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

~~Many organizations are collecting large quantities of dark data with the hopes of one day, converting it into business intelligence (Ajis & Baharin, 2019). This conversion will not happen magically, as it requires careful planning that begins with the identification of specific questions and research objectives. These goals need to leverage a framework of interconnected processes that transform data into predictive models that describe hidden relationships across different heterogeneous data sources. Data practitioners meet this need with data mining, an automated process of knowledge discovery from large volumes of data, involving databases, statistics, and artificial intelligence (Giraldo Mejia et al., 2017).~~

# Section 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).

# Section II: Organizational Use of Data Mining

~~All competitive organizations require a strategy to operationalize their data into business intelligence so that they can adapt to dynamic market conditions (Knabke & Olbrich, 2018).~~ Contoso, a financial investment firm, relies on several automated strategies to filter the sea of market data into a manageable number of options. According to Snee (2015), a high-quality data model should be practical and explainable. For example, Fonskea and Liyange (2008) propose a data mining strategy that relies on tracking correlation of related companies and profiting from deviations. Consider FedEx and UPS

# Explain Challenges Experienced Using Data Mining

# 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.