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 procedure for data mining

## How would you implement the data mining project

Sun et al. (2018) describe the challenges associated with data mining Electronic Medical Records (EMR). They state these records contain both structured and unstructured information, which requires distinct tools and approaches. For instance, analyzing demographic metadata and prescription refills from standardized forms is relatively trivial compared to MRI photographs or free-formed clinical notes. The authors use a circular data processing feedback loop across data collection, preprocessing, mining, and then evaluation. This approach allows more specialized exploration to augment and extend generalized observations. For example, the system notices that Alice has high blood pressure and that feedback causes another process to query her parent’s medical information for hereditary markers. It turns out that her father died of cardiac arrest, which drives more feedback to avoid specific prescriptions. Another challenge comes from the decentralized heterogeneous data sources that need to feed into this system. Sun et al. propose a data cleaning process that normalizes identifiers, deduplicates records, and also anonymizes subsets of shareable information. They note that normalizing identifiers requires a separate system due to the complexity caused by spelling mistakes, locale preferential terms, and other disambiguation scenarios.

Having distinct phases makes the data mining system more maintainable and also addresses the criticality of reproducible results. One area that they could improve is giving more focus to the specific health aspects they are attempting to uncover. Few pharmaceutical or insurance organizations have a sufficient budget to address everything always. Instead, most research begins with a problem statement and a narrow focus, such as reducing heart failure or pancreatic cancer. Along with identifying the research scope, they would also benefit from clarifying the mining strategies (e.g., family tree graphs versus temporal attribute analysis), as that will influence the shape of persisted records. Another observation is that Sun et al. describe a medical data management system, though data mining needs to be both analytical and graphical (Snee, 2015). Without an efficient reporting and exploration strategy, it can be difficult to have a conversation about the results. Their data management system also ends at an OLAP data warehouse, though many data formats (e.g., audio and video) are easier to explore within a data lake (McKendrick, 2019) or purpose-built NoSQL solution (e.g., graph database).

## What are the limitations of data mining