Analysis of Data Mining Techniques

Dr. Sheel Ghule, Manoj Dhobale

Abstract –

As we experienced a revolution in information availability and usage via the internet, most of the businesses and organizations began to collect data related to their own operations. While the database technologists have been seeking efficient means of storing, retrieving and manipulating data, the machine learning community has focused on developing techniques for learning and acquiring knowledge from the data.

One of the most important and difficult tasks in the whole Knowledge Discovery in Databases (KDD) process is to choose the right data mining technique, as the commercial software tools provide more and more possibilities together and the decision requires more and more expertise on the methodological point of view. Indeed, there are a lot of data mining techniques available for an environmental scientist wishing to discover some model from her/his data. This diversity can cause some troubles to the scientist who often have not a clear idea of what are the available methods, and moreover, use to have doubts about the most suitable method to be applied to solve a concrete domain problem. In this work, a classification of most common data mining methods is presented in a conceptual map which makes easier the selection process. It is oriented to provide model/algorithm selection support, suggesting the user the most suitable data mining techniques for a given problem.

Keywords: Knowledge Discovery from Databases, Data Mining, Association, Clustering, Prediction

Introduction

The development of information technology has generated large amount of databases and huge data in various areas. The research in databases and information technology has given rise to an approach to store and manipulate this precious data for further decision making. Data mining is a process of extraction of useful information and patterns from huge data. It is also called as knowledge discovery process, knowledge mining from data, knowledge extraction or data /pattern analysis.

In its simplest form, data mining automates the detection of relevant patterns in a database, using defined approaches and algorithms to look into current and historical data that can then be analyzed to predict future trends. Because data mining tools predict future trends and behaviors by reading through databases for hidden patterns, they allow organizations to make proactive, knowledge-driven decisions and answer questions that were previously too time-consuming to resolve.

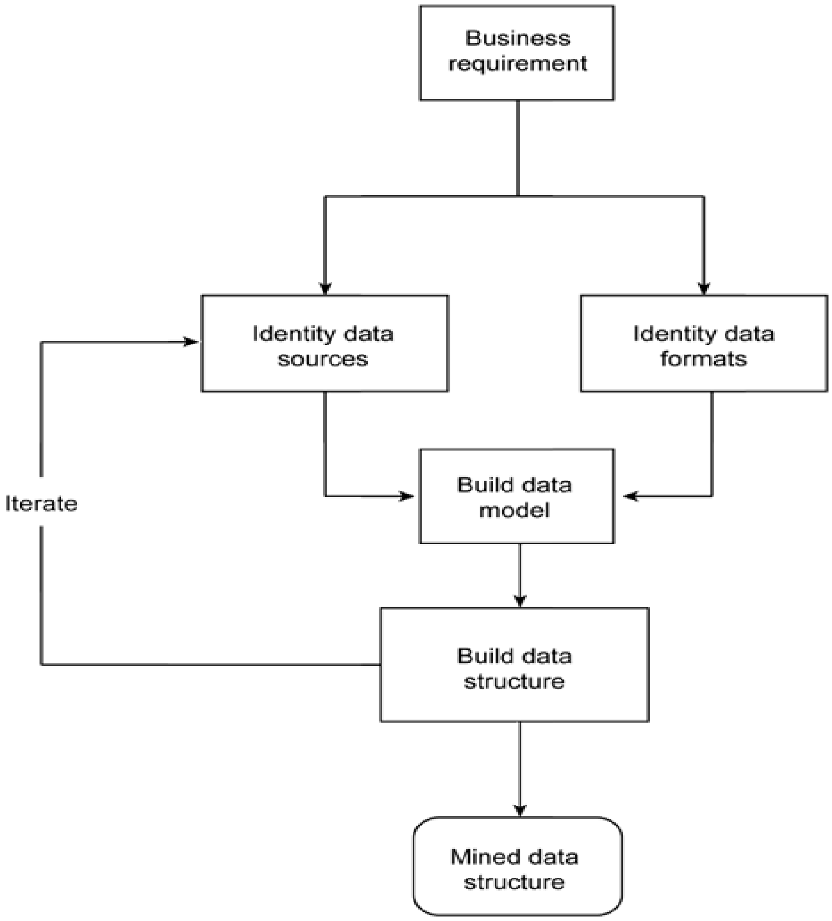
Data mining as a process

Fundamentally, data mining is about processing data and identifying patterns and trends in that information so that you can decide or judge. Data mining principles have been around for many years, but, with the advent of *big data*, it is even more prevalent.

Big data caused an explosion in the use of more extensive data mining 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. With large data sets, it is no longer enough to get relatively simple and straightforward statistics out of the system. With 30 or 40 million records of detailed customer information, knowing that two million of them live in one location is not enough. You want to know whether those two million are a particular age group and their average earnings so that you can target your 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 report. Figure outlines the process.

Figure: Outline of the process



The process of data analysis, discovery, and model-building is often iterative as you target and identify the different information that you can extract. You must also understand how to relate, map, associate, and cluster it with other data to produce the result. Identifying the source data and formats, and then mapping that information to our given result can change after you discover different elements and aspects of the data.

DATA MINING TECHNIQUES

Several core techniques that are used in data mining describe the type of mining and data recovery operation. Unfortunately, the different companies and solutions do not always share terms, which can add to the confusion and apparent complexity.

Let's look at some key techniques and examples of how to use different tools to build the data mining.

Association

Association is one of the best known data mining technique. In association, a pattern is discovered based on a relationship of a particular item on other items in the same transaction. For example, the association technique is used in market basket analysis to identify what products that customers frequently purchase together. Based on this data businesses can have corresponding marketing campaign to sell more products to make more profit.

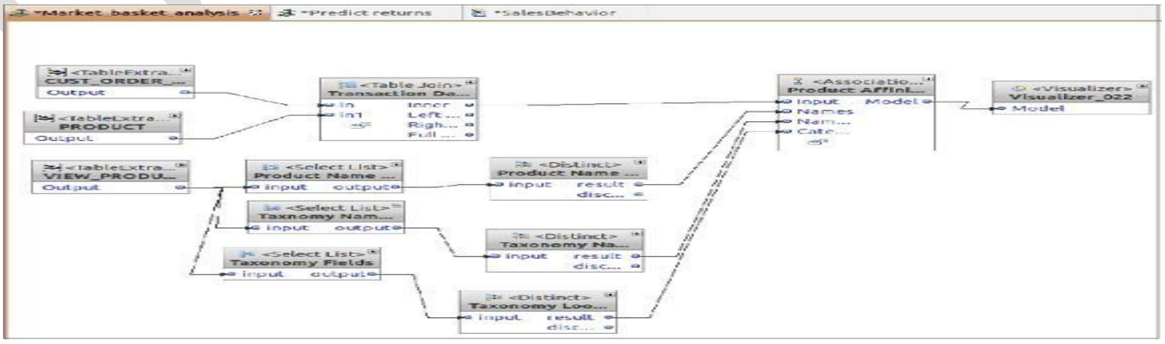
Applications: market basket data analysis, cross-marketing, catalog design, loss-leader analysis, etc.

Types of association rules: Different types of association rules based on

- Types of values handled
  - Boolean association rules
  - Quantitative association rules
- Levels of abstraction involved
  - Single-level association rules
  - Multilevel association rules
- Dimensions of data involved
  - Single-dimensional association rules
  - Multidimensional association rules

Building association or relation-based data mining tools can be achieved simply with different tools. For example, within InfoSphere Warehouse a wizard provides configurations of an information flow that is used in association by examining your database input source, decision basis, and output information. Figure shows an example from the sample database.

Figure: Information flow that is used in association



Classification

You can use classification to build up an idea of the type of customer, item, or object by describing multiple attributes to identify a particular class. For example, you can easily classify cars into different types (sedan, 4x4, convertible) by identifying different attributes (number of seats, car shape, driven wheels). Given a new car, you might apply it into a particular class by comparing the attributes with our known definition. You can apply the same principles to customers, for example by classifying them by age and social group.

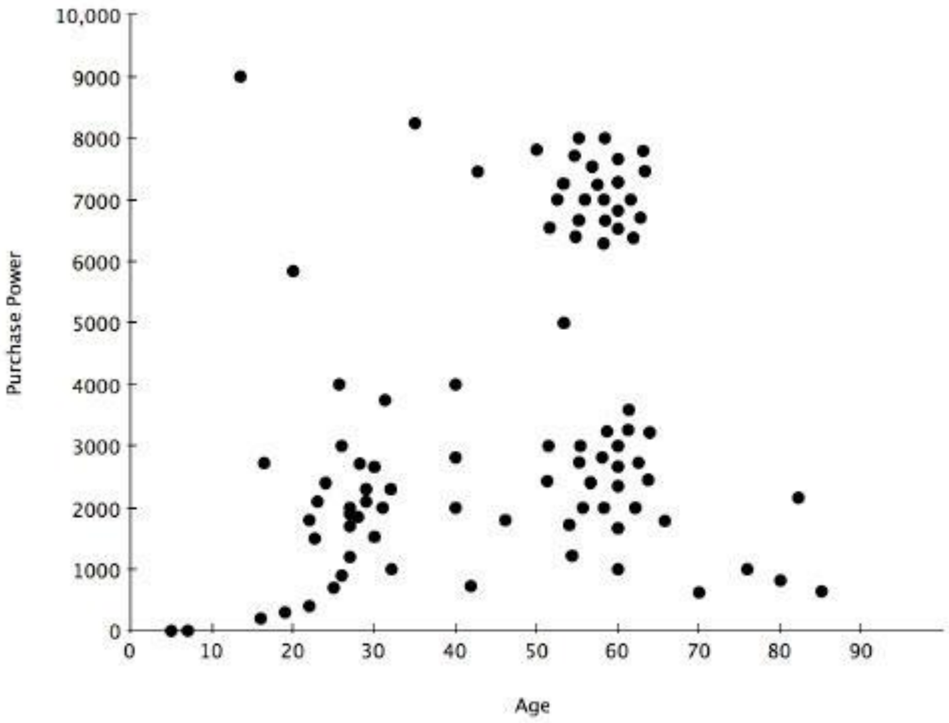
Additionally, you can use classification as a feeder to, or the result of, other techniques. For example, you can use decision trees to determine a classification. Clustering allows you to use common attributes in different classifications to identify clusters.

Clustering

By examining one or more attributes or classes, you can group individual pieces of data together to form a structure opinion. At a simple level, clustering is using one or more attributes as your basis for identifying a cluster of correlating results. Clustering is useful to identify different information because it correlates with other examples so you can see where the similarities and ranges agree.

Clustering can work both ways. You can assume that there is a cluster at certain point and then use our identification criteria to see if you are correct. The graph in Figure shows a good example. In this example, a sample of sales data compares the age of the customer to the size of the sale. It is not unreasonable to expect that people in their twenties (before marriage and kids), fifties, and sixties (when the children have left home), have more disposable income.

Figure: Clustering



In the example, we can identify two clusters, one around the US$2,000/20-30 age group, and another at the US$7,000-8,000/50-65 age group. In this case, we've both hypothesized and proved our hypothesis with a simple graph that we can create using any suitable graphing software for a quick manual view. More complex determinations require a full analytical package, especially if you want to automatically base decisions on *nearest neighbor* information.

Plotting clustering in this way is a simplified example of so called *nearest neighbor* identity. You can identify individual customers by their literal proximity to each other on the graph. It's highly likely that customers in the same cluster also share other attributes and you can use that expectation to help drive, classify, and otherwise analyze other people from your data set.

You can also apply clustering from the opposite perspective; given certain input attributes, you can identify different artifacts. For example, a recent study of 4-digit PIN numbers found clusters between the digits in ranges 1-12 and 1-31 for the first and second pairs. By plotting these pairs, you can identify and determine clusters to relate to dates (birthdays, anniversaries).

Prediction

Prediction is a wide topic and runs from predicting the failure of components or machinery, to identifying fraud and even the prediction of company profits. Used in combination with the other data mining techniques, prediction involves analyzing trends, classification, pattern matching, and relation. By analyzing past events or instances, you can make a prediction about an event.

Using the credit card authorization, for example, you might combine decision tree analysis of individual past transactions with classification and historical pattern matches to identify whether a transaction is fraudulent. Making a match between the purchase of flights to the US and transactions in the US, it is likely that the transaction is valid.

Sequential patterns

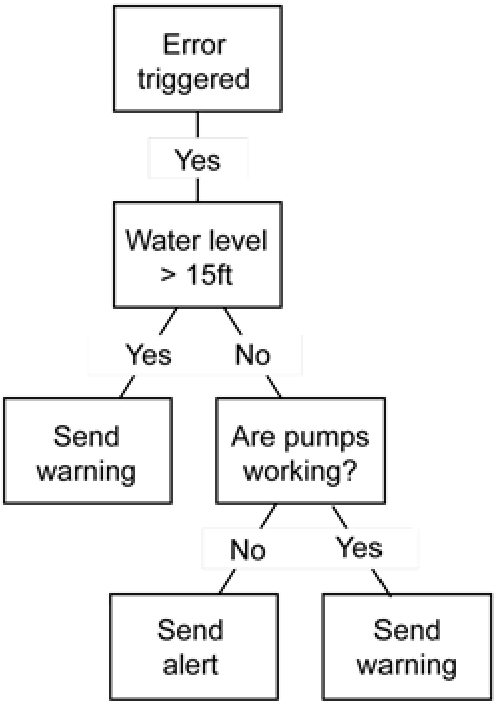
Often used over longer-term data, sequential patterns are a useful method for identifying trends, or regular occurrences of similar events. For example, with customer data you can identify that customers buy a particular collection of products together at different times of the year. In a shopping basket application, you can use this information to automatically suggest that certain items be added to a basket based on their frequency and past purchasing history.

Decision trees

Related to most of the other techniques (primarily classification and prediction), the decision tree can be used either as a part of the selection criteria, or to support the use and selection of specific data within the overall structure. Within the decision tree, you start with a simple question that has two (or sometimes more) answers. Each answer leads to a further question to help classify or identify the data so that it can be categorized, or so that a prediction can be made based on each answer.

Figure shows an example where you can classify an incoming error condition.

Figure: Decision tree



Decision trees are often used with classification systems to attribute type information, and with predictive systems, where different predictions might be based on past historical experience that helps drive the structure of the decision tree and the output.

Combinations

In practice, it's very rare that you would use one of these exclusively. Classification and clustering are similar techniques. By using clustering to identify nearest neighbors, you can further refine your classifications. Often, we use decision trees to help build and identify classifications that we can track for a longer period to identify sequences and patterns.

Conclusion

Data mining is a “decision support” process in which we search for patterns of information in data. In other words, Data mining has importance regarding finding the patterns, forecasting, discovery of knowledge etc in different business domains. Data mining techniques such as classification, clustering, prediction, association and sequential patterns etc. it helps in finding the patterns to decide upon the future trends in businesses to grow.

Using data mining to understand and extrapolate data and information can reduce the chances of fraud, improve audit reactions to potential business changes, and ensure that risks are managed in a more timely and proactive fashion.

References

1. http://dataminingtools.net/wiki/introduction_to_data_mining.php

2. Aamodt A. and E. Plaza, "Case-Based Reasoning: Foundational Issues, Methodological Variations, and System Approaches," AI Communications, vol. 7, 1994, pp. 39-59.

3. http://www.slideshare.net/huongcokho/data-mining-concepts

4. Fayyad U, et al 1996: From Data Mining to Knowledge Discovery: An overview. In Advances in Knowledge Discovery and Data Mining. AAAI/MIT Press.

5. Gibert K, A. García-Rudolph, G. Rodríguez-Silva 2008: The role of KDD Support-Interpretation tools in the conceptualization of medical profiles: An application to neurorehabilitation. Acta Informatica Medica 16(4) 178-182

6. Spate J, K. Gibert, M. Sànchez-Marrè, E. Frank, J. Comas, I. Athanasiadis, R. Letcher 2006. Data Mining as a tool for environmental scientist. In procs 1srts iEMSs Workshop DM-TES 2006 Third Biennal Meeting: "Summit on Environmental Modelling and Software". Burlington, VE USA.

7. Vazirigiannis M, M. Halkidi, D. Gunopulos 2003: Uncertainty handling and quality assessment in data mining. Springer-Verlag.

8. https://iaonline.theiia.org/data-mining-101-tools-and-techniques

9. http://www.dataminingtechniques.net