Section 1: Week 1: Analyze an Organization’s Data Mining Assets

Nate Bachmeier

TIM-8130: Data Mining

February 9th, 2020

North Central University

Analyze an Organizations Data Mining Assets

## I: Cite Specific Examples of Data Mining Techniques

The four major categories of data mining are association rule mining, clustering techniques, classification techniques, and regression modeling (Barua & Mondal, 2019). Association rules are patterns that take the form of ‘if X then Y,’ such as a person that buys bread is likely also to purchase butter. Clustering and classification are similar strategies that both attempt to group similar items into buckets. The key difference is that classification knows the bucket labels ahead of time (supervised) while clustering does not (unsupervised). For instance, a teacher gives their class a quiz then maps the students into groups by their assessment score (e.g., A/B, C/D, and F) is a classification problem. If they mapped the students on their favorite color, then the groups are not deterministic, and it’s a clustering scenario. Regression modeling tries to find a mathematical equation that explains the observations. A classic example estimates housing prices by considering the features like square footage, age of the house, and the number of rooms. Across these high-level categories, numerous scenario-specific algorithms can be applied to different data sets. For instance, Apriori-based Algorithms rely on the concept that subsets of frequent itemsets must also be frequent itemsets to prune the search space and timely report recommendations (Edureka, 2016) (Giraldo Mejia et al., 2017). Another use case comes from Self-Organizing Maps that cluster or categorize arbitrary data for anomaly detection (Sonmez et al., 2018). Then consider Ant Colony Optimization and Genetic Algorithms, which combine random guessing and regression modeling to iterate toward optimal solutions (Mirjalili, 2018) (LeiosOS, 2017). Other strategies exist to handle count-less other challenges like dimension reduction (e.g., Principal Component Analysis) and brute force discovery (e.g., Parameter Sweeping) (Starmer, 2017).

## II: Organizational Use of Data Mining

Many financial investment firms rely on different automated strategies to filter the sea of market data into a manageable number of options. For example, Fonskea and Liyange (2008) propose a data mining strategy that relies on tracking correlation of related companies (e.g., FedEx and UPS) and profiting from deviations. In this case, both shipping companies are likely to experience similar political and economic headwinds. Bhoopathi and Rama (2017) propose an Apriori-like algorithm that attempts to derive trading signals based on implicit associations between instruments. Hargreaves and Yi (2012) use a decision tree model to filter the Australian index from 2000 companies down to a high-quality basket of the top six. Finally, George and Changat (2017) transform daily quotes into a connected graph to determine the criticality of market intradependencies.

## III: Explain Challenges Experienced Using Data Mining

There’s a joke that ‘70% of all statistics are made-up,’ which infers that without properly evaluating correlation versus causation, the model is unlikely to work in practice. Carver (2007) touches on this point with guidance that researchers focus on relevance, not “just seeing what we want to see.” Snee (2015) echos this point that high-quality models are both practical and explainable. In Fonseka and Liyanage’s solution, they accurately predict prices would fall, and impressive feat had it not considered any context other than the Great Recession. Bhoopathi and Rama’s association rules discovered tight relationships between Intuit (creator of TurboTax) and International Flavors and Fragrance. While there a statistical model that can justify that decision, there’s no economic reason to believe it. George and Changat experienced similar contextual sensitive challenges as they only explored 2016.

## IV: Develop a Code of Procedures for Data Mining

## What are the limitations of data mining

## How would you implement the data mining project

# Section I: Literature Review

## A practical approach to Data Mining (2015)

When a data mining project is beginning, two critical aspects are (1) the identification of the problem statement and (2) discovering facts that will support or refute the answer (Snee, 2015). Without a clear understanding of either the project is doomed to fail. For instance, if a researcher wants to prevent global warming, they would need to identify a specific scenario and relevant scientific measurements. These measurements could be erroneous due to miscalibrated equipment or poor information governance policies. It can be challenging to detect these inaccuracies without some domain-specific knowledge of the subject matter. Snee recommends extracting descriptive statistics of a data set before investing too much time into it. An example data set might contain hourly sensor readings—are the measurements at consistent intervals, and are the range of value rational? Garbage-in leads to garbage-out that has no value and will be completely arbitrary.

## Data Processing and Text Mining Technologies on Electronic Medical Records (2018)

Electronic Medical Records (EMR) contain “a treasure trove of information for large scale analysis of health information (Sun et al., 2018).” Professionals are already gaining insights from the structured sections, such as the patient demographics and prescription metadata, though large sections of the file are unstructured. For instance, free form clinical and surgical notes can be difficult to index though highly relevant to automated diagnostic processes.