Oct 27th, 3:00 PM - 4:30 PM

A Critical Overview of Data Mining for Business Applications

George Telonis
University of Patras, qtelonis1@otenet.gr

Peter P. Groumpos
University of Patras, groumpos@ece.upatras.gr

A Critical Overview of Data Mining for Business Applications

George Telonis

[1]University of Patras,,
26500 Rion, Greece
{gtelonis@otenet.gr, gtelonis@ferrycenter.gr}

Abstract. Over the past 2-3 decades there has been a huge increase in the amount of data being stored in databases as well as the number of database applications in business and the scientific domain. This explosion in the amount of electronically stored data was accelerated by the success of the relational model for storing data and the development and maturing of data retrieval and manipulation technologies. While technology for storing the data developed fast to keep up with the demand, little stress was paid to developing software for analysing the data until recently when companies realized that hidden within these masses of data was a resource that was being ignored. The huge amounts of stored data contains knowledge on a good number of aspects of their business waiting to be harnessed and used for more effective business decision support. Data mining methods seem very appropriate to extract this useful information. A good number of them are presented and briefly analyzed. Possible applications in utilizing these techniques are outlined. An overview of both the data mining techniques and potential application to solve many challenging problems of the society is been carefully analyzed and presented. Very interested future research directions are briefly presented.

Keywords: Data mining, Big Data, Management Systems, Business Applications

Introduction

Everybody looks to a world that does not remain the same. No one can deny that the world is changing, and changing very fast. Technology, education, science, environment, health, communication, entertainment, eating habits, - there is hardly anything in life that is not changing. It is no more possible to live in the way we have been living so far. It seems that now the entire fabric of life will have to be changed. It must be planned in an entire different new way.

The rapid proliferation of the Internet and related technologies has created an unprecedented opportunity for enterprises to collect massive amounts of data regarding customers and all aspects of their business operations. Yet the reality is that most organizations today are:

1) "data rich" but "information and knowledge poor" and
2) not harnessing the full potential of their data, which is perhaps the second most important asset after human capital.

Internet based applications such as social media, webs, its usage tracking and online reviews as well as more traditional technology applications like RFID, Supply Chain
Management (SCM), Enterprise Resource Planning (ERP) and Customer Relationship Management (CRM) provide access to vast amounts of data regarding customers, suppliers, competitors as well as a firm's own activities and business processes.

Being able to unlock the insights and knowledge trapped in such raw data constitutes a key lever and for competitive advantage in hypercompetitive business. Therefore the challenging problem is what knowledge the raw data contain and how can be extracted from them. There are a number of methods such as learning techniques. [1-2].The last few years Data Mining methods have become an attractive technique for extracting new knowledge. [3-4]. Data Mining can be considered to be an inter-disciplinary field involving concepts from Machine Learning, Database Technology, Statistics, Deep Learning, Mathematics and Cognition, among others. Data Mining is often accused of being a new buzz world for Database Management System (DBMS) reports.
Data Mining in various forms is becoming a major component of business operations.
Almost every business process today involves some form of data mining. Customer Relationship Management, Supply Chain Optimization, Demand Forecasting, Assortment Optimization, Business Intelligence, and Knowledge Management are just some examples of business functions that have been impacted by data mining techniques. [5-7]
In this paper an extensive critical overview of Data Mining and its different aspects are analyzed and presented with an emphasis in Business applications.

Why Data Mining from the Big Data Driven World is a Scientific Challenge to Business

In today's digital world, we are surrounded with big data that is forecasted to grow 50%/year into the next decade. Furthermore over the past 2-3 decades there has been a huge increase in the amount of data being stored in databases as well as the number of database applications in business and the scientific domain. This explosion in the amount of electronically stored data was accelerated by the success of the relational model for storing data and the development and maturing of data retrieval and manipulation technologies. While technology for storing the data developed fast to keep up with the demand, little stress was paid to developing software for analyzing the data until recently when companies realized that hidden within these masses of data was a resource that was being ignored. . The ironic fact is, we are drowning in data but starving for knowledge. Why? All this data creates noise which is difficult to mine – in essence we have generated tons of amorphous data, but experiencing failing big data initiatives and especially creating new knowledge. No one denies that valuable and useful knowledge is deeply buried inside the big data driven world (BDDW). If we do not have powerful tools or techniques to mine such data, it is impossible to gain any benefits from such data.
The huge amounts of stored data contains knowledge on a good number of aspects of their business waiting to be harnessed and used for more effective business decision support. Database Management Systems (DMS) used to manage these data sets at present only allow the user to access information explicitly present in the databases i.e. the data. The data stored in the database is only a small part of the 'iceberg of information' available from it. Contained implicitly within this data is knowledge about a number of aspects of their business waiting to be harnessed and used for more effective business decision support. This extraction of knowledge from large data sets is called Data Mining or Knowledge Discovery in Databases and is defined as the non-trivial extraction of implicit, previously unknown and potentially useful information from data. Almost in parallel with the developments in the database field, machine learning research was maturing with the development of a number of sophisticated techniques based on different models of human learning. Learning by example, cased-based reasoning, learning by observation and neural networks are some of the most popular learning techniques that were being used to create the ultimate thinking machine.

While the main concern of database technologists was to find efficient ways of storing, retrieving and manipulating data, the main concern of the machine learning community was to develop techniques for extracting knowledge from data. It soon became clear that what was required for Data Mining was a formal connection between technologies developed in the big database and machine learning communities.

Data Mining is an important analytic process designed to explore data. Much like the real-life process of mining diamonds or gold from the earth, the most important task in data mining is to extract non-trivial nuggets from large amounts of data. This is very important for the business community. Extracting important knowledge from a mass of data can be crucial, sometimes essential, for the next phase in the analysis: the modeling. Many assumptions and hypotheses will be drawn from your models, so it's incredibly important to spend appropriate time "massaging" the data, extracting important information before moving forward with the modeling. Today most "Business" lack mathematical models that can rely to take important decisions for the well been of the business.

Data Mining can be considered to be an inter-disciplinary field involving concepts from Machine Learning, Database Technology, Statistics, Deep Learning, Mathematics, Cognition, Clustering and Visualization among others. Data Mining is often accused of being a new buzz world for Database Management System (DBMS) reports. This is not true. Using a DBMS Report a company could generate reports such as:

- Last month's sales for each service type
- Sales per service grouped by customer sex or age bracket
- List of customers who lapsed their insurance policy

However, using Data Mining techniques the following questions may be answered

- What characteristics do my customers that lapse their policies have in common and how do they differ from my customers who renew their policy?
- Which of my motor insurance policy holders would be potential customers for my house content Insurance policy?

Clearly, Data Mining provides added value to DBMS reports and answers questions that DBMS reports cannot answer. Thus let us address the issue of data mining methods and techniques.

Data Mining Methods and Techniques: A Brief Overview

Although the definition of Data Mining (DM) seems to be clear and straightforward, we may be surprised to discover that many people mistakenly relate to Data Mining different algorithms, techniques and/or tasks. Among them are generating histograms, issuing Structural Query Language (SQL) queries to a database, a data control language (DCL), a data manipulation language (DML), a declarative language (4GL), and visualizing and generating multidimensional shapes of a relational table.

For example: DM is not about extracting a group of people from a specific city in our database; the task of DM in this case will be to find groups of people with similar hobbies or preferences in our data. Similarly, DM is not about creating a graph of, say, the number of people that have lung cancer against smoking heavily —data mining's task in this case could be something like: is the chance of getting lung cancer higher if you smoke heavily?

The main tasks of DM are twofold:

A) **Create predictive power (CPP)**, using features to predict unknown or future values of the same or other feature and

B) **Create a descriptive power (CDP)**, find interesting, human-interpretable patterns that describe the data.

There are a number of different data mining techniques. Most of them are of the A) type.

Regardless the type that they are representing, it is important to understand well they depend on the different business problem and provides a different insight. Knowing the type of business problem that we are trying to solve, will determine the type of data mining technique that will yield the best results.

From the many and different techniques for this paper eight (8) techniques have been chosen as been considered the most common been used. Short description of each one is provided next.

1. **Regression analysis.**
   Regression models are the mainstay of predictive analytics. In statistical terms, a regression analysis is the process of identifying and analyzing the relationship among variables. It can help you understand the characteristic value of the dependent variable changes, if any one of the independent variables is varied. This means one variable is dependent on another, but it is not vice versa. It is generally used for prediction and forecasting. The linear regression model analyzes the relationship between the response or dependent variable and a set of independent or predictor variables. That relationship is expressed as an equation that predicts the response variable as a linear function of the parameters. It is important to emphasize that this process provides information ONLY of the LINEAR dependence between two variables.
2. **Clustering analysis**.
   Cluster analysis, or clustering, is a way to categorize a collection of "objects," such as survey respondents, into groups or clusters to look for patterns. The cluster is actually a collection of data objects; those objects are similar within the same cluster. That means the objects are similar to one another within the same group and they are rather different or they are dissimilar or unrelated to the objects in other groups or in other clusters. There are various ways to cluster. Regardless of the method, the purpose is generally the same: to use cluster analysis to partition into a group of segments and target markets to better understand and predict the behaviors and preferences of the segments. Clustering is a valuable predictive-analytics approach when it comes to product positioning, new-product development, usage habits, product requirements, and selecting test markets. A result of this analysis can be used to create customer profiling.
3. **Classification analysis**
   Data mining techniques classification is the most commonly used data mining technique which contains a set of pre classified samples to create a model which can classify the large set of data. This technique helps in deriving important information about data and metadata (data about data). This analysis is used to retrieve important and relevant information about data, and metadata. It is used to classify different data in different classes. Classification is similar to clustering in a way that it also segments data records into different segments called classes. But unlike clustering, here the data analysts would have the knowledge of different classes or cluster in advance. So, in classification analysis you would apply algorithms to decide how new data should be classified. A classic example of classification analysis would be our Outlook email. In Outlook, they use certain algorithms to characterize an email as legitimate or spam.
   There are two main important processes involved in this technique:

A) **Learning** – In this process the data are analyzed by different algorithms and
B) **Classification** – In this process the data is used to measure the precision of the classification rules

4. **Association Rule Learning (ARL).**
   It refers to the method that can help you identify some interesting relations (dependency modeling) between different variables in large databases. This technique can help you unpack some hidden patterns in the data that can be used to identify variables within the

data and the concurrence of different variables that appear very frequently in the dataset. Association rules are useful for examining and forecasting customer behavior. It is highly recommended in the retail industry analysis. This technique is used to determine shopping basket data analysis, product clustering, catalog design and store layout. In IT, programmers use association rules to build programs capable of machine learning.

5. **Rule Induction**.

It refers to the method that can help you identify some interesting relations (dependency modeling) between different variables in large databases. This technique can help you unpack some hidden patterns in the data that can be used to identify variables within the data and the concurrence of different variables that appear very frequently in the dataset. Association rules are useful for examining and forecasting customer behavior. It is highly recommended in the retail industry analysis. This technique is used to determine shopping basket data analysis, product clustering, catalog design and store layout. In IT, programmers use association rules to build programs capable of machine learning.

6. **Memory-based reasoning (MBR)/Case-based reasoning.**

This technique has results similar to a neural network's but goes about it differently. MBR looks for "neighbor" kind of data rather than patterns. It solves new problems based on the solutions of similar past problems. MBR is an empirical classification method and operates by comparing new unclassified records with known examples and patterns.

7. **Decision Trees.**

Decision trees use real data-mining algorithms to help with classification. A decision-tree process will generate the rules followed in a process.Decision trees are useful for helping you choose among several courses of action and enable you to explore the possible outcomes for various options in order to assess the risk and rewards for each potential course of action.Such an analysis is useful when you need to choose among different strategies or investment opportunities, and especially when you have limited resources.

8. **Anomaly Analysis or Single out-Detection**

This refers to the observation for data items in a dataset that do not match an expected pattern or an expected behavior. Anomalies are also known as outliers, novelties, noise, deviations and exceptions. Often they provide critical and actionable information. An anomaly is an item that deviates considerably from the common average within a dataset or a combination of data. These types of items are statistically aloof as compared to the rest of the data and hence, it indicates that something out of the ordinary has happened and requires additional attention. This technique can be used in a variety of domains, such as intrusion detection, system health monitoring, fraud detection, fault detection, event detection in sensor networks, and detecting eco-system disturbances. Analysts often remove the anomalous data from the dataset top discover results with an increased accuracy. This technique finds application mostly in health systems and big supply chain products.

All of these techniques can help analyze different data from different business perspectives. Now we have the knowledge to decide the best technique to summarize data into useful information – information that can be used to solve a variety of business problems to increase revenue, customer satisfaction, or decrease unwanted cost.

Data Mining for Business Applications

Data mining (DM) is already incorporated into the business processes in many sectors. This Technology is well established in applications such a targeted marketing, customer fault detection and market basket analysis (see below). It is also emerging as an important new

technology in a wide range of new applications areas, such as social networks, games, sensor and smart network, social media, intelligent e-mails and TripAdvisor.

Data Mining is primarily used today by companies with a strong consumer focus — retail, financial, communication, and marketing organizations, to "drill down" into their transactional data and determine pricing, customer preferences and product positioning, impact on sales, customer satisfaction and corporate profits. With data mining, a retailer can use point-of-sale records of customer purchases to develop products and promotions to appeal to specific customer segments.

It is very interesting to mention a few applications of DM for Business. Some of them are:

1. Market Basket Analysis
   Market basket analysis is a modelling technique based upon a theory that if you buy a certain group of items you are more likely to buy another group of items. This technique may allow the retailer to understand the purchase behavior of a buyer.
2. Financial Banking
   With computerised banking everywhere huge amount of data is supposed to be generated with new transactions. DM can contribute to solving business problems in banking and finance by finding patterns, and correlations in business information and market prices that are not immediately apparent to managers because the volume data is too large.
3. Health and Medical Systems
   Data mining holds great potential to improve health and Medical systems. It uses data and analytics to identify best practices that improve care and reduce costs. Researchers use data mining approaches like multi-dimensional databases, machine learning, soft computing, data visualization and statistics. Mining can be used to predict the volume of patients in every category. Processes are developed that make sure that the patients receive appropriate care at the right place and at the right time. Data mining can also help healthcare insurers to detect fraud and abuse.
4. Mining Geology and Mineral Resource Estimation and Consulting Services
   One area that has a big data base is the area of Geology. Data Mining can find a wide range of applications. The four pillars of Mineral Resource estimation are data quantity and quality, geological understanding, grade analysis, and adherence to international reporting codes and standards (e.g. JORC; CIM NI 43-101; SAMREC; SME; PERC). All four of these aspects are equally important in the public and responsible declaration of a Mineral Resource Estimation. We build and maintain long lasting relationships with our clients and industry partners. Our estimates have assisted hundreds of projects from early stage grass roots exploration through to production. Collection, management, and analysis of mining and exploration of mass data is a must. In .addition, geological models need to be developed, that can capture thousands of geological changes for millions of years.
5. Tourism and Culture
   Aggregating the largest collection of travel search and booking data sources from around the world is an area that Data Mining can find immediate applications. Simplifying the view of the complex behaviors, intentions and preferences that influence travel and specific visits to certain archeological sites is excellent for using DM techniques . While traditional business intelligence (BI) products benchmark historic performance and future booked revenue, they don't provide a real-time view of who is searching and booking travel online. Those solutions ignore the different ways that consumers make their travel decisions. Using DM and Artificial Intelligence the Tourism and culture business can benefit a lot.

Many other applications can be named in which DM techniques can be used. These are but not limited: Education Data Mining (EDM), Customer Relationship Management (CRM), Financial Banking, Fraud Detection, Law enforcement, market research through Customer Segmentation,

Corporate Surveillance, Bio- Informatics and many more. There are no any scientific field that is not collecting data and everyday are available for further use of them.

REMARK: **It must be stressed that in most of the above and not only DM techniques and Artificial Intelligence usage is on an embryonic stage. There is long way to go before the benefits of using these new methods of DM and AI can become apparent.**

Summary and Future Research

Big data caused an explosion in the use of more extensive Data Mining (DM) techniques, partially because the size of the information is much larger and because the information tends to be more varied and extensive in its very nature and content. In many cases these data are also fuzzy. Therefore data mining (DM) is more than running some complex queries on the data you stored in your database. You must work with your data, reformat it, or restructure it, regardless of whether you are using SQL, document-based databases such as Hadoop, or simple flat files. Identifying the format of the information that you need is based upon the technique and the analysis that you want to do. After you have the information in the format you need, you can apply the different techniques (individually or together) regardless of the required underlying data structure or data set.

With large data sets, it is no longer enough to get relatively simple and straightforward statistics out of the system. With 40 or 50 million records of detailed customer information, knowing that five million of them live in one location is not enough. We want to know whether those five million are a particular age group and their average earnings so that we can target our customer needs better.

These business-driven needs changed simple data retrieval and statistics into more complex Data Mining. The business problem drives an examination of the data that helps to build a model to describe the information that ultimately leads to the creation of the resulting and useful report.

Creating new knowledge from the Big Data Driven World which is formed from the different business communities is a scientific challenge. Data Mining (DM) principles have been around for many years, but, with the advent of **big data**, it is even more important. Many different techniques have been developed the last two decades. It is important to keep in mind that each business application will use a specific DM technique. One technique that might work perfectly for one application might not even be useful to another on. In this paper eight (8) such techniques have been presented and briefly analyzed. A number of applications that DM have and/or will been used were presented. However there is still a lot of work to be done.

Therefore future research in Data Mining (DM) theory and methodologies is an open field. More vigorous mathematical modelling of this fast moving scientific area is needed. Theories from Artificial Intelligence (AI), Deep learning (DL) and Fuzzy Cognition should be used searching for new methods to solve many everyday problems. Generic solutions that could be used to more than one application are highly desired. New learning algorithms are needed. The role of experts should be taken into serious considerations. Since many data are fuzzy and contain misleading information fuzzy logic and cognitive methods need to be investigated and how they can be useful to business applications. Certain business sectors that are very active and depend heavily on personal behavior need to be investigated separately and with a much greater attention. Specific application driven software tools are needs to be developed.

References

1. Sarder, R., Effective Learning Methods: How to develop the most effective learning method, paperback, March 3, 2011.
2. Prince, M., J., and Felder R., M., "Inductive Teaching and Learning Methods: Definitions, Comparisons and Research Bases", Journal of Engineering Education, pp.123-138 April 2006.
3. Hand, D. J., Mannila, H., & Smyth, P., Principles of data mining. MIT press, 2001.
4. Han J, Kamber M and Pei J, Data Mining: Concepts and Technique. 3rd Edition, Morgan Kaufman,2011.
5. Witten I. H., Frank E., Hall M.A., J. Pal C., J., Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufman, 2016.
6. Kantardzic M., Data Mining: Concepts, Models, Methods, and Algorithms. 2nd Edition, Jonh Wiley and Sons. Inc. 2011.
7. Berry M., J., A., and Linoff G., F., Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management, John Wiley & Sons INC, 2004.