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

Nate Bachmeier

TIM-8130: Data Mining

February 9, 2020

North Central University

Analyze an Organizations Data Mining Assets

## 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 related strategies that both attempt to group similar items into buckets. The critical 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) is a classification problem. If they mapped the students on their favorite color, then the groups are not deterministic, and it is 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 are available for 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

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, such as when X goes up, then Y follows. Hargreaves and Yi (2012) use a decision tree model to filter the Australian index on fundamental data (e.g., return on equity) 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 interdependencies between participants.

## Section III: Explain Challenges Experienced Using Data Mining

There is 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. Both Fonseka & Liyanage’s and George & Changat’s did not account for the contextually sensitive results caused by the Great Recession occurring in parallel. Bhoopathi and Rama’s association rules discovered tight relationships between Intuit (creator of TurboTax) and International Fragrance—with no economic justification. Aside from Hargreave and Yi, none of these approaches even had a basis in modern market theory. For instance, correlations between price movements did not account for volume. The authors also limited their asset analysis to only primary assets, instead of expanding into secondary assets. George & Changat determined that banks were the most critical aspect of their network, but did not investigate interest rates, GDP, nor consumer credit statistics. Bhoopathi and Rama could have transformed the data with a moving average to smooth out noise, decreasing false-positive rules.

# Section IV: Develop a procedure for data mining

## Implementing 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 (e.g., neural network image classification) or free-formed clinical notes (e.g., NLP). 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. 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 on the specific health aspects they are attempting to uncover. Few pharmaceutical or insurance companies 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. However, data mining needs to be both analytical and graphical (Snee, 2015). Without an efficient reporting and exploration strategy, it can be challenging to have a conversation about the results. Their data management system also ends at an OLAP data warehouse. However, many data formats (e.g., audio and video) are more comfortable to explore within a data lake or purpose-built NoSQL solution (Barua & Mondal, 2019) (McKendrick, 2019).

## What are the limitations of data mining

Many organizations are collecting vast 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. Even with well-formulated plans, the necessary measurements could be erroneous due to miscalibrated equipment and poor information governance policies. It can also be challenging to detect these inaccuracies without some domain-specific knowledge of the subject matter. Gaining the domain-knowledge is often complicated and relies on cross-organizational communication, and that can introduce political constraints due to competing for business priorities (Al-Sai, Abdullah, & Husin, 2019). After overcoming these obstacles, the tooling relies on non-standard data interfaces, dynamically typed languages, and a lack of built-in parallelism (Zeehan et al., 2019).

# References

Ajis, A., & Baharin, S. (2019). Dark Data Management as a frontier of Information Governance. 34-36.

Al-Sai, Z., Abdullah, R., & Husin, M. (2019). Big Data Impacts and Challenges: A Review. *2019 IEEE Jordan Internation Joint Conference on Electrical Engineering and Information Technology (JEEIT).*

Barua, H., & Mondal, K. (2019). A Comprehensive Survey on Cloud Data Mining (CDM) Frameworks and Algorithms. *CM Computing Surveys. Sep2019, Vol. 52, Issue 5, p1-62. 62p*, 1-62.

Bhoopathi, & Rama. (2017). A Novel Framework for Stock Trading Analysis Using Casual Relationship Mining. *2017 Third International Conference on Advances in Electrical, Electronics, Information, Communication, and Bio-Informatics (AEEICB)* (pp. 136-141). AEEICB.

Carver, S. (2007). Feedback Loop. *Sage Reference: Encyclopedia of Social Psychology*.

Edureka. (2016). *What is the Apriori Algorithm*? Retrieved from YouTube: https://youtu.be/guVvtZ7ZClw

Fonseka & Liyanage. (2008). A Data mining algorithm to analyze stock market data using lagged correlation. *2008 International Conference on Information and Automation for Sustainability (ICIAFS)*.

George & Changat. (2017). A network approach for Stock market data mining and portfolio analysis. *2017 International Conference on Networks & Advances in Computational Technologies (NetACT).*

Giraldo Mejia et al. (2017). Knowledge-based model to support decision-making when choosing between two association data mining techniques. *Revista Lasallista de Investigación. Jul-des, 2017, Vol. 14, Issue 2*, 41-50.

Hargreaves & Yi. (2012). Does the use of Technical & Fundamental Analysis improve Stock Choice? *2012 International Conference on Statistics in Science, Business, and Engineering (ICSSBE).*

LeiosOS. (2017). *How algorithms evolve*. Retrieved from YouTube: https://www.youtube.com/watch?v=qiKW1qX97qA

McKendrick, J. (2019). Data Lakes and Data Warehouses, Working Tandom. *DATABASE TRENDS AND APPLICATIONS OCTOBER/NOVEMBER 2019*.

Mirjalili, A. (2018, October 4). *How the Ant Colony Optimization algorithm works*. Retrieved from YouTube: https://www.youtube.com/watch?v=783ZtAF4j5g

Snee, R. (2015). Practical Approach to Data Mining: I Have All These Data; Now What Should I Do? *Quality Engineering, Volume 27*, 477-487.

Sonmez et al. (2018). Anomaly Detection Using Data Mining Methods in IT Systems: A Decision Support Application. *Sakarya University Journal of Science, 22(4)*, 1109-1123.

Starmer, J. (2017). *What is Principal Component Analysis*? Retrieved from YouTube: https://www.youtube.com/watch?v=HMOI\_lkzW08

Sun et al. (2018). Data Processing and Text Mining Technologies on Electronic Medical Records: A Review. *Journal Of Healthcare Engineering 2018 April*.

Zeehan et al. (2019). Machine Learning at Microsoft with ML.NET. *International Conference on Knowledge Discovery and Data Mining.*