12

Data Mining Applications: Promise and Challenges

Rukshan Athauda[1], Menik Tissera[2] and Chandrika Fernando[2]
[1]*The University of Newcastle*
[2]*Sri Lanka Institute of Information Technology*
[1]*Australia*
[2]*Sri Lanka*

1. Introduction

Data mining is an emerging field gaining acceptance in research and industry. This is evidenced by an increasing number of research publications, conferences, journals and industry initiatives focused in this field in the recent past.

Data mining aims to solve an intricate problem faced by a number of application domains today with the deluge of data that exists and is continually collected, typically, in large electronic databases. That is, to extract useful, meaningful knowledge from these vast data sets. Human analytical capabilities are limited, especially in its ability to analyse large and complex data sets. Data mining provides a number of tools and techniques that enables analysis of such data sets. Data mining incorporates techniques from a number of fields including statistics, machine learning, database management, artificial intelligence, pattern recognition, and data visualisation.

A number of definitions for data mining are presented in literature. Some of them are listed below:

- "Data mining is the process of discovering meaningful new correlations, patterns and trends by sifting through large amounts of data stored in repositories, using pattern recognition technologies as well as statistical and mathematical techniques" (Gartner Group, 1995).
- "Data mining is the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner" (Hand et al., 2001).
- "Data mining is an interdisciplinary field bringing together techniques from machine learning, pattern recognition, statistics, databases, and visualization to address the issue of information extraction from large data bases" (Cabena et al., 1998).
- "The extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data" (Han & Kamber, 2001).

We present application of data mining (also known as "Data Mining Applications") as an "experiment" carried out using data mining techniques that result in gaining useful knowledge and insights pertaining to the application domain. Figure 1 below depicts this process.

Source: Data Mining and Knowledge Discovery in Real Life Applications, Book edited by: Julio Ponce and Adem Karahoca,
ISBN 978-3-902613-53-0, pp. 438, February 2009, I-Tech, Vienna, Austria

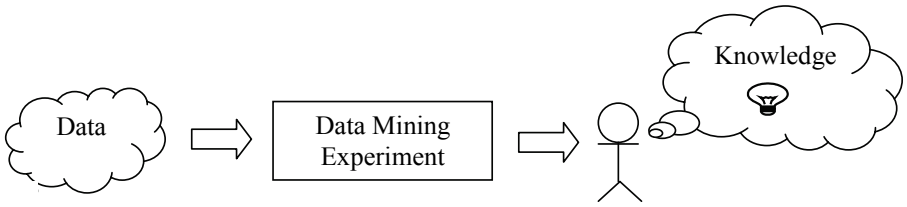

Fig. 1. A Data Mining Experiment

The analogy between conducting an experiment and applying data mining will become evident as we discuss further the issues and challenges faced in data mining applications.

Data mining (DM) application shows the promise to unlocking previously unknown, interesting knowledge from vast collections of data in many different application domains. DM has been successfully applied in a number of different domains (for e.g. astronomy (Weir et al., 1995), decision support (Wang & Weigend, 2004), business (Ester et al., 2002), IT security (Lee et al., 2001; Julisch & Dacier, 2002), medical and health research (Antonie et al.,2001; Au et al., 2005; Yu et al., 2004; Yan et al., 2004; Daxin et al.,2004; Liang & Kelemen ,2002; Chun et al., 2003), text mining (Hearst, 1999), marketing (Kitts et al., 2000), financial forecasting (Li & Kuo, 2008), education (Druzdzel & Glymour, 1994; Tissera et al., 2006; Ma et al., 2000), fraud detection (Senator et al., 1995; Rosset et al., 1999), and others (Fernando et al., 2008; Last et al., 2003; Fayyad et al., 1996c). Although a number of successful data mining applications exist, applying data mining is neither trivial nor straight-forward. Due to the inherent exploratory nature in data mining, there is a significant risk with little or no guarantee of successful results or Return on Investment (RoI) at the onset of a data mining initiative. Conducting a DM experiment is resource intensive and requires careful planning, critical analysis and extensive human judgement.

Today, a number of process models exists which aims to provide structure, control and standard methodology in applying data mining (Fayyad et al.,1996a; Cabena et al., 1998; Anand & Buchner, 1998; Chapman et al. 2000; Cios et al., 2000; Han & Kamber, 2001; Adriaans & Zantinge, 1996; Berry & Linoff, 1997; SAS Institute, 2003; Edelstein , 1998; Klösgen & Zytkow, 2002 ; Haglin et al. 2005). These process models are also known as Knowledge Discovery and Data Mining (KDDM) process models (Kurgan & Musilek, 2006). The KDDM process models outline the fundamental set of steps (typically executed iteratively with loopbacks) required in data mining applications covering the entire lifecycle in applying data mining from the initial goal determination to the final deployment and maintenance of the discovered knowledge.

Although KDDM process models provide a high-level structure to conduct a data mining experiment, following a KDDM process model by itself does not guarantee success. Applying data mining is an iterative process with extensive human intervention, whereby the DM team gains insights to the patterns and trends in data through the application of DM techniques with each iteration. This results in more-or-less a trial-and-error approach. The authors believe that a number of fundamental questions needs to be resolved in order to provide a more predictive, less risky approach with higher certainty of success in applying data mining. This, in turn, will enable fulfilling the promise of data mining resulting in a proliferation of data mining applications. The chapter aims to bring to light some of these fundamental questions.

Presently, research in data mining has mainly focused on different approaches for data analysis (techniques such as clustering, classification, associations, neural networks, genetic algorithms and others). A number of scalable algorithms enabling analysis of heavy volumes of data are proposed (Apriori (Agrawal et al., 1993); Auto Regression Trees (Meek et al., 2002); CURE (Guha et al.,1998); Two-Dimensional Optimized Association Rules (Fukuda et al.,1996); ML-T2LA (Han & Fu, 1995) and many others (Wu et al. 2008)). These contributions have provided a rich set of techniques for data analysis. However, in comparison, we observe only a few research work in data mining that discusses practical aspects pertaining to data mining application (Feelders et al., 2000; Karim, 2001). There is a dire need for research focusing these challenges and issues in data mining application.

In comparison to many other fields of study, data mining is evolving and still in its infancy. The analogy between data mining and the field of software development is presented to illustrate this evolution in perspective. Initially, research in software development focused on the development of programming languages and techniques. Later research focused on issues pertaining to large scale software development and methodologies with an entire discipline of software engineering in existence today. Similarly, the field of data mining is still in its infancy where presently more focus is on different techniques and data mining tasks in data analysis. A predictive, controlled approach to data mining application is yet to be realised.

This chapter aims to bring attention to some of the fundamental challenging questions faced in applying data mining with the hope that future research aims to resolve these issues. This chapter is organised as follows: Section 2 briefly discusses the KDDM process models and basic steps proposed for applying data mining. Section 3 discusses the fundamental questions faced during data mining application process. Section 4 provides a discussion and recommendations for conducting a data mining experiment. Section 5 concludes the paper.

2. Knowledge discovery and data mining process models

KDDM process models provide the contemporary guidelines in applying data mining. A number of KDDM process models have been discussed in literature (Fayyad et al.,1996a; Cabena et al., 1998; Chapman et al. 2000; Anand & Buchner, 1998; Cios et al., 2000; Han & Kamber, 2001; Adriaans & Zantinge, 1996; Berry & Linoff, 1997; SAS Institute, 2003; Edelstein , 1998; Klösgen & Zytkow, 2002 ; Haglin et al. 2005). The basic structure for KDDM process models was initially proposed by Fayyad et al. (Fayyad et al.,1996a, Fayyad et al.,1996b) (popularly known as the "KDD Process") with other models proposed later. A survey and comparison of prominent KDDM process models are presented in (Kurgan & Musilek, 2006).

The KDDM process models outline the fundamental set of steps (typically executed iteratively with loopbacks) required in data mining applications. KDDM process models span the entire lifecycle in applying data mining, from the initial goal determination to the final deployment of the discovered knowledge. It is noteworthy to point out that data mining is considered as a single step in the overall process of applying data mining in KDDM process. The different KDDM process models are similar except mainly for terminology, orientation towards research vs. industry contexts and emphasis on different steps. In (Kurgan & Musilek, 2006), the prominent KDDM process models (Fayyad et al.,1996a; Cabena et al., 1998; Chapman et al. 2000; Anand & Buchner, 1998; Cios et al., 2000)

are compared. Figure 2 illustrates a comparison table presented in (Kurgan & Musilek, 2006) mapping the different steps in the prominent process models. It is evident that all process models follow more-or-less a similar set of steps. Some of the similaritites outlined in (Kurgan & Musilek, 2006) are given below:

- *All process models follow a set of steps*: "All process models consist of multiple steps executed in a sequence, which often includes loops and iterations. Each subsequent step is initiated upon the successful completion of a previous step, and requires a result generated by the previous step as its inputs".
- *All process models cover entire lifecycle of a data mining experiment*: "Another common feature of the proposed models is the span of covered activities. It ranges from the task of understanding the project domain and data, through data preparation and analysis, to evaluation, understanding, and application of the generated results."
- *All process models are iterative in nature*: "All proposed models also emphasize the iterative nature of the model, in terms of many feedback loops and repetitions, which are triggered by a revision process."

Further discussion on KDDM proces models are presented in (Kurgan & Musilek, 2006).

Model	Fayyad *et al.*	Cabena *et al.*	Anand & Buchner	CRISP-DM	Cios *et al.*	Generic model
Area	Academic	Industrial	Academic	Industrial	Academic	N\A
No of steps	9	5	8	6	6	6
Refs	(Fayyad *et al.*, 1996d)	(Cabena *et al.*, 1998)	(Anand & Buchner, 1998)	(Shearer, 2000)	(Cios *et al.*, 2000)	N\A
Steps	1 Developing and Understanding of the Application Domain	1 Business Objectives Determination	1 Human Resource Identification 2 Problem Specification	1 Business Understanding	1 Understanding the Problem Domain	1 Application Domain Understanding
	2 Creating a Target Data Set	2 Data Preparation	3 Data Prospecting 4 Domain Knowledge Elicitation	2 Data Understanding	2 Understanding the Data	2 Data Understanding
	3 Data Cleaning and Preprocessing		5 Methodology Identification	3 Data Preparation	3 Preparation of the Data	3 Data Preparation and Identification of DM Technology
	4 Data Reduction and Projection		6 Data Preprocessing			
	5 Choosing the DM Task					
	6 Choosing the DM Algorithm					
	7 DM	3 DM	7 Pattern Discovery	4 Modeling	4 DM	4 DM
	8 Interpreting Mined Patterns	4 Domain Knowledge Elicitation	8 Knowledge Post-processing	5 Evaluation	5 Evaluation of the Discovered Knowledge	5 Evaluation
	9 Consolidating Discovered Knowledge	5 Assimilation of Knowledge		6 Deployment	6 Using the Discovered Knowledge	6 Knowledge Consolidation and Deployment

Fig. 2. Comparison of steps in prominent KDDM process models (Source: Kurgan & Musilek, 2006)

Due to the differences in terminology and the number of steps between different KDDM process models, we adopted the Generic Model presented in (Kurgan & Musilek, 2006) to describe the KDDM process:

1. *Application Domain Understanding*: In this step, the high-level goals of the data mining project (i.e. business/customer objectives) are determined and explicitly stated. The context in which the experiment is conducted is understood. The business/customer goals are mapped to data mining goals. The preliminary project plan is developed.
2. *Data Understanding*: This step pertains to identifying the relevant data sources and parameters required by the relevant data mining tasks. It includes elicitation of relevant domain knowledge about the data sets, selection and merging of data, identifying quality issues (completeness, redundancy, missing, erroneous data etc.). Exploring data

at this stage may also lead to hypothesis creation and determination of data mining tasks (Chapman et al., 2000).

3. *Data Preparation and Identification of DM Technology*: At this stage, all necessary tasks needed to perform data mining are finalised. Data mining techniques and algorithms are decided. The data sets are pre-processed for specific data mining tasks. Pre-processing includes selecting, cleaning, deriving, integrating and formatting data in order to apply specific data mining tasks.
4. *Data Mining*: In this step, the data mining tasks are applied to the prepared data set. At this stage, a DM model is developed to represent various characteristics of the underlying data set. A number of iterations in fine-tuning the model may take place with the involvement of data mining experts.
5. *Evaluation*: In this phase, the results of the generated models are visualised and evaluated. Both domain and data mining expertise is incorporated for visualisation and interpretation of results to find useful and interesting patterns. Based on the evaluation of results, the decisions to iterate any of the previous or current steps, or to conclude the project is made.
6. *Knowledge Consolidation and Deployment*: At this stage, the final report documenting the experiment and results are published. In addition, incorporation and formulation of how to exploit the discovered knowledge is taken into consideration. Deployment of discovered knowledge includes plan for implementation, monitoring and maintenance.

Further discussion and comparison of individual steps of the different KDDM process models are presented in (Kurgan & Musilek, 2006) (see Table 2. pp. 9-12 of Kurgan & Musilek, 2006).

Although the KDDM process models provide a structure and a set of high-level steps in applying data mining, there are a number of issues, challenges and decisions faced by the DM team during the project. Decisions taken by the DM team at each stage has a significant impact on the outcome of the experiment. Typically these issues, challenges and decisions are not covered in generic KDDM process models. Research focusing on "guidelines" and "best-practices" to assist in resolving some of these challenging questions would be of significant value – especially to reduce risks and uncertainty associated in conducting a data mining experiment. The authors beleive that successful strides in these areas will enable proliferation of data mining applications in many different domains. The next section highlights some of these issues and challenges faced during data mining applications.

3. Challenges in data mining application

Data mining by definition is exploratory in nature – that is, we are in search for previously unknown, hidden and interesting patterns in data. The fact that we are in search for unknown, hidden knowledge makes the outcome of data mining aplication difficult to predict at the onset of a DM project making it a risky and an uncertain endeavour. "Does interesting, relevant knowledge exist?", "What types of knowledge are we looking for?", "What method should we consider in order to find what we are looking for?", "How do we know whether we haven't missed any interesting 'knowledge' in the data set?" are some of the fundamental questions that pertain to data mining application. At present, these questions are answered based on the judgement of the DM team. To assist in these judgements, the iterative nature of KDDM process allows the DM team to try out certain data mining tasks, if failed, to re-tract and repeat until satisfactory results are achieved (or in

the worst case, resources are exhausted). This approach makes contemporary data mining application, a risky endeavour and typically follows a trial-and-error process.

In this section, we present some of the challenging questions faced by the DM team during DM application:

Application Domain Understanding: In this step, the overall goal for a data mining experiment is determined. The significance of the impact of determining goals on the outcome of the data mining experiment is highlighted in (Chapman et al., 2000) as follows: "A possible consequence of neglecting this step is to expend a great deal of effort producing the right answers to the wrong questions".

Although the objective of this step is clear, determing a goal for a data mining application may not necessarily be simple or straight-forward. A challenging question that is faced by the DM team is "How do you determine the 'correct'/ 'valid'/ 'best' objective or goal for a data mining project?"

The fact that data mining is exploratory in nature makes it difficult. In a typical engineering project, the objectives/goals of the project are straight-forward and easily determined. In a data mining experiment, due to its inherent exploratory nature, the outcomes are typically unknown. How to determine the suitable goals for a data mining experiment can be a challenging task.

We observe three main approaches taken in practice to determine goals for a DM initiative:

- *Using domain knowledge to determine goals for the data mining experiment*: The use of domain knowledge to determine suitable goals for data mining is popular. For instance in (Chapman et al., 2000), asking questions such as "What is the motivation of the project?" (see p35 in Chapman et al., 2000) refers directly to finding out what interests domain experts and therefore considered useful and interesting. This will assist in aligning the project goals with what the domain experts consider to be useful, interesting and relevant. Hence it reduces the risk and increases the possibility to mine knowledge that is useful, interesting and relevant. This is the approach typically considered first in a data mining experiment in an industry context.

This was the approach taken by the authors in (Fernando et al., 2008). A key goal of the DM intiative was to determine the correlation between production and export levels of tea in Sri Lanka.

- *Exploring the data to determine goals*: In this approach, it is more data driven, whereby exisiting data is examined at first (types of data, available data etc.) and understood. Initially, goals may be broad (i.e. "What kind of analysis is possible?"). Typically a preliminary analysis takes into consideration, domain knowledge, available data and data mining techniques, to determine a suitable goal(s). Careful analysis of preliminary results enables to find clues on the promising direction to be taken before further mining is performed. The goals are selected with the aim of increasing the possibility of discovering interesting knowledge.

  At the preliminary stage, if a particular path of analysis is not promising, different techniques may be considered. This may result in a more-or-less trial-and-error approach, however, without involving much effort as it is a preliminary analysis. The DM team gets a "look-and-feel" of the possibilities with data mining before focusing/selecting a particular direction/goal.

This approach was taken by authors in (Tissera et al., 2006). In this instance, the authors had access to a data set from an educational institute containing student information, including student performances of various courses in the undergraduate degree programmes. A

particular goal/aim did not exist initially. However, after examining the data, data mining techniques and tools, and considering the domain context, the authors determined that finding related courses in the undergraduate programme as a suitable goal/direction for the project based on student performance.

- *Blind search*: In this approach, certain data mining algorithms are executed on data sets with the hope of finding interesting "nuggets" of knowledge, without having an initial pre-conceived goal/direction. At the evaluation phase, the patterns returned by the different algorithms are considered for relevancy and usefulness. However, this approach is more suitable in a research context than an industry setting – where typically a project's resources/goals need strict justification prior to initiating. In (Chapman et al., 2000), this aspect is described as follows: "Blind search is not necessarily useless but a more directed search towards business objectives is preferable".

A hybrid of the above-mentioned approaches may be considered in determing a suitable goal for data mining.

Other challenges faced in data mining goal determination stage include:

- *How do we know the selected goals are "valid"?*: For instance, we may like to find "patterns" or "interesting" clusters of customers. For us to gain such knowledge, such patterns must exist. It is difficult to validate the existence of such patterns prior to embarking on a DM experiment. Otherwise, we may even try to find patterns that do not exist (i.e. patterns which are simply a product of random fluctuations, rather than representing any underlying structure (Hand, 1998)). This process is also known as "data dredging" or "fishing" in statistics (Hand, 1998). However, to complicate matters further, not discovering a pattern using a certain DM technique doesn't necessarily mean that such a pattern/correlation does not exist!
- *How do we ensure that "good" goals are not missed*: During the selection of goals, the data mining team focuses on selected goals. These goals were selected based on the DM team's judgement. It is possible that there exist hidden patterns and knowledge undiscovered in the vast data sets. How can we ensure that all relevant "knowledge" is discovered via "valid" goal selection?

Today's state-of-art mining methodologies, selects goals for a data mining application based on judgement of the data mining team considering domain knowledge, data available and data mining techniques. All KDDM process models emphasise the iterative nature of the process which a data mining application is conducted. Typically, goals are selected, an experiment is conducted, based on results at each stage, a step is revisited or moves to the next step. The iterative nature of KDDM process models allows retracting and considering different approaches/paths (goals, techniques and methods) in conducting a data mining experiment as a way to address this uncertainty. This approach, however is not optimal and results in a trial-and-error process which is resource-intensive and risky with no guarantee of favourable results. Approaches to minimise unsuccessful attempts and provide certain guarantees would be highly beneficial.

Answers to questions such as:

- How to select valid goals in a data mining project? What is a suitable goal?
- How to ensure that valid goals are not missed when conducting a data mining application? When do we stop looking? How do we know nothing interesting exists in the deluge of this data?

are still challenging open research problems.

Data Understanding: The major goal of this step is to understand the data sources, data parameters and quality of data. Data sets are selected for analysis in later steps of the KDDM process.

The major challenge at this stage includes:

- How to determine whether the relevant data sets and parameters are selected for the data mining tasks?

Since data mining considers patterns which are hidden or unknown, how do we ensure that the data parameters omitted in the data understanding and selection process does not result in ignoring parameters which have the possibility to affect the outcome of the DM experiment.

We observe that both domain expertise and technical expertise are required in order to determine the relevant data sets and also determine whether certain types of data are useful for mining. As in the previous stage, answers to these questions depend on the experience and judgement of the data mining team conducting the experiment. The iterative nature of conducting a data mining experiment enables to re-consider and re-do a step. However, as previously stated this leads to a trial-and-error approach taking considerable time and effort.

Another challenge is completeness: Whether successful or not in the initial data selection and data mining steps, how do we know we haven't missed any data parameters or considered some data source (maybe even external to the organisation) that, if considered, may have resulted in discovery of valuable knowledge?

Challenging questions such as

- How to determine the 'ideal' data set and parameters to satisfy a data mining goal? How do we ensure that all relevant data parameters and data set are considered?
- How do we ensure that data sets with the potential to discover 'useful' patterns are considered and not missed/omitted? What is the best way to identify which data parameters and data sets are relevant and will enable discovery of 'novel' nuggets of knowledge?

are still open research issues.

Data Preparation and Identification of DM Technology: This pertains to preparation of data sets to perform data mining tasks and finalising of the data mining techniques and tools.

Selection of data mining techniques, algorithms and tools is one of the crucial and significant questions that need to be decided by the DM team. There are a number of possible data mining techniques such as classification, clustering, association rule mining, machine learning, neural networks, regression and others. Also, a plethora of data mining algorithms exists. The DM team based on their judgement selects the data mining techniques and algorithms to apply. The question of

- What DM techniques and algorithms to apply that will result in useful and relevant knowledge?

is one of the fundamental open research questions in Data Mining Applications.

Preprocessing is a popularly used term to mean the preparation of data for data mining tasks. Preprocessing takes a significant effort, typically most of the effort in a data mining experiment (see Kurgan & Musilek, 2006 for a discussion on relative efforts spent of specific steps in the KDDM process). There are several reasons; the main reason being the fact that data being collected by enterprises and in different domains is not originally planned for analysis and therefore often contains missing, redundant, inconsistent, outdated and sometimes erroneous data.

As stated in (Edelstein, 1998), data quality is highly important in a data mining application. "GIGO (garbage in, garbage out) is very applicable to data mining. If you want good

models, you need good data." (Edelstein, 1998). The DM team makes a number of decisions at preprocessing stage which has a significant impact on the results of the DM tasks (for instance, sampling techniques, quantitative vs. qualitative data, omission/reduction of dimensions, data imbalance and others). We observe a number of research efforts addressing these issues in literature (imbalance data sets - e.g. (Nguyen et al., 2008); misssing values - e.g. (Farhangfar et al., 2007), and others).

Data Mining: In this step, data mining techniques are applied. The primary goals of data mining in practice tend to be prediction and description (Fayyad et al., 1996b). The DM techniques fit models to the data set. These models either are used to describe the characteristics of the data set (descriptive – for e.g., clustering), with the hope of discovering novel and useful patterns and trends in data; or used to predict future or unknown values of other variables of interest (predictive – for e.g., regression).

The DM team typically iterates through process to find a best-fit model for the data by adjusting various parameters of the model (e.g. threshold values). It is possible that the DM team may even modify the DM technique itself in order to determine a best-fit model. For instance in (Julisch & Dacier, 2002), the Attribute Oriented Induction technique is modified to gain favourable results in prediction. In order for the DM team to modify an existing or develop a new DM technique (such as SBA scoring function in (Ma et al., 2000)), requires an in-depth understanding of the DM technique.

A number of challenging issues may be faced by the DM team at this stage as outlined below:

- What are the optimal values in the model that would provide the 'best-fit' for the data set?
- Are there any other models that provide a better fit (in terms of accuracy, performance, etc.)?

At present, the DM team makes a judgement call when faced with these challenging questions, typically after applying DM techniques iteratively to the data set.

Evaluation: At the evaluation step of the data mining experiment, the results of the particular data mining tasks are visualised and interpreted. Although DM tasks reveal patterns and relationships, this by itself is not sufficient. Domain knowledge and data mining expertise is required to interpret, validate and identify interesting and significant patterns. The DM team incorporates domain expertise and data mining expertise in evaluating and visualising models in order to identify interesting patterns and trends.

In addition to evaluation of results, a number of significant decisions are considered at this stage with respect to the progress of the project:

- *Iterate?*: As discussed earlier, DM tasks may be iterated adjusting the model. Also, during evaluation, it is possible to discover a novel pattern or promising trend. Further investigation may be required for validation and verification purposes. This may be considered as a part of the existing project or as an entirely different DM initiative.

For instance, when conducting the DM experiment discussed in (Fernando et al., 2008), a spike in tea prices is observed in 1996. To determine the reasons causing this spike requires further investigation and can be considered as an entirely different DM project. Similarly, in (Tissera et al., 2006), when minining for correlated courses, the results revealed that the capstone project course did not demonstrate a strong relationship to any other course. The reasons for this fact can be investigated as a part of the exisiting DM project or fresh DM initiative.

- *Conclude project?*: The decision to conclude the project may be considered at this stage of KDDM process due to a number of reasons:
  - If satisfactory results are achieved in applying data mining, the decision to conclude the project can be considered at this stage. Note that lack of new patterns or correlations can itself be an insight to the non-existence of dependencies in parameters in the data set.
  - If unsatisfactory results are achieved, or due to the lack further resource commitment (personnel, funds etc.) for future analysis and minimal possibility for novel discoveries, the DM project may be concluded.

Knowledge Consolidation and Deployment: At this stage of the KDDM process, the results are published and the main stakeholders of the DM project are informed. Also, strategies to incorporate and exploit the discovered knowledge is considered. The implementation of the discovered knowledge and monitoring is also taken into consideration.

It is important to ensure that the support and "buy-in" of the project's stakeholders are maintained throughout the project. This aspect will assist in speedy acceptance and action on results of the project. A DM experiment conducted in isolation will require more convincing to build trust and acceptance of results prior to deployment.

4. Discussion

It is evident that a number of challenges and issues are faced during data mining applications. Some of them include:

- How do we determine goals for a DM application?
- How do we select the data that will result in achieving the goal?
- What type of DM technique(s) should be considered?
- How complete is the exploration? Did we miss any useful "nuggets"?

The responses to these challenging questions have a major impact on the direction taken and results obtained in a DM initiative. At present, there aren't definitive answers or processes to determine answers to these challenging questions. It is more-or-less a judgement made by the DM team.

Today, to address this uncertainty, data mining applications are conducted in an iterative manner. This process allows the DM team to develop a better understanding of the data, domain and data mining technqiues and gain insights with each iteration. The insights gained assists in decision making. The iterative nature of data mining application is reflected in the KDDM process models as well. However, a disadvantage of this approach is that the iterative nature results in a trial-and-error process to data mining applications, which is resource-intensive. Finding approaches that enable a more predictive, controlled, and risk-averse methodology to data mining applications is a challenge that remains to the DM research community. The authors believe that such methodologies require addressing the challenging questions presented in section 3.

At the beginning of the chapter, the authors described Data Mining Application as an "experiment" where the goal is discovery of knowledge from large data sets. The analogy between a research experiment and application of DM is evident. Similar to a scientific experiment, where researchers explore the unknown, data mining applications require the DM team to explore for unknown knowledge hidden in the vast data sets. This analogy assists us in determining conducive environments for conducting a DM experiment. Based

on the discussion thus far, we present some recommendations for conducting a data mining experiment.

- *Expertise required in a DM project team*: We believe that selecting the appropriate team with relevant expertise is crucial as human judgement plays a major role in a data mining application. The team must include domain expertise and a high-level of technical expertise. Domain knowedge will play a crucial role throughout the DM application especially in goal determination and evaluation of results. Also, technical expertise, especially relevant to data mining techniques and algorithms, is required. As demonstrated in many DM applications (for instance, Julisch & Dacier, 2002; Ma et al., 2000; Tissera et al., 2006), the DM models are tweaked, modified and even new techniques are developed to enable best-fit models that describe the data. This requires an extensive and in-depth understanding of the data mining models.
- *Managment stakeholders perspective of a DM project*: The authors believe that DM project's management stakeholders' understanding of the exploratory nature and uncertainties associated with a DM initiative is beneficial. This is especially true in an industrial context, where typical projects are deterministic. This fact will enable the project's managing stakeholders to effectively support/facilitate a DM project and manage expectations/outcomes of a DM initiative.

5. Conclusion

Data mining shows promise in enabling the discovery of hidden, previously unknown knowledge from vast data sets. Today, proliferation of databases technology has made it possible to collect and access such data sets in many different domains. Application of data mining to these large data sets has the possibility to unravel knowledge that can impact many different fields of study in society.

To achieve such proliferation in data mining applications, we believe a consistent, risk-averse and predictable methodology is required. Today, a number of Knowledge Discovery and Data Mining (KDDM) process models are proposed in literature with the aim of providing structure, control and standard methodology in applying data mining. KDDM process models outline the fundamental steps (executed iteratively) in applying data mining covering the entire lifecycle from goal determination to deployment of the discovered knowledge. Although KDDM process models provide a high-level structure to conduct a data mining application, following a KDDM process model by itself does not guarantee success. The exploratory nature of data mining makes DM application a risky, resource-intensive and an uncertain endeavour with no guarantee of success.

To address this issue, we believe some challenging questions need to be answered by the data mining research community. This chapter brings to attention some of the fundamental questions that need to be addressed for a more predictive, less-risky and controlled approach to data mining. We believe that significant strides in resolving these fundamental questions will enable a proliferation of data mining applications in many different application domains.

6. References

Adriaans, P. & Zantinge, D. (1996) *Data Mining*, Addison-Wesley.

Agrawal, R., Imieliski, T. and Swami, A. (1993), Mining association rules between sets of items in large databases, *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pp. 207-216, Washington, D.C., USA.

Anand, S. S. & Buchner, A. G. (1998) *Decision Support Using Data Mining*, Trans-Atlantic Publications.

Antonie, M.-L.; Za¨ıane, O. R. & Coman, A. (2001) Application of data mining techniques for medical image classification, *Second Intl. Workshop on Multimedia Data Mining (MDM/KDD'2001)*, pp. 94-101, San Francisco, USA.

Au, W.-H.; Chan, K. C. C.; Wong, A. K. C. & Wang, Y. (2005) Attribute Clustering for Grouping, Selection, and Classification of Gene Expression Data, *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 2, 2, 83 - 101.

Berry, M. J. A. & Linoff, G. (1997) *Data Mining Techniques: For Marketing, Sales, and Customer Support*: Wiley, 0471179809.

Cabena, P.; Hadjinian P.; Stadler R.; Verhees, J. & Zanasi, A. (1998) *Discovering data mining: From Concept to Implementation,* Prentice-Hall, Inc.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. & Wirth, R. (2000) CRISP-DM 1.0: Step by step data mining guide, CRISP-DM Consortium, pp. 79.

Chun, T., Aidong, Z. & Jian, P. (2003) Mining phenotypes and informative genes from gene expression data, *Proceedings of the ninth ACM SIGKDD international conference on Knowledge Discovery and Data Mining*, pp. 655-660, Washington, D.C.

Cios, K. J.; Teresinska, A., Konieczna, S.; Potocka, J. & Sharma, S. (2000), A knowledge discovery approach to diagnosing myocardial perfusion, *IEEE Engineering in Medicine and Biology Magazine*, 19, 4, 17-25.

Daxin, J.; Jian, P.; Murali, R.; Chun, T. & Aidong, Z. (2004) Mining coherent gene clusters from gene-sample-time microarray data, *Proceedings of the tenth ACM SIGKDD international conference on Knowledge Discovery and Data Mining*, pp.430-439, Seattle, WA, USA.

Druzdzel, M. & Glymour, C. (1994) Application of the TETRAD II program to the study of student retention in U.S. colleges, *Proceedings of the AAAI--94 Workshop on Knowledge Discovery in Databases (KDD-94)*, pp. 419-430, Seattle, WA.

Edelstein, H. (1998) Data Mining – Let's Get Practical, *DB2 Magazine*, 3, 2.

Ester, M.; Kriegel, H.-P. & Schubert, M. (2002), Web site mining: a new way to spot competitors, customers and suppliers in the world wide web, *Proceedings of the eighth ACM SIGKDD international conference on Knowledge Discovery and Data Mining*, pp. 249-258, Edmonton, Alberta, Canada.

Farhangfar, A.; Kurgan, L. A. & Pedrycz, W. (2007) A Novel Framework for Imputation of Missing Values in Databases, *IEEE Transactions on Systems, Man and Cybernetics, Part A*, 37, 5, 692-709.

Fayyad, U.; Piatetsky-Shapiro, G. & Smyth, P. (1996a), The KDD Process for Extracting Useful Knowledge from Volumes of Data, *Communications of the ACM*, 39,11, 27-34.

Fayyad, U. M.; Piatetsky-Shapiro G.& Smyth P. (1996b), From data mining to knowledge discovery in databases, *AI Magazine*, 17, 3, 37-54.

Fayyad, U.; Haussler D. & Stolorz P. (1996c), Mining scientific data, Communications of the ACM, 39, 11, 51-57.

Feelders, A.; Daniels, H. & Holsheimer, M. (2000), Methodological and practical aspects of data mining, Information & Management, 37, 271-281.

Fernando, H. C.; Tissera, W. M. R. & Athauda, R. I. (2008), Gaining Insights to the Tea Industry of Sri Lanka using Data Mining, *Proceedings of the International MultiConference of Engineers and Computer Scientists (IMECS 2008)*, pp. 462-467, Hong Kong.

Fukuda, T.; Morimoto, Y.; Morishita S. & Tokuyam, T. (1996), Data Mining Using Two-Dimensional Optimized Association Rules: Scheme, Algorithms, and Visualization, *Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data*, pp. 13-23, Montreal, Canada.

Gartner Group (1995), Gartner Group Advanced Technologies and Applications Research Note http://www.gartner.com

Guha, S.; Rastogi, R. & Shim, K. (1998), CURE: An Efficient Clustering Algorithm for Large Databases", *Proceedings of the 1998 ACM SIGMOD International Conference on Management of Data*, pp. 73-84. Seattle, WA, USA.

Haglin, D.; Roiger, R.; Hakkila, J. & Giblin, T. (2005), A tool for public analysis of scientific data, *Data Science Journal*, 4, 30, 39-53.

Han, J. & Fu, Y. (2005), Discovery of Multiple-Level Association Rules from Large Databases, *Proceedings of the 21st International Conference on Very Large Data Bases (VLDB 95),* pp. 420-431, Zurich, Swizerland.

Han, J. & Kamber, M. (2001), *Data Mining: Concepts and Techniques*, Morgan Kaufmann Publishers.

Hand, J. D.(1998), Data Mining: Statistics and More?, *The American Statistician*, 52, 2, 112-118.

Hand, D. J.; Mannila, H. & Smyth, P. (2001), *Principles of Data Mining*: MIT Press.

Hearst, M. A. (1999), Untangling text data mining, Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational LinguisticsAssociation for Computational Linguistics, pp. 3-10., College Park, Maryland, USA.

Julisch, K. & Dacier, M. (2002), Mining Intrusion Detection Alarms for Actionable Knowledge, *The 8th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD*), pp. 366-375, Edmonton, Alberta, Canada.

Karim, K. H. (2001), Exploring data mining implementation, *Communicaitons of the ACM*, 44, 7, 87-93.

Klösgen, W. & Zytkow, J.M. (2002). The knowledge discovery process, In : *Handbook of data mining and knowledge discovery*, Klösgen, W. & Zytkow, J.M. (Ed.), 10-21, Oxford University Press, 0-19-511831-6, New York.

Kurgan, L. A. & Musilek P. (2006), A Survey of Knowledge Discovery and Data Mining Process Models, *The Knowledge Engineering Review*, 21, 1 ,1-24.

Last, M.; Friedman, M. & Kandel, A. (2003), The data mining approach to automated software testing, *Proceedings of the ninth ACM SIGKDD international conference on Knowledge Discovery and Data Mining*, pp. 388-396, Washington, D.C., USA.

Lee, W.; Stolfo, S. J.; Chan, P. K.; Eskin, E.; Fan, W.; Miller, M.; Hershkop, S. & Zhang, J. (2001), Real time data mining-based intrusion detection, *DARPA Information Survivability Conference & Exposition II, 2001 (DISCEX '01)*, pp. 89-100. Anaheim, CA, USA.

Li, S.-T & Kuo, S.-C. (2008), Knowledge discovery in financial investment for forecasting and trading strategy through wavelet-based SOM networks, *Expert Systems with Applications*, 34, 935-951.

Liang, Y. & Kelemen, A. (2002), Mining heterogeneous gene expression data with time lagged recurrent neural networks, *Proceedings of the eighth ACM SIGKDD*

international conference on Knowledge Discovery and Data Mining, pp. 415-421, Edmonton, Alberta, Canada.

Kitts, B.; Freed, D. & Vrieze, M. (2000), Cross-sell: a fast promotion-tunable customer-item recommendation method based on conditionally independent probabilities, *Proceedings of the International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, pp. 437-446 Boston, Massachusetts, USA.

Nguyen, T. H.; Foitong, S.; Udomthanapong, S. & Pinngern, O. (2008) Effects of Distance between Classes and Training Datasets Size to the Performance of XCS: Case of Imbalance Datasets, *Proceedings of the International MultiConference of Engineers and Computer Scientists 2008 (IMECS 2008)* pp. 468-473, Hong Kong.

Ma, Y.; Liu, B.; Wong, C. K.; Yu, P. S. & Lee, S. M. (2000), Targeting the right students using data mining, *Proceedings of the sixth ACM SIGKDD international conference on Knowledge Discovery and Data Mining*, pp. 457-464, Boston, Massachusetts, USA.

Meek, C.; Chickering, D.M. & Heckerman, D. (2002), Autoregressive Tree Models for Time-Series Analysis, *Proceedings of the Second SIAM International Conference on Data Mining*, pp. 229-244, Arlington, VA, USA.

Rosset, S.; Murad, U.; Neumann, E.; Idan, Y. & Pinkas, G. (1999), Discovery of fraud rules for telecommunications - challenges and solutions, *Proceedings of the International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, pp. 409-413, San Diego, California, USA.

SAS Institute (2003), *Data mining Using SAS Enterprise Miner*, SAS Publishing, 1590471903.

Senator, T. E.; Goldberg, H. G.; Wooton, J.; Cottini, M. A.; Khan, A. F. U.; Klinger; C. D., Llamas; W. M.; Marrone M. P. & Wong, R. W. H. (1995), The Financial Crimes Enforcement Network AI System (FAIS): Identifying Potential Money Laundering from Reports of Large Cash Transactions, *AI Magazine*, 16, 21-39.

Tissera, W. M. R.; Athauda, R. I. & Fernando, H. C. (2006), Discovery of Strongly Related Subjects in the Undergraduate Syllabi using Data Mining, *International Conference on Information and Automation (ICIA 2006)*, pp. 57-62, Colombo, Sri Lanka.

Wang H. & Weigend, A. S. (2004), Data mining for financial decision making, *Decision Support Systems*, 37, 4, 457-460.

Weir N.; Fayyad, U. M. & Djorgovski, S. (1995), Automated star/galaxy classification for digitized POSS-II, *Astronomical Journal*, 109, 2401-2412.

Wu X. ;Kumar V., Quinlan J. R., Yang J. G. Q., Motoda H., McLachlan G. J., Ng A., Liu B., Yu P. S., Zhou Z.-H., Steinbach M., Hand D. J. & Steinberg D. (2008), Top 10 algorithms in data mining, *Knowledge and Information Systems*, 14, 1, 1-37.

Yan, L.; Verbel D. & Olivier, S. (2004), Predicting prostate cancer recurrence via maximizing the concordance index, *Proceedings of the tenth ACM SIGKDD international conference on Knowledge Discovery and Data Mining*, pp. 479-485, Seattle, WA, USA.

Yu, L. T. H.; Chung, F-l; Chan, S. C. F. & Yuen, S.M.C. (2004), Using emerging pattern based projected clustering and gene expression data for cancer detection, *Proceedings of the Second Conference on Asia-Pacific Bioinformatics*, Vol. 29, pp. 75-84, Dunedin, New Zealand.