Homework 1

ERIN WITMER

CS440

ACM Reference Format:
Erin Witmer. 2019. Homework 1. 1, 1 (January 2019), 5 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 PART I

Textbook problems (Chapter 1, 3rd edition)

1.1 What is data mining? (1.1)

- *Is it a fad?* According to the textbook, data mining is the "process of discovering interesting patterns from massive amounts of data." Over the past decade, there has been a proliferation of devices with the capacity to gather a substantial amount of data. Over the same period of time, the cost to store data has decreased asymptotic to zero and computer processing speeds have increased exponentially. Due to these trends, businesses, governments, researchers and other interested parties have access to massive amounts of data and the computing power to process data in a cost-effective way. In my opinion, data mining– the process by which this massive amount of data is analyzed for valuable insights– is not another hype. When properly executed, insights derived from the data mining process have the potential to aid in everything from increasing the profitability of a company to diagnosing a disease. As data becomes more prolific, the data mining process becomes more important. The process is not simply number crunching. Care must be taken to clean, integrate and pre-process the data to avoid "garbage in, garbage out". Patterns that emerge need to be evaluated for spurious correlations and the golden rule of statistics should not be forgotten: *correlation does not equal causation.*
- *Is it a simple transformation or application of technology developed from databases, statistics, machine learning and pattern recognition?* Per the textbook, data mining incorporates many of the techniques from all of these domains. Data mining is closely related to all of these disciplines. It is not uniquely a transformation or application of one of these disciplines but rather an evolution of all of them in response to the trends outlined above and the need for "effective, scalable, and flexible data analysis in our society". The focus in data mining on effectiveness, scalability and flexibility is noteworthy. As data sets become massive and diverse, data mining techniques must be computationally efficient, handle various data types and produce effective results. This is distinct from the more constrained and controlled worlds of traditional disciplines such as statistics.
- *Describe the steps involved in data mining when viewed as a process of knowledge discovery.*

Author's address: Erin Witmer, ewitmer@ur.rochester.edu.

XXXX-XXXX/2019/1-ART $15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

- `Data cleaning`: Removing noise and inconsistent data from a data source.
- `Data integration`: Combining multiple data sources, typically storing it in a data warehouse, ideally in a way that makes analyzing and querying the data relatively straightforward.
- `Data selection`: the process of extracting the relevant data from the database, for example through an SQL query.
- `Data transformation`: Taking the raw data from the data selection process and aggregating, summarizing or consolidating it in a way that can be mined.
- `Pattern discovery`: Analytical methods are applied to the preprocessed data to extract patterns.
- `Pattern evaluation`: Not all patterns are interesting. This step helps determine how interesting a pattern is based on whether the pattern is easily understood by humans, is valid on new test data, is potentially useful, and is novel. "An interesting pattern represents knowledge".
- `Knowledge presentation`: The knowledge derived from the data mining process must be presented to the end users in a way that is meaningful and useful. This process may entail data visualization or summarization.

1.2 Define each of the following data mining functionalities. Give examples of each data mining functionality, using a real-life database that you are familiar with. (1.3)

A real-life example of a database is the Wegmans shopping app (I am familiar with this database as a user, not an administrator). On the backend, the data warehouse likely includes a product database, customer database and a transaction database. Definitions are paraphrased from the text.

- `Characterization`: Mines the general characteristics or features of a target class of data. An example of this might be summarizing the characteristics of customers who regularly spend over 100 dollars per week. The general characteristics of these customers may include people of a certain gender, age, household size, household income or zip code. This general profile is an example of characterization.
- `Discrimination`: the general characteristics or features of a target class versus one or more other classes. An example might be comparing products that have had over 20 percent month over month sales increase vs. over 10 percent month over month sales decrease. The general characteristics of the first may include "fermented foods and beverages" versus the second "gluten-free foods and beverages".
- `Association and Correlation Analysis`: The analysis of itemset that typically occur together or in sequence. For example, analysis of my transactions at Wegmans might reveal that if I buy cold medicine, it is also highly likely I will buy tissues and tea.

- `Classification`: The process of finding a model that describes and distinguishes data classes or concepts. An example of this might be training a neural network of my historical purchases to come up with a predictive model to predict whether or not I would buy a new product.

- `Regression`: This is similar to the classification process, but typically the outcome is a continuous rather than discrete variable. An example might be predicting the amount of money I will spend next week based on my historical weekly spending.

- `Clustering`: Analyzes data objects without labels. For example, my purchase history might cluster me with other health-conscious, price-sensitive, mothers of young childrenâĂŤeven though those specific labels are not available in my profile..

- `Outlier analysis`: Identifying data that does not comply with the general behavior of the dataset or model. For example, if I were to buy cat food, this would be an outlier (I am allergic to cats).

1.3 Present an example of where data mining is crucial to the success of a business. What data mining functionalities does this business need (e.g. think of the kinds of patterns that could be mined)? Can such patterns be generated alternatively by data query processing or simple statistical analysis? (1.4)

Stitch Fix is a 3 billion dollar personal styling company that sends customized boxes of clothing to customers. The company has been explicit about the role data mining has played in the success of their business. They have been very transparent with the way they leverage data in nearly every aspect of their business operations. Here is an overview of how they utilize data in their business: https://algorithms-tour.stitchfix.com. "Our business model enables unprecedented data science, not only in recommendation systems, but also in human computation, resource management, inventory management, algorithmic fashion design and many other areas. Experimentation and algorithm development is deeply engrained in everything that Stitch Fix does." Stitch Fix utilize all of the data mining functionalities outlined in the text in various capacities. When a user signs up, they fill out a style profile and this information can be mined to categorize customers. A sophisticated classification model ranks the articles of clothing based on the likelihood the customer will like the clothing.

In addition to machine generated recommendations, human stylists are involved in the selection process, but association and correlation analysis is used to match the customer with the human stylist most likely to select styles the customer likes. The warehouse the customer will be receiving the shipment from is determined by a regression analysis that incorporates the cost of shipping and inventory availability. Clustering analysis is used to understand the state of the customer, where they are in the customer lifecycle and how their needs can be anticipated based on where they are. The data mining methods utilized by Stitch Fix are far more sophisticated and multi-dimensional than simple statistical analysis or data queries. Stitch Fix has massive amounts of multidimensional and varied data. Their data includes everything from a customerâĂŹs body measurements to the images that the customer has pinned on Pinterest. The insights generated by the data are the core differentiator and arguably the greatest asset of their business.

1.4 Explain the difference and similarities between: (1.5)

Definitions are taken from the textbook.

- Discrimination and classification:

  - Similarities: Both classification and discrimination generally deal with categorical data and discrete variables rather than continuous variables.
  - Differences: “Discrimination is the comparison of the general features of the target class versus the general features of one or more contrasting classes.” “Classification is the process of finding a model the describes and distinguishes data classes or concepts.” For example, discrimination would be comparing the basic features of shirts versus pants, while classification is the process of finding a model to determine from a picture whether it is more likely to be a shirt or a pair of pants. Discrimination is a descriptive process, whereas classification is a predictive process.
- Characterization and clustering:
  - Similarities: Both characterization and clustering are used to identify homogenous groups of data. For example, you could use either of these methods for identifying and targeting a group of customers with a specific marketing message.
  - Differences: “Characterization is the summarization of the general characteristic or features of a target class.” In clustering, “objects are clustered or grouped based on the principle of maximizing intraclass similarity and minimizing interclass similarity.” Clustering analyzes data objects without class labels but may be used to generate labels for groups of data. Characterization extracts the general features of a specified group of data.
- Classification and regression:
  - Similarities: Classification and regression are both forms of predictive analysis. Training data is used to establish a predictive model. New data is then fed through the model to predict the label or outcome.
  - Differences: The primary difference between classification and regression is that “classification predicts categorical (discrete, unordered) labels, regression models continuous valued functions.”

1.5 Outliers are often discarded as noise. However, one person’s garbage could be another’s treasure. For example, exceptions in credit card transactions can help us detect the fraudulent use of credit cards. Using fraudulence detection as an example, propose two methods that can be used to detect outliers and discuss which one is more reliable. (1.7)

Two methods that could be used to detect fraudulent use of credit cards are clustering analysis and classification analysis. A clustering analysis would be more reliable because it is better at detecting an anomaly if one element of the object vector is off and it does not require labels such as "fraudulent" and "valid" to detect an outlier. For example, if the fraud detection model incorporates the following factors: [time of purchase, amount of purchase, type of establishment, physical location] and I typically purchase a 2 dollar Starbucks coffee in the Rochester area at 3pm on Mondays, clustering analysis is more likely to detect fraud if my card is used to purchase a 2 dollar Starbucks at 3pm on a Monday in Los Angeles. One element of the vector being way off would be identified as an outlier in a clustering analysis. A classification analysis would not work as well because there are typically very few data points with the “fraudulent” label for a purchase history. As a result the model would not likely do as good a job classifying “fraudulent” vs. “valid” transactions.

1.6 What are the major challenges of mining a huge amount of data (e.g. billions of tuples) in comparison with mining a small amount of data (e.g. data set of a few hundred tuple)? (1.9)

The main challenge presented when mining a huge amount of data is the efficiency and scalability of the data mining algorithm used to store and process the data effectively. According to the textbook: "data mining algorithms must be efficient and scalable in order to effectively extract information from huge amounts of data." In other words, finding an algorithm that can process the data effectively, efficiently and increasingly in real-time, is the main. Another challenge is dealing with the increasing diversity of data types. Data is no longer of a uniform data type and stored neatly in a relational database. It is diverse, dynamic and distributed. One way to deal with these challenges is to utilize parallel and distributed data-intensive mining algorithms. Similar to a classical merge sort algorithm, the data is partitioned, processed, and then merged.

REFERENCES

Han, Jiawei, et al. Data Mining: Concepts and Techniques. Morgan Kaufmann, 2012.