I.J.Modern Education and Computer Science, **2012, 2, 15-22**
Published Online March 2012 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijmecs.2012.02.03

Improve Teaching Method of Data Mining Course

Sakha'a Al Manaseer
Department of Information Systems, King Abdul Aziz University, Jeddah, Saudi Arabia
Email: selmanaseer@kau.edu.sa

Areej Malibari
Department of Information Systems, King Abdul Aziz University, Jeddah, Saudi Arabia
Email: aamalibari1@kau.edu.sa

***Abstract*— it is clearly perceived that most of theoretical information we teach to students are lost after graduation, mostly because abstract information do not last in students' minds much longer than the final exams, as they are not related to practical aspects and uses. Though labs are useful, they are not enough as they only offer exercises, but the course benefits will highly increase if students implement theories and exercises on a real life project. Data mining (DM) is one of the core courses which Information Systems students start to use after graduation once they start their career life. This paper focuses on improving the teaching ways of the course we shall implement, in order to attain a better understanding and comprehension by the students to make it more useful in their future real life careers, and it demonstrates the improvements in students' marks average after applying main concepts to real data**

***Index Terms*— Data Mining, DM Project, DM Lab.**

I. Introduction

All the time, huge amounts of data are collected and saved in large databases, whether on purpose or not. Essentially, data mining (DM) utilizes these large data warehouses to extract the information [1]. As mentioned earlier, Data mining is a core course in IS undergraduate program, which means that it is fatal to make sure that students really understand all DM basic concepts, and are able to apply the appropriate techniques in real applications using real data. In this paper we discuss the importance of changing the traditional teaching method in order to improve the students' skills and make sure that students know how Data Mining course can be utilized in actual areas.

Data mining students take this course as a set of chapters, with each chapter covering only one task of Data Mining with a set of definitions and techniques and where they apply. In my students' case I noted that they had a problem when applying tasks in real situations, this led me to conclude that the old teaching method is not sufficient anymore and we need to replace this method or enhance it with the one suggested in this paper. After finishing one chapter we directly applied the main concepts in that chapter on a real data "Project", and then we conducted the exam on students after they finished the project. This will show how this way improved students' work and comprehension.

This paper particularly guides the Data Mining lecturer to better ways of teaching the course in the form of "project", which should cover all Data Mining tasks. It familiarizes the students with DM concepts and applications and enables them to apply these concepts and applications on real data. This will provide students with the experience that can enhance their perception and

understanding of the course content. Moreover, the students' learning outcomes will be better tested and evaluated. Furthermore, the students' satisfaction about the course outcomes will start to increase when they know they can work with real data efficiently.

Before we start our work, we will highlight the importance and benefits of Data Mining as well as the main phases of Data Mining.

II. THE IMPORTANT AND BENEFITS OF DATA MINING

The main reason for the rapid growth of mining is mainly caused by the growing need to analyze large amount of data exist on internal documents [2]. Huge amount of organizations' information is stored in unstructured textual form, which illustrates the real need for automated extraction of useful knowledge from the textual data. Data mining applications have been broadly implemented to solve the problem in education, banking, marketing strategies, and production industries, which shows that data mining has large impact in the society [10]. According to, the use of data mining offers numerous benefits; some of these benefits are illustrated in TABLE I.

TABLE I.

DATA MINING BENEFITS [3]

Benefits	Description
Increased value of corporate Information	Deploying DM increases the value of corporate investments in existing information systems. It allows companies to more effectively gather and use the corporate knowledge buried in large collections of data, eliminating the need for costly reintegration or replacement of older data repositories
Lower integration costs	Many application data require a tremendous amount of consulting or integration before it can be deployed.
Increased productivity of knowledge workers	DM makes it easier for workers to find information within large corporate repositories of data; they allow users to navigate the information they need by following the key concepts that are important to them.
Improved Competitiveness	Data mining can facilitate better, faster prediction to application future.

According to table I, we let student's select real applications to really understand the huge benefits of the DM course earn better understanding of DM concepts and know how to apply data mining course in real work.

III. DATA MINNG PHASE

DM plays an essential role in the knowledge discovery process: in order to reach high levels of DM benefits that discover knowledge for each case; both Lecturer and student must follow the following sequence of Data mining phases [4].

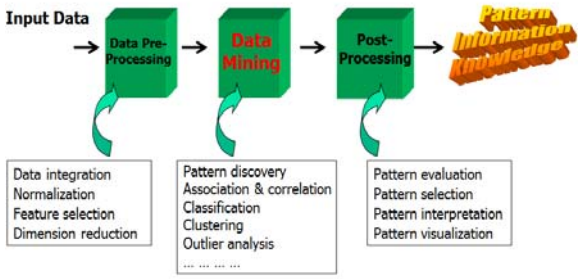

Figure 1. Data Mining phases [4]

Students at the DM class only go through theories and have just limited experience of real applications. In the first lecture students are introduced to Snapshot about the course, at the same time the students are divided into groups and asked to select an area of application and a final goal of DM. the students then discuss their selection with their lecturer to verify the objective, the task and approve the goals. table II shows the groups of students in the course (1 semester, 2011):

TABLE II.

APPLICATION OF DATA MINING [3]

Group Number	Application Area
Group1	Customer Relationship Management
Group2	Data Mining Application in Higher Education
Group3	Employment Placement and Job Market Online
Group4	Data Mining and Customer Relationships Management
Group5	Saher System

The above table shows the areas that each group of students selected. This paper shows how our new teaching method will improve students' skills by applying data Mining task to the above application. Students can choose any application (the above one is only an example to our students' task). So any student can improve their skill by choose any application.

In general, steps such as data selection, data cleaning, pre-processing, and data transformation should be applied on data just before performing the mining process.

A. *Data Preprocessing*

The lecturer explains data preprocessing techniques then breaks for one lecture to allow each group of students to apply all the major tasks discussed in lessons. The lecturer then publically goes through the students' groups and discusses their work in terms of [4]:

- Data cleaning: each group fills all missing values to his/her project, smooth noisy data, identify or remove outliers, and resolve inconsistencies
- Data integration: each group integrates all needed data for their project of multiple databases, data cubes, or files
- Data transformation: each group applies Normalization and aggregation techniques

- Data reduction: each group reduces representation in volume but produces the same or similar analytical results
- Data discretization: each group forms an ordinal categorical attribute as Static or dynamic depend on real data

Now each group should have its data ready to store, and at the same time the students surely understand how to manipulate main concepts to storing data; like: collect data, clean it, integrate it from multiple sources, represent data as mining results, and normalize it to be ready to be stored.

All previous concepts can be applied at any area because each group can share real data with other groups. After storing lessons is finished, the lecturer shall proceed to the second phase.

B. Data Mining

Now the lecturer should start with the second phase with groups and with same application area. Second phase is 'data mining' or (knowledge discovery from data). Data mining includes a set of phases.

B.a Clasification

Classification is another common task of data mining. The main task is to classify a collection of documents into a given set of categories. And when a new document is presented, the objective is to place this document in the appropriate category.

Systems that construct classifiers are among the most commonly used tools in data mining. Such systems take as input a collection of cases, each belonging to one of a small number of classes and described by its values for a fixed set of attributes, and output a classifier that can accurately predict the class to which a new case belongs [11]. Effective and scalable methods have been developed for classification phases such as Decision trees, naïve Bayesian classifications, support vector machines, neural networks, rule-based classifications, pattern-based classifications and logistic regressions. Students learn all these method and each group applies a different algorithm according to real data so that the student knows exactly how each algorithm works and the advantages and disadvantage of each algorithm.

Data classification is a two-step process. In the first step, a classifier is built describing a predetermined set of data classes or concepts. This is the learning step (or training phase). While the training data set is used to build the model, on the other hand testing data set is used to validate the model [7]. In the second step, The class labels of training data is unknown and a set of measurements is given with the aim of establishing the existence of classes or clusters in the data. See Figure 2.

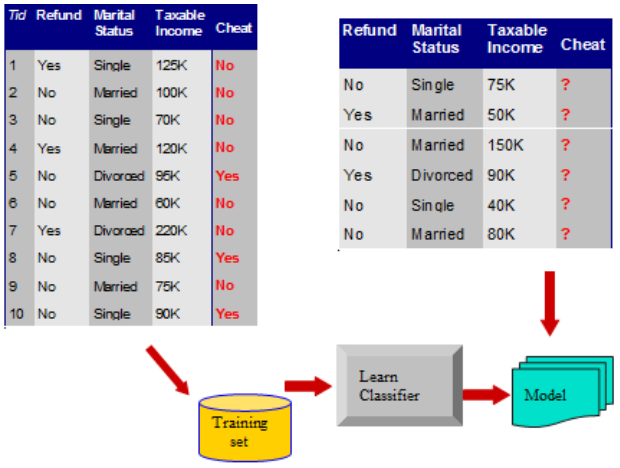

Figure 2. Classification [4]

Figure 2 describes classifier which memorizes the entire training data and performs classification only if the attributes of the test object match one of the training examples exactly. An obvious drawback of this approach is that many test records will not be classified because they do not exactly match any of the training records. A more sophisticated approach clustering algorithm finds a group of k objects in the training sets that are closest to the test object.

B.b Clustering

Cluster analysis is a field that comprises a wide range of data analysis fields. Clustering is a fully automatic process through which a collection of

documents are classified into groups. The documents within each group are more closely related to one another than documents assigned to different groups [5]. Figure 3 illustrates the overall task of cluster analysis.

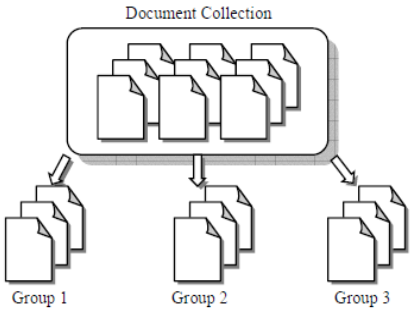

Figure 3. Clustering [5]

Clustering of unsupervised learning data means that the class labels of training data is unknown, so these data will not be classified using classification methods because they do not exactly match any of the training records but it is easy to classify them using clustering techniques. Clustering has a lot of algorithm; all of them share the same idea that is generated candidates. The most famous algorithms are partitioning algorithm, hierarchical algorithm and Density-based algorithm.

A partitioning can be used to divide the entire various partitions, and then evaluate them under some criterion e.g., minimizing the sum of square errors then each partition is mined separately, typical methods are k-means, k-medoids and CLARANS. Hierarchical algorithm create a hierarchical decomposition of the set of data (or objects) using some criterion e.g., distance matrix, typical methods are Diana, Agnes, BIRCH, ROCK and CAMELEON. Density-based approach can be used to create various partitions of the set of data (or objects). Based on connectivity and density, typical methods are DBSACN, OPTICS and Den Clue.

Clustering includes a lot of technics to produce high quality clusters, the quality of a clustering method is also measured by its ability to discover some or all of the hidden patterns and similarity of patterns ; good cluster with high intra-class similarity and low inter-class similarity.

Now each group goes to cluster its project data and collect it and apply all algorithms to find out the degree of similarity and dissimilarity to the objects related in other clusters. Clusters are differentiated by using similarities between data according to the characteristics found in the data and grouping similar data objects into clusters [6]. So students will apply more than one algorithm and the result is that the student will learn the exact meaning of clustering.

B.c Association

Discovery of interesting associations, generating rules and correlations between item sets in transactional and relational databases, this task is known as association rule mining [8]. A popular area of application is market-basket analysis, which studies the buying habits of customers by searching for sets of items that are frequently purchased together (or in sequence) see Figure 4. Association rule mining can be viewed as a two-step process [4].

1. Find all frequent items sets: items frequently gathered in a transaction data set with satisfying minimum support, such as buying first a PC, then a digital camera, and then a memory card, if it occurs frequently in a shopping history database.
2. Generate strong association rules from the frequent items sets: these rules must satisfy minimum support (How to mine such patterns and rules efficiently in large datasets?) and minimum confidence(How to use such patterns for classification, clustering, and other applications?).

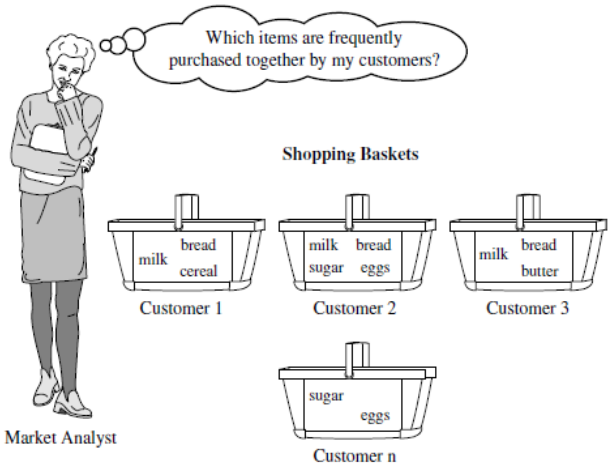

Figure 4. Association Example

For example, determine which items are frequently purchased together within the same transactions, see Figure 4. A typical association rule:

Milk → eggs [1/4, 1/3] (support, confidence)

A confidence, or certainty, of 0.333% means that if a customer buys milk, there is a 0.333% chance that she will buy eggs as well. A 25% support means that 25% of all of the transactions under analysis showed that milk and eggs were purchased together. Now each group find association rule to the real data and represent it as support and confidence.

C. Post-Procrssing

Lecturer should explain data post processing then let each group evaluate and visualize all real Data according to the type of application selected.

D. Survey To Find Soluation

A survey is a very important research tool that shows the benefits to problem solver of any type. Surveys can assist decision-makers in developing strategies to achieve the all-important goals. Results can play a key role in identifying areas of the related field that require corrective action and improvement. Keeping in view the importance of survey [9]. Results may directly affect the identification of pass or fail of new technology. The survey was conducted based on following questions and distributed to 35 students then results compiled as shown in the TABLE III.

Q1: Do you think u can apply all Data Mining task to any real actual? (Yes/No).

Q2:Do u believe that the combination between theoretical and practical sides during the semester is better to improve your skills? (Yes/No).

Q3: Do u believe that the new teaching way will help you in skipping your fear in applying it to career life? (Yes/No).

TABLE III.

RESULT OF SURVEY

Q#	Yes (%)	No (%)
1	88%	12%
2	79	21%
3	95	35%

Based on the improving teaching method of data mining course described in the survey, most of student believe that they can apply Data Mining in real application and they prefer learning theory with practical at same time to increase deep knowledge. A majority of them believe that this improvement in data mining course could be improve their skills.

IV .ANALYSIS AND CONCLUSION

I developed a mid-term exam to students and noted that the student results were good in general but they have weakness in numerical Questions. Figure 5 shows students' results.

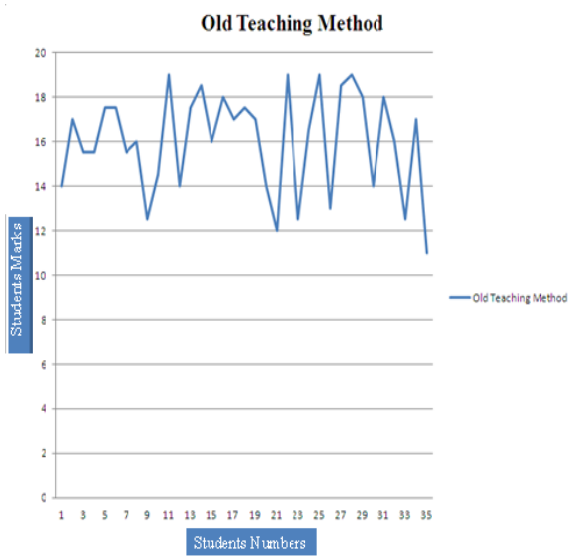

Figure 5. Student Result with Old Teaching Method

As a result to that I allowed my students select an application and apply Data Mining task with real data after each lecture. After applying the new method, I developed an exam to evaluate and collect students' results. The improvements were very clear. Figure 6 shows students' results.

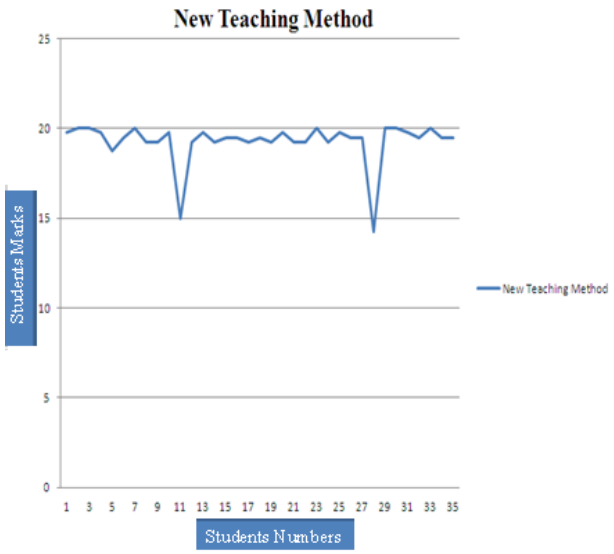

Figure 6. Student Result After Improving Teaching Method

We can obviously note that the suggested teaching method has improved the skills of students and made them ready to face career life and apply DM to any given application. I believe I achieved my goal to deliver the information and help my students to get the use of DM by this simple improvement. See Figure 7 that shows a simple comparison between the two methods.

Now students have the theoretical knowledge along with practical experience to work with. This course is very important for students after graduation to find a job in their field, so when we improve student skills in this course it would make them a better IS students and improve the chances for the graduates to compete more in the market. Benefits can be gained from improving teaching methods are many, for example if the student is ready to work with new project they would directly hit the market with minimum need for after graduation training, this is our goal.

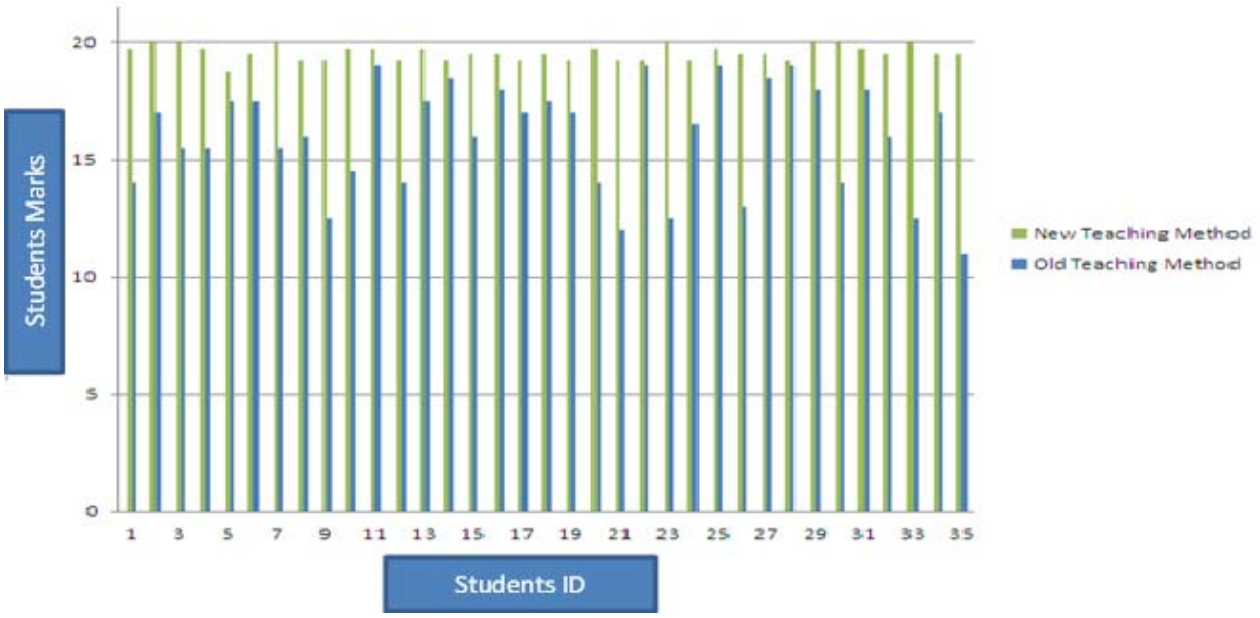

Figure 7. Comparison

REFERENCES

[1] http://www.laits.utexas.edu/~norman/BUS.FOR /course. Mat/Alex/.

[2] H. Karanikas and B. Theodoulidis, "Knowledge Discovery in Text and Text Mining Software," Department of Computation, UMIST, Manchester, UK, Technical Report Centre for Research in Information Management (CRIM) , 2002.

[3] F. Troulakis, "Text Mining," Department of Computation, UMIST, Manchester, Unpublished MSc thesis , 2000.

[4] Jiawei Han, Micheline Kamber and Jian Pei, *Data Mining, 2nd ed.,* Morgan Kaufmann, 2005.

[5] R. Feldman and J. Sanger, *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*, London: Cambridge University Press, 2007.

[6] Pang-Ning Tan, Michael Steinbach and Vipin Kumar, *Introduction to Data Mining*, Addison Wesley, 2005.

[7] G. Stefanson, "Business-to-Business Data Sharing: A source for integration of supply chains," *Int'l Journal of Production Economics*, 2002.

[8] Farrukh Saleem and Areej Malibari, "Data Mining Course in Information System Department– Case Study of King Abdul Aziz University," *3rd International Congress on Engineering Education (ICEED)*, December 2011, Malaysia.

[9] Dr. Ibrahim Abdulmohsin Albidewi and Abdul Rauf, "Improvement in Course Curriculum forSoftware Project management," *3rd International Congress on Engineering Education*, December, 2011, Kuala Lumpur Malaysia.

[10] http://www.sigkdd.org/curriculum.php, accessed date, 25th March, 2011.

[11] XindongWu, Vipin Kumar, J. Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey J. McLachlan, Angus Ng · Bing Liu, Philip S. Yu, Zhi-Hua Zhou, Michael Steinbach, David J. Hand and Dan Steinberg, "Top 10 algorithms in data mining," *Knowl Inf Syst* , 2007.

[12] Cabena, Hadjnian, Stadler, Verhees and Zanasi, *Discovering Data Mining from Concept to Implementation*, Prentice Hall, 1997.

I. Sakha'a El Manaseer is a lecturer in Information system Department, King Abdul Aziz University and IEEE KAU Student Branch Counselor. She has master degrees in Computer Science, Al-Balqa Applied University. In teaching, she has been focusing on improving teaching way in Data Mining course. In research, her current interests in Advanced Artificial Intelligence technology. include Genetic Algorithms (GA), Differential Evolution (DA). She is a member of IEEE.

D. Areej Malibari is an assistant professor at King Abdulaziz University. She is also the head of the information systems department (girls section), Faculty of Computing and information Technology. She completed her PhD in Computer Science from University of Essex in 2010, Master of Computer Science in 2003 from University of Essex. Her research interest is mainly in applications of agents in e-commerce. However, since 2011 she joined a research group with the main interest in data mining and fuzzy logic.

