Rebecca Jackson

An Introduction to Data Mining

Data mining can be defined as "the analysis of 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, Mannila, & Smyth, 2001). Data mining has become a necessary tool in data analysis due to the technological advances in data retrieval and data storage. These technological advances have resulted in a vast amount of data available to researchers in all areas. Researchers wish to use this data to their advantage in order to locate patterns and relationships in the data. This would allow them to create models and to predict patterns which can be of use in many different applications. Researchers cannot simply attempt to look at each individual entry in the data set as these data sets are quite large with terabytes of data being stored in them. These goals have led to the creation of the field of data mining.

Data mining's ability to locate patterns in relationships in data is useful to a vast variety of research areas. In business, can be used to monitor transactions, which allows the business to be able to better predict their sales. This means that a business will be less likely to sell out or have a surplus of a particular item. It is also used to monitor credit card records, telephone calls, and even government statistics. In other fields, however, data mining has more unique and interesting uses. It is used in various areas of science through different databases such as astronomical images and molecular structures. It is also used in the medial field when it comes to our medical records. With data being recorded in most areas, data mining has a vast number of applications in real world scenarios.

Data sets are a crucial tool in data mining and contain the measurements to be analyzed. Data sets can be referred to as "training data, sample, [or] database" (Hand et al., 2001). We think of data sets as a collection of sets which all contain the same number of measurements. These sets are referred to as objects while the corresponding measurements are called "individuals, entities, cases, objects, or records" (Hand et al., 2001). These sets are typically represented by a data matrix. A data matrix is a $n \times p$ matrix where n refers to the number of objects or rows and p refers to the number of measurements or columns. While a $n \times p$ matrix is used to simply represent the concept of data sets, in reality, data sets are much more complex.

One way in which the representation of data mining results can be characterized is by distinguishing between a model and a pattern. A model "is a global summary of a data set" while a pattern "describes a structure relating to a relatively small part of the data" (Hand et al., 2001). A model can be used to predict the values for unknown variables and when a data set has missing measurements, a model can still make a statement about the object. Patterns on the other hand, characterize which parts of the data behave in certain ways. Models and patterns are often used in conjunction with one another as in order to detect a behavior in a model, we must first have a description on the behavior from the pattern. Models and patterns are both used as representations of data mining results.

While data mining results are characterized as either models or patterns, the objectives of data mining, or tasks, can be categorized at least generally as either exploratory data analysis, descriptive modeling, predictive modeling, discovering patterns and rules, or retrieval by content. The goal of exploratory data analysis or EDA is to explore the data though there is not necessarily an idea of what they are looking for. EDA techniques are

usually more interactive and visual in nature than other methods. Low dimensional data sets have many graphical models which can be used in EDA. The goal of descriptive modeling is to provide a description of all of the data, which is also considered a generalization of the data. Descriptive modeling can be seen in models for "the overall probability distribution of the data, partitioning of the p-dimensional space into groups, and models describing the relationships between variables" (Hand et al., 2001). Predictive modeling has the goals of classification and regression. The goal is to create a model that allows a value to be predicted using the known values from the data. The predicted value may not be numerical in nature though will determine a potential future result. The goal of discovering patterns and rules differs from the categories listed above in that is not part of the creation of models. A problem with this task is that outliers must be detected in order be able to more clearly classify and recognize the patterns in the data. The last category of data mining tasks is retrieval by content. The goal here is to take a known pattern which is of interest and find similar patterns in the data set. This task is frequently used when analyzing a data set consisting of text. While the five categories of tasks are different, there are many common components between them. Many tasks look into the similarities or distance between any two data vectors, though any approach to analyzing data sets can be beneficial.

In order to analyze data for the completion of the tasks listed above, one must first develop a data mining algorithm. A data mining algorithm will have four basic components. The first component is a model or pattern structure. This is the structure we want to represent the data. The next component is the score function. The job of the score function in the algorithm is to judge the quality of a fitted model. An example of a well-known score function is the squared error score function defined by $\sum_{i=1}^{n}(y(i) - \hat{y}(i))^2$ (Hand et al., 2001). The

third component of a data mining algorithm is an optimization and search method. The goal of this component is to optimize the score function and to search through different models and patterns in order to get the result that yields the best model or pattern for the data set. The final component of a data mining algorithm is the data management strategy. This component controls the handling of data access or "the ways in which the data are stored, indexed, and accessed" during the search and operation component of the algorithm (Hand et al., 2001). These four components form the basis for a data mining algorithm.

Even though statistics may not be sufficient to address all aspects and issues in data mining, it plays a key role. Statistics is frequently most ineffective with large data sets as straightforward facts and information about the data may not be available. Statistics cannot make conclusion about incomplete data sets which is where data mining is necessary in drawing conclusion. Another fundamental difference between data mining and statistics is what is considered a large data set. For a statistician, a large data set contains somewhere between a hundred and a thousand data points while with data mining, data sets with gigabyte and terabytes of entries are common. Large data sets are analyzed through sample sets when the task is to create a model or by summarizing records to generate sufficient statistics. Ultimately, there are many problems which occur and must be overcome when using a statistical analysis of a data set. Even through these problems exist, statistical analysis is still a key part of data mining.

Without time constraints, for any data set, we are able to find a model which will be of an arbitrary fit. While this is certainty true, the factors of complexity and size mean that this is not always a task that can be completed in a reasonable amount of time. Because of this, different methods of data mining must be used which look at the meaning of the data rather

than simply just the data itself. Data mining is a field of study which can be of benefit to many. It is used in all aspects of research from business to science. It allows researchers to locate patterns in data and to form models of the data. This allows them to make predictions for unknown variables. Data mining has revolutionized the way in which people interact with and analyze data making it more beneficial to everyone.

References

Hand, David J.; Mannila, Heikki; Smyth, Padhraic. Principles of Data Mining (Adaptive Computation and Machine Learning series). The MIT Press. Kindle Edition.