44(6):2325–2384

Grossman. Robert, Kasif. Simon, Moore. Reagan, Rocke. David., & Ullman. Jeff, (1998). Data Mining Research: Opportunities and Challenges, *A Report of three NSF Workshops on Mining Large, Massive, and Distributed Data*, (Draft 8.4.5)

Hunt EB, Marin J, Stone PJ, (1996). Experiments in Induction, Academic Press, New York

Jain AK, Dubes RC, (1988). Algorithms for clustering data, Prentice-Hall, Englewood Cliffs

Johnson, P., Ekstedt, M., (2007). In Search of a Unified Theory of Software Engineering, International Conference on Software Engineering Advances (ICSEA 2007)

Khan, Dost Muhammad., & Mohamudally, Nawaz., (2010). 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,* Vol. 1, Number 2, 31 August 2010, pp(85-95), ISSN: 2092-8637

Khan, Dost Muhammad., & Mohamudally, Nawaz., (2010). An Agent Oriented Approach for Implementation of the Range Method of Initial Centroids in K-Means Clustering Data Mining Algorithm, *International Journal of Information Processing and Management,* Volume 1, Number 1, July 2010 (pp 104-113), ISSN: 2093-4009

Khan, Dost Muhammad., & Mohamudally, Nawaz., (2011). An Integration of k-means clustering and Decision Tree (ID3) towards a more Efficient Data Mining Algorithm, Journal of Computing, Volume 3, Issue 12, 2011, pp 76-82, ISSN: 2151-9617

King, G., R. Keohane, and S. Verba, (1994). *Designing Social Inquiry: Scientific Inference in Qualitative Research*, Princeton University Press

Kuhn, T., (1962). *The Structure of Scientific Revolution*, University of Chicago Press

Li, Xining., Ni, JingBo., (2007). Deploying Mobile Agents in Distributed Data Mining, International workshop on High Performance Data Mining and Applications (HPDMA 2007) China

Liu, Bing, (2007). Web Data Mining Exploring Hyperlinks, Contents, and Usage Data, ISBN-13 978-3-540-37881-5, Springer Berlin Heidelberg New York pp. 124 -139

Lloyd SP, (1982). Least squares quantization in PCM. Unpublished Bell Lab. Tech. Note, portions presented at the Institute of Mathematical Statistics Meeting Atlantic City, NJ, September 1957. Also, IEEE Trans Inform Theory (Special Issue on Quantization), vol IT-28, pp 129–137

Lovász. L. , Pelikán. J., Vesztergombi. K. (2003). Discrete Mathematics: Elementary and Beyond, pp. 38, Springer, ISBN: 0-387-95585-2

MacQueen, J.B., (1967). Some Methods for classification and Analysis of Multivariate Observations, *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley, University of California Press, p: 281-297

MathWorksheetsGo (2011). Evaluating Functions. at URL: www.MathWorksheetsGo.com

Mendeleev, D., (1869). *Principles of Chemistry*

Mielikäinen, Taneli., (2004). Inductive Databases as Ranking, In: DaWak, Zaragoza, Spain

Mohamudally, Nawaz., Khan, Dost Muhammad. (2011). Application of a Unified Medical Data Miner (UMDM) for Prediction, Classification, Interpretation and Visualization on Medical Datasets: The Diabetes Dataset Case, P. Perner (Ed.): ICDM 2011, LNAI 6870, Springer-Verlag Berlin Heidelberg pp. 78–95

Moore Andrew W., (2001). VC-dimension for characterizing classifiers, Carnegie Mellon University at URL: www.cs.cmu.edu/~awm

Newell, A., (1990). *Unified Theories of Cognition*, Harvard University Press

Newton, I., (1687). *Philosophiae Naturalis Principia Mathematica*.

Peng, Y., Kou, G., Shi, Y., Chen, Z., (2008). A Descriptive Framework for the Field of Data Mining and Knowledge Discovery, International Journal of Information Technology and Decision Making, Vol. 7, Issue: 4, Page 639-682

Peter Stone, & Manuela Veloso (1997). Multiagent Systems: A Survey from a Machine Learning Perspective, URL: http://www.cs.cmu.edu/afs/cs/usr/pstone/public/papers/97MAS-survey/revised-survey.html

Popper K., (1980). *The Logic of Scientific Discovery*, Hutchinson, 1st impression 1959

Quinlan JR, (1979). Discovering rules by induction from large collections of examples, In: Michie D (ed), Expert systems in the micro electronic age. Edinburgh University Press, Edinburgh

Quinlan JR, (1993). C4.5: Programs for machine learning, Morgan Kaufmann Publishers, San Mateo

Sarabjot S. Anand, David A. Bell, John G. Hughes, (1996). EDM: A general framework for Data Mining based on Evidence Theory, Data & Knowledge Engineering 18 (1996) 189-223, *School of Information and Software Engineering, Faculty of lnformatics, University of Ulster (Jordanstown), UK*

Schwann, T., (1839). *Mikroskopische Untersuchungen uber die Ubereinstimmung in der Structur und dent Wachsthum der Thiere und Pflanzen*

Singh. Shivanshu K., Eranti. Vijay Kumer., & Fayad. M.E., (2010). Focus Group on Unified Data Mining Engine (UDME 2010): *Addressing Challenges, Focus Group Proposal*

Steinbach M, Karypis G, Kumar V, (2000). A comparison of document clustering techniques. In: Proceedings of the KDD Workshop on Text Mining

Two crows, (1999). Introduction to Data Mining and Knowledge Discovery, ISBN: 1-892095-02-5, Third Edition by Two Crows Corporation

Unified Theory or Theory of Everything (TOE), (2001). URL: http://searchcio-midmarket. techtarget.com/ definition/unified-field-theory

US Census Bureau, (2009). Iris, Diabetes, Vote and Breast datasets at URL: www.sgi.com/ tech/mlc/db

Wooldridge, M, Jennings, N. R., Kinny, D., (2000). "The Gaia Methodology for Agent-Oriented Analysis and Design", Kluwer Academic Publishers

Wu. Xindong, Kumar. Vipin, Quinlan., & J. Ross, et al, (2008). Top 10 algorithms in data mining, SURVEYPAPER, *Knowl Inf Syst* (2008) 14:1–37

Yang. Qlang., & Wu. Xindong, (2006). 10 CHALLENGING PROBLEMS IN DATA MINING RESEARCH, *International Journal of Information Technology & Decision Making,* Vol. 5, No. 4 (2006) 597–604

IntechOpen