DATA MINING

A Simple Guide for the Beginners!

This paper introduces the subject of data mining in simple lucid language and moves on to build more complex concepts. Start here if you are a beginner.

Author: Akash Mitra
Dated: 21 Mar 2012
Source: http://www.dwbiconcepts.com

Data Mining. I have an allergy to this term.

Not because I hate the subject of data mining itself, but because this term is so much over-used and misused and exploited and commercialized and often conveyed in inaccurate manner, in inappropriate places and often with intentional vagueness.

So when I decided to write about what is data mining, I was convinced that I need to write about what is NOT data mining first, in order to build a formal definition of data mining.

What is Data Mining? (And what it is not)

Here is the Wikipedia definition of data mining:
"Data mining … is the process of discovering new patterns from large data sets"

Now the question is: what does the above definition really mean and how does it differ from finding information from databases? We often store information in databases (as in data warehouses) and retrieve the information from the database when we need it. Is that data mining?
Answer is 'no'. We will soon see why is it so.

Let's start with the big picture first. Data mining is basically one of the steps in the process of knowledge discovery in database (KDD). Knowledge discovery process is basically divided in 5 steps:

- Selection
- Pre-processing

- Transformation
- Data Mining
- Evaluation

“Selection” is the step where we identify the data, “pre-processing” is where we cleanse and profile the data, “transformation” step is required for data preparation, and then is data mining. Lastly we use “Evaluation” to test the result of the data mining.

Notice here the term – “Knowledge” as in Knowledge Discovery in Database (KDD). Why did you say “Knowledge”? Why not “information” or “data”?

This is because there are differences among the terms “data”, “information” and “knowledge”. Let’s understand this difference through one example.

> You run a local departmental store and you log all the details of your customers in the store database. You know the names of your customers and what items they buy each day.
>
> For example, Alex, Jessica and Paul visit your store every Sunday and buys candle. You store this information in your store database. This is data.
> Any time you want to know who are the visitors that buy candle, you can query your database and get the answer. This is information. You want to know how many candles are sold on each day of week from your store, you can again query your database and you’d get the answer – that’s also information.
>
> But suppose there are 1000 other customers who also buy candle from you on every Sunday (mostly – with some percentage of variations) and all of them are Christian by religion. So, you can conclude that Alex, Jessica and Paul must be also Christian.

Now the religion of Alex, Jessica and Paul were not given to you as data. This could not be retrieved from the database as information. But you learnt this piece of information indirectly. This is the ”knowledge” that you discovered. And this discovery was done through a process called “Data Mining”.

Now there are chances that you are wrong about Alex, Jessica and Paul. But there are fare amount of chances that you are actually right. That is why it is very important to “evaluate” the result of KDD process.

Now, the reason I gave you this example is I wanted to make a clear distinction between knowledge and information in the context of data mining. This is important to understand our first question – why retrieving information from deep down of your database is not same as data mining. No matter how complex the information retrieval process is, no matter how deep the information is located at, it's still not data mining.

As long as you are not dealing with predictive analysis or not discovering "new" pattern from the existing data – you are not doing data mining.

What are the applications of Data Mining?

When it comes to applying data mining, your imagination is the only barrier (not really ☺ there are technological hindrances as well as we will see later). But it's true that data mining is applied in almost any fields starting from genetics to human rights violation. One of the most important applications is in "Machine Learning". Machine learning is a branch of artificial intelligence concerned with the design and development of algorithms that allow computers to evolve behaviors based on empirical data. Machine learning makes it possible for computers to take autonomous decisions based on the data available from past experiences. Many of the standard problems of today's world are being solved by the application of machine learning as solving them otherwise (e.g. through the deterministic algorithmic approach) would be impossible given the breadth and depth of the problem.

Let me start with one example of the application of data mining that enables machine-learning algorithm to drive an autonomous vehicle. This vehicle does not have any driver and it moves around the road all by itself.
The way it maneuvers and overcomes the obstacles is by applying the images that it sees (through a VGA camera) and then using data mining to determine the course of action based on the data of its past experiences.

There are notable applications of data mining in the subjects such as –

Voice recognition

Think of Siri in iPhone. How does it understand your commands? Clearly it's not deterministically programmable as every body has different tone and accent and voice. And not only it understands, it also adapts better with your voice as you keep using it more and more.

Classification of DNA sequences

DNA sequence contains biological information. One of the many approaches of DNA sequencing is through sequence mining where data mining techniques are applied to find statistically relevant patters, which are then compared with previously studied sequences to understand the given sequence.

Natural Language processing

Consider the following conversations between customer (Mike) and shop-keeper (Linda).

> Mike: You have playing cards?
> Linda: We have one blue stack from Jackson's and also one other from Deborah
> Mike: What is the price?
> Linda: Jackson's $4 and Deborah's $7.
> Mike: Okay give me the blue one please.

Now consider this. What if "Linda" was an automated machine? You could probably have the same kind of conversations still, but it would probably had much more unnatural.

> Mike: You have playing cards?
> Robot: Yes.
> Mike: What type of playing cards do you have?
> Robot: We have Jackson's and Deborah's playing cards.
> Mike: What are the colors of the playing cards?
> Robot: Which Company's playing card do you want to know the color of?
> Mike: What is the color of Jackson's playing cards?
> Robot: Blue.
> Mike: What are the prices of Jackson's and deborah's playing cards?
> Robot: Jacksons' playing cards cost you $4 and Deborah's playing cards cost you $7.
> Mike: Ok, then can I buy the blue ones?
> Robot: We do not have any product called 'blue ones'.
> Mike: Can I have the blue color playing cards please?
> Robot: Sure!

I know the above example is a bit of overshoot, but you got the idea. Machines do not understand natural language. And it's a challenge to make them understand the same. And until we do that we wont be able to build a really useful human-computer interface.

Recently, real advancement on natural language processing is done after the application of data mining. Prior implementations of language-processing tasks typically involved the direct hand coding of large sets of rules. But the machine-learning paradigm instead used general learning algorithms — often, although not always, grounded in statistical inference — to automatically learn such rules through the analysis of large corpora of typical real-world examples.

Methods of data mining

Now if the above examples interest you then let's continue learning more about data mining. One of the first tasks that we have to do next is to understand the different approaches that are used in the field of data mining. Below list shows most of the important methods:

Anomaly Detection

This is the method of detecting patterns in a given data set that does not conform to an established normal behavior. This is applied in number of different fields such as – network intrusion detection, share market fraud detection etc.

Association Rule Learning

This is a method of discovering interesting relations between variables in large databases. Ever seen "Buyers who bought this product, also bought these:" type of messages in e-commerce websites (e.g. in Amazon.com)? That's an example of Association Rule learning.

Clustering

Clustering is the method of assigning a set of objects into groups (called clusters) so that the objects in the same cluster are more similar (in some sense or another) to each other than to those in other clusters.
Cluster analysis is widely used in market research when working with multivariate data. Market researchers often use this to create customer segmentation, product segmentation etc.

Classification

This method is used for the task of generalizing known structure to apply to new data. For example, an email program might attempt to classify an email as legitimate or spam.

Regression

Attempts to find a function, which models the data with the least error. The above example of autonomous driving uses this method.

Next we would learn about each of these methods in greater detail with examples of their application. Mean while, let me know if you have any question / suggestion on this article.

Please visit www.dwbiconcepts.com to find more.