Information Technology and Quantitative Management (ITQM2013)

A Unified Theoretical Framework for Data Mining

Dost Muhammad Khan\textsuperscript{a}\textsuperscript{*}, Nawaz Mohamudally\textsuperscript{b}, D K R Babajee\textsuperscript{c},

\textsuperscript{a}Department of Computer Science \& IT, The Islamia University of Bahawalpur, PAKISTAN
dostmuhammad\_khan@yahoo.com
\textsuperscript{b}School of Innovative Technologies \& Engineering (SITE), University of Technology, MAURITIUS
almohamudally\_utm.intnet.mu
\textsuperscript{c}Scientific \& Academic Research Council, African Network for Policy Research \& Advocacy for Sustainability, MAURITIUS,
dkrbabajee@gmail.com

Abstract

The pattern extraction and discovery of useful information from a dataset are the foremost purposes of data mining; the outcome of this process is the ‘knowledge’ which is helpful in taking the decision. For the past decade there have been multiple attempts and strong beliefs in the development and the formulation of the unified data mining frameworks that would answer to the fundamental versions related to the discovery of knowledge. In this paper we are presenting a novel unified framework for data mining conceptualized through the composite functions. The framework is further illustrated with a variety of real life datasets using different data mining algorithms.

© 2013 The Authors. Published by Elsevier B.V.
Selection and/or peer-review under responsibility of the organizers of the 2013 International Conference on Information Technology and Quantitative Management

Keywords: Composite Function; Unified Theory (UT); Unified Data Mining Theory (UDMT); Multiagent System (MAS); Data Mining Algorithms; Datasets; Unified Data Mining Framework

1. Introduction

Data mining is helpful to solve many challenges and problems in different fields with many success stories like Fraud Detection, Genomic Data, Text Mining and many more [31]. But still there are numerous challenges in the field of data mining research itself which should be addressed. These challenges are: Unified Data mining Processes, Scalability, Mining Unbalanced, Complex and Multiagent Data, Data mining in Distributed

\textsuperscript{*}Corresponding author. Tel.: +92-346-887-0217; fax: +92-62-9250232.
E-mail address: dostmuhammad\_khan@yahoo.com
and Network setting and Issues of Security, Privacy and Data Integrity in data mining as enumerated by Qiang Yang et al [42]. In this research paper, emphasis is land on the unified data mining processes.

Data mining is to discover hidden features and patterns from the large and complex datasets. According to Y.Y.Yao, the data mining is an intermediate system between a dataset and an application which transforms data into the useable knowledge [2]. Michael Berry classifies the data mining into two types: One type is the directed data mining and the other one is the undirected data mining. The directed data mining which is also called the supervised data mining explains and categorizes the data by using the target field of the dataset and the undirected data mining which is also called the unsupervised data mining finds the patterns without the target filed [17]. The knowledge can only be extracted through the integration of supervised and unsupervised data mining [15]. There is unanimity that data mining is not a single-step process and knowledge discovery is the result of successive processes. The scientific community attempts to develop and formulate a unified data mining theory that would answer to the fundamental questions related to the discovery of knowledge.

Another reason behind to formulate a unified theory is that the current data mining algorithms and techniques can solve the individual discrete data mining tasks such as clustering, classification, visualization, regression and association rule learning; called the single-step process. Since the available data mining tools and suites are based on the single-step process; therefore, tools are failed to produce ‘knowledge’. The researchers are now emphasizing that the discovery of ‘knowledge’ from a dataset is not a single-step process rather than it is a multi-step process. Although a unified data mining theory is not yet formulated, there is a little research being done to develop the unified data mining frameworks [14,43]. We propose a theoretical framework that unifies the clustering, classification and visualization, the data mining tasks and will quest the issues related to the discovery of knowledge. A composite functions approach is applied in the proposed framework and in order to test the concept we opt for agent programming, a multiagent system (MAS) precisely. In [18-20] it is demonstrated that a MAS approach is useful due to its scalability, robustness and simple programming. Furthermore, a MAS is helpful to speed up the performance and operation of the system. The proposed framework is not only validated mathematically but also with the illustrations on a variety of real life datasets with different data mining algorithms.

The rest of the paper is organized as follows. Based on existing unified data mining frameworks and data mining process models, section 2 proposes unified theoretical data mining framework is presented. Section 3 illustrates of this framework with results. Finally, the conclusion is drawn in section 4.

2. A Unified Theoretical Framework for Data Mining

According to Y.Y.Yao, data mining is relatively a new field and its own theories are not yet formulated [41]. The focus of the extensive research in the field of data mining is on the development and improvement of the existing algorithms and the evaluation of the discovered knowledge as a single-step process. A few attention is given to formulate data mining as a multi-step process that is a unified data mining. Now with the passage of time the field of data mining is being matured and the researchers have started to develop the unified data mining frameworks [32-41]. We discuss different unified data mining frameworks.

In [1-10,46] different unified data mining frameworks are presented and different approaches like relational algebraic, set theoretic, information theoretic and hybrid are applied in these frameworks. The data mining processes such as data partitioning, aggregation and data transformation are unified in some frameworks. In some frameworks, the missing data and variables and large data in a dataset are automatically handled and then different measures are taken to select the interesting pattern. Some frameworks unify different processes to parameterize a dataset based on intrinsic geometry. In [11] a multiagent system (MAS) theoretical framework for the knowledge extraction from a dataset is proposed where two approaches are used to extract the knowledge; one is by means of combination of clustering and classification and the second is through the combination of clustering and visualization data mining processes. The Boolean algebraic and the composition
of functions approaches are applied to validate the proposed theory. The oracle data mining (ODM) unifies clustering and classification [12] and the Self-Organizing Map (SOM), a tool developed by Teuvo Kohonen, is used for the visualization of high dimensional data; unifies the clustering and visualization [13]. In stead of separately performing the unification of data mining processes, the proposed framework in [11] unifies the both. In other words, the framework of [11] consists of the both ODM and SOM. In [14,15] the frameworks where different data mining algorithms are integrated, are presented which leads towards the knowledge extraction. But there is no theory behind these frameworks only the empirical results are demonstrated.

The most common data mining process models include CRISP-DM: Cross Industry Standard Process for Data Mining [27]; Six Sigma data mining process model [30]; The SEMMA data mining process model [30]; Scientific Data Mining Process Model [21]; Knowledge Discovery Process (KDP) Model [22-25]; Hybrid Data Mining Process Model [26]; Data Mining Process Model [29].

Based on the above, a unified theoretical framework for data mining is presented by formulating a unified data mining theory (UDMT) where the data mining processes; clustering, classification and visualization are unified by means of composition of functions. The proposed unified theoretical framework is based on the following assumptions which are also called the steps for knowledge extraction from a dataset:

**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.

The unified theory (UT) to support UDMT has set of requirements, comprehensiveness, preciseness, consistency, correctness and provides a solid foundation to attain the same conclusion and principle is called a unified theory 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 extracted through the composition of the functions. The composition of the functions is illustrated in Fig. 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. First, apply the clustering algorithms, then classification algorithms, followed by the visualization algorithms and finally the interpretation of the results of visualization. The selection of the appropriate and right result(s) will give the knowledge. Thus, data mining processes are unified through the composition of functions. The eq. (1) illustrates the composition of the functions.

\[(h \circ g \circ f): S \rightarrow C\]  

The order of the functions must be the same as given in eq. (1). In [16] the Mathematical Formulation of the Unified Data Mining Theory is discussed and the proposed theory is illustrated with different data mining algorithms and datasets and two cases are also discussed. Furthermore, a multi-step, multiagent system (MAS) tool called a Unified Data Mining Tool (UDMTool) is developed based on the proposed theory. The architecture, flowchart and framework of the tool are elaborated through different diagrams in [16]. In [31] the UDMTool is compared with MS SQL Server and ODM (single-step data mining tools). The comparison reveals that an automated and a multi-step data mining tool provide better facilities for knowledge extraction than the single-step data mining tools.

Moreover, the fitness of the given dataset is determined by using Akaike Information Criterion (AIC), a model selection criterion and the value of AIC is mapped with the logarithm value of complexities \( O \) of data mining algorithms. This approach leads to select the appropriate data mining algorithm(s). The formula to compute the value of AIC is given in eq. (2).

\[AIC = -2(\text{log(likelihood)}) + 2k\]  

Where \(2k\) and \(\text{log(likelihood)}\) are the number of parameters and the log of the \text{likelihood} of the model respectively. \(-2(\text{log(likelihood)})\) is also called the ‘Model Accuracy’ and \(2k\) is referred as ‘Number of Parameters’. Therefore, we can write the formula as given in eq. (3).

\[AIC = -\text{Model Accuracy} + \text{No. of Parameters}\]  

Now the question arises that is there any statistical linear relation, which is also called the linear model, between the AIC and the \( O \). The answer is ‘yes’. We try to investigate the linear relation of linear model between the AIC and the \( O \) of data mining algorithms. Fig. 2 depicts this linear model.

![Diagram](image-url)  

Fig. 2. The Linear Relation between the AIC and the \( O \)
3. Illustration and Results

To illustrate the proposed UDMT with different Data Mining Algorithms and datasets, let $S$ be a sample image of different colors.

1. $f(SOM) : S \rightarrow A$

The algorithm takes two inputs, the dataset $S$ and a metric of weights $W$, small standardized random values, used as initial values to start the algorithm. The algorithm has no targeted output. The function $f$ is illustrated in eq. (4).

$$f(S, W) = C_k(c_k)$$

Fig. 3 demonstrates the mapping of set $S$ and set $A$ using SOM algorithm in function $f$. The required parameters of this algorithm are a sample dataset $S$ and a metric of initial weights $W$. This is a 2:1 mapping because all the members of set $S$ are mapped with the single member of set $A$; therefore $f(SOM) : S \rightarrow A$ is onto or surjective. Hence $f$ is a function because it satisfies the conditions of the function; therefore SOM data mining algorithm is a function. In this algorithm a sample data and initial weights are inputted which creates a cluster on the basis of the best matching unit of neighbourhood as shown in an image in the set $A$ of Fig. 3.

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

The C4.5 algorithm takes the clusters of the dataset and classifies the data and produces the classifiers/rules of the cluster. The function $g$ is illustrated in eq. (5).

$$g(C_k) = R_k$$
Fig. 4. The mapping of function $g$ using Classification Algorithm (C4.5) for $S$ in illustration 3

Fig. 4 demonstrates the mapping of set $A$ and set $B$ using C4.5 algorithm in function $g$. The algorithm takes the set of clusters $C_1(c_1,c_2)$ of dataset $s_1$ and produces the further classified image $R_1(r_1,r_2)$. This is a 1:1 mapping because the members of the set $A$ are mapped with the corresponding members of set $B$; therefore $g(C4.5): A \rightarrow B$ is both one-to-one and onto or bijective. Hence $g$ is a function because it satisfies the conditions of the function; therefore C4.5 algorithm is a function.

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

The algorithm Data Visualization takes the classified data as input and produces 2D graphs as output for visualization. The function $h$ is illustrated in eq. (6).

$$h(R_k) = V_k \tag{6}$$

Fig. 5. The mapping of function $h$ using Data Visualization (2D Graph) for $S$ in illustration 3
Fig. 5 demonstrates the mapping of set \( B \) and set \( C \) using Data Visualization algorithm in function \( h \). The algorithm takes the classified image \( R_1(r_1,r_2) \) 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 a 1:1 mapping because the members of the set \( B \) are mapped with the corresponding members of the set \( C \); therefore \( h(\text{DataVisualization}) : B \rightarrow C \) is both one-to-one and onto or bijective. Hence \( h \) is a function because it satisfies the conditions of the function, therefore Data visualization is a function.

This illustration proves the knowledge can be extracted from the given dataset \( S \) through the proposed unified data mining theory using SOM, C4.5 and Data Visualization algorithms. The results is that 2D graph \( V_1(v_1,v_2) \) is ‘knowledge’ which can be interpreted as the white parts represent the similarity which forms a cluster and the black parts show the non similarity which forms a different cluster.

The proposed unified theoretical framework provides 2D/3D graphs as well as the decision rules of each cluster by unifying the data mining tasks such as clustering, classification and visualization which facilitate the user to extract knowledge. The proposed framework consists of five steps where data mining tasks are in a predefined order, clustering is its first data mining task followed by classification and visualization and these data mining tasks are also unified in such a way that the input of one step is the output of next step. Hence we can say all the limitations outlined from the existing frameworks and data mining process models are properly addressed in the proposed framework. Now we draw a comparison between the existing approaches with our proposed approach in Table 1.

Table 1. A Comparison of Existing and the Proposed Approach

<table>
<thead>
<tr>
<th>Criteria / Approach</th>
<th>Existing</th>
<th>Proposed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coverage of Data Mining Tasks</td>
<td>1. Clustering &amp; Visualization or 2. Clustering &amp; Classification. (Different names are used in different approaches)</td>
<td>Clustering, Classification &amp; Visualization.</td>
</tr>
<tr>
<td>Data Mining Algorithms</td>
<td>Data mining algorithms are implemented through different ways such as SQL etc.</td>
<td>Data mining algorithms are used for each of the data mining tasks using a MAS. But there is a condition that data mining algorithm must be a function.</td>
</tr>
<tr>
<td>Ease of Use</td>
<td>It is easy to use but it seems a difficult to implement data mining tasks in different ways. It is difficult to add new data mining tasks in these frameworks.</td>
<td>It is simple to use and easy to understand because the data mining tasks are carried out in straightforward way and covers all the possible data mining tasks. The user can add new data mining tasks.</td>
</tr>
<tr>
<td>Overview</td>
<td>Without properly applying the data mining algorithms and tasks it is difficult to extract knowledge but at the end knowledge can be extracted through the proposed approaches. Some of them provide knowledge in the form of clusters with characteristics of each cluster and some provides knowledge in the form of either decision rules or 2D graphs. The user has to select any of them or the both.</td>
<td>Provides knowledge directly in the form of 2D graphs which is extracted on the basis of results of clustering and classification data mining tasks and also facilitates the user by providing the information about clusters and decision rules of the dataset.</td>
</tr>
<tr>
<td>Conclusion</td>
<td>Although it is simple, easy and in some cases a lot of computation is required which can confuse the experienced user; but also the extraction of knowledge is not straightforward.</td>
<td>A single comprehensive framework which contains all the data mining tasks. It is simple, easy and the extraction of knowledge is straightforward.</td>
</tr>
</tbody>
</table>

In the existing unified frameworks different data mining tasks are unified through different approaches. It is not easy to add new data mining tasks in these frameworks. Some of the existing frameworks address only a specific problem such as in a framework where clustering and visualization is unified only focuses the graphical problem i.e. dimensionality, multi-resolution indexing scheme, bipartite graph, video etc. On the other hand, the proposed framework is a single comprehensive framework which unifies almost all data mining tasks through the composite functions, a novel approach. The proposed framework is flexible i.e. it is easy and simple to add new data mining task. The only limitation of the proposed framework is that data mining algorithm must be a function and there is no such condition in the existing approaches. We conclude that both
approaches are suitable for the extraction of knowledge from the dataset.

4. Conclusion

In this paper, a unified theoretical framework is presented for data mining by formulating a unified data mining theory (UDMT) where the data mining processes; clustering, classification and visualization are unified by means of composite functions. We first verify that the chosen data mining algorithms are functions and then utilize these algorithms as functions. The first function takes the set of data applies clustering algorithm (K-means) and produces the set of clusters. The second function takes the set of clusters applies the classification algorithm (C4.5) and produces the set of rules and finally the third function takes the set of rules applies data visualization algorithm and produces the set of 2D graphs. The functions are mapping from dataset (set of data) to the clusters (set of clusters) to the rules (set of rules) to the set of 2D graphs. From the set of 2D graphs one can interpret and evaluate the results to extract the knowledge. In this way the knowledge is extracted from the given dataset through the composition of the data mining processes; clustering, classification and visualization followed by interpretation and evaluation of the results. The “bijectivity” of each function is also determined. The classification and visualization functions are bijective and the clustering functions are surjective. The proposed theory is illustrated through variety of datasets like numeric, text and image with different arrangements and combinations of data mining algorithms. These illustrations demonstrate the proper working of the theory. The proposed theory is based on the assumption that the first data mining process is clustering followed by the classification and visualization. Therefore, the same order of the functions must be followed otherwise the theory will not work. We also compare the proposed approach with the exist approaches applied in different frameworks. The comparison reveals that the proposed approach is comprehensive and covers all the limitations of the existing frameworks. At this stage we conclude that the obtained results are satisfactory and consistent. However, in order to validate the proposed unified theory, further tests can be conducted.

Acknowledgement

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

References

