Data Mining

CSC 595 Business Intelligence and Data Warehouse

Matthew Dunaway

**Problem and Motivation**

Since I do not come from an IT or database background, I had to first define what data mining is. To answer this question, I chose Wikipedia for its broad coverage of most topics and its citation of articles for more specific topics in data mining. **Data mining** is the process of extracting relevant data from a huge collection of digital data repositories. (Chakrabarti et al., 2020) The goal of data mining is to extract textual patterns from a dataset, transform that information into relevant data, and draw conclusions from that data. The data analysis from data mining is also known as **knowledge discovery in databases**, or KDD. (Wikipedia, 2022) Data mining is set up as an automatic process to sift through ambiguous or unknown data to find information patterns that can be surprising, interesting, sequential, or an anomaly. This process requires a software engineer to create programming, or code, to tell a program what to look for in data sets or how to organize a data set search. This search code is better known as **data mining algorithms**. The patterns and associations that these algorithms search for make up a broad array of **data mining techniques**. The patterns are then analyzed, and predictions can be made for several business processes. This research paper will discuss the history of data mining to better understand it’s procedures and purpose. It will then discuss the procedures and give examples based on my own data mining project. Last, the paper will review data mining solutions and which software is best suited for which purpose.

**History**

**The roots of data mining begin in 1763, when a British mathematician named Thomas Bayes created a statistical formula for conditional probability. His basis for the formula was a thought experiment about an unseen ball being thrown on a table. The first ball is thrown but it is unknown which side of the table it landed. A volunteer then continues throwing balls on the table, telling him whether the ball landed to the right or left of the first ball. The more balls that land on the right of the first, the better the chance the first ball landed on the left side of the table. The formula was then postulated as Initial Belief + New Data = Improved Belief. Later in 1791, a mathematician named Pierre Laplace rediscovered the theorem and tested it on birth record data. He was the first scientist to mathematically prove that humans tend to have more males than females born across all races and regions. He published the general theorem for Bayes Theorum as P(C|E) = [ P(E|C)Pprior(C) ] / [ΣP(E|C')Pprior(C'). This probability formula is the backbone for event prediction today. (Hayes, 2022)**

**In 1805, regression analysis was just beginning to take shape with the least square’s method pioneered by Legendre and Gauss. Least squares method is simply the smallest sum of the squares of errors that is used to draw a straight line through a set of data points on an x, y axis. This form of regression is known as linear regression and is the first known method to fit a predictive model to a known dataset based on dependent and independent variables. (Kenton, 2022) The formula, Y = a + bX + u, finds a relationship between variables caused by an event and variables not known to be related to an event. From simple linear regression, an entire field of statistical study has emerged with regression being at the center of choosing variables of events to predict outcomes.**

**In 1936, Allan Turing created an abstract of the first programmable machine called the universal Turing machine. By 1952, he had a working prototype, the automatic computing engine. At the same time, Von Neumann produced a similar computer, the electronic discrete variable computer. The difference was in how they each handled coded instructions. The ENIAC functioned on a central processing unit and became the popular option. While Von Neumann is credited as the creator of the first electronic computer, the ENIAC, he gave credit to Turing for his work in computable numbers and stored program computing. (Copeland & Proudfoot, n.d.) In 1958, the first artificial neural network was invented by Frank Rosenblatt called the Perceptron. It was invented to model the human brain’s ability to learn through optic nerve input. (Kay, 2001) Each processor models a neuron and can be physical or a digital processor. To begin with, all the processors are tasked with solving a problem or question but are set to random values to represent not knowing anything. Through the input and output of data, these processors begin to discover patterns and relationships in the data and refine their values. After this, the neural network can make predictions based on learned patterns and relationships of their individual values to answer the question or problem posed. Another method that was used to give the processors data and evaluate answers as correct or incorrect for a given problem. The stored data and predictions would then be weighted higher or lower for right and wrong answers. Neural networks set the stage for artificial learning** through pattern recognition and data analysis**. (Kay, 2001)**

**In 1970, General Electric hired Bachman to automate its business processes. It needed an efficient way to store and recall business data. Bachman came up with an integrated data store (IDS) that was an intermediary between the data and the program. With this, programs didn’t have to be rewritten every time a database was updated or a different one was accessed. This allowed for multiple departments of a business to be accessed through one program and larger databases being collated. (Haigh, 2016) Many systems improved upon his design until today we have an efficient database management system to peruse large data for statistically significant data analysis results.**

**Next was genetic algorithms theorized in 1975 by John Holland. He studied the crossing over action of chromosomes and modeled an algorithm with symbols to mix and breed possible data solutions with variety introduced through symbol mutation. A fitness function is used on each generation of algorithms and set to breed again for a higher function score.** **(Hosch, n.d.) This innovation sets the foundation for data mining because the technology world has hardware and software capable of learning. If not for Charles Bachman though, the capacity to store large amounts of data would have never reached a level for data mining to be practical.**

**Data mining was finally formally recognized and defined in 1989 by The Association for Computing Machinery. They outlined a process for discovering useful knowledge from large collections of data through automation. Before this point, businesses relied on manual knowledge discovery by individuals sifting through data. Eventually, datasets became too large for efficient knowledge discovery and data mining pioneers were already experimenting with data mining. ACM decided that the IT industry needed to agree on the fundamental process of data mining to keep information theory accurate. (Techopedia, 2017) The process they came up with is identify customers KDD goal, understand the data domain and data required, select your target data for discovery, clean the data to resolve formatting and missing fields, filter the data of unnecessary variables, match customer goals with data mining methods, choose the data mining method to suggest hidden patterns, choose the data algorithm to discover the hidden patterns, search for patterns of interest in a specific representational form, interpret knowledge from mined patterns, use and incorporate the new knowledge, and document the knowledge for interested parties. (Techopedia, 2017) Through the 1990’s and 2000’s, data mining became popular for industrial, commercial, and financial applications with the rapid improvement of database capacity and processor speed. In 2003, data mining finally became a public topic with the Oakland A’s use of data mining to staff a roster of a major league baseball team for 33% of the average MLB team salary and make the MLB playoffs. This was proof to the world that data mining could be a feasible application for any business transaction or scientific observation looking to cut costs and increase performance based on already observed data. Today, data mining is available to almost anyone with a dataset and a computer able to download and run SQL server. It plays a role in analytic decision-making process of universities, businesses, hospitals, civic institutions, and the worlds’ militaries.**

**Data Mining Procedure**

**We have already defined data mining and its history. Now, lets discuss how data mining works. Data mining follows an industry standard approach called the CRISP-DM method. It stands for Cross Industry Standard Process for Data Mining- Daimler Benz. It was created in 2000 by a group of data analysts working for Daimler Benz that wanted to market the emerging field of data mining to the public but needed an industry recognized standard to let their customers know they were doing it right. These data analysts sent an industrywide invite to a conference on data mining and got the input of all the scientists in the field. Finally, the standard was published, and it is followed by all professionals in the field. (Chapman et al., 2000) Phase 1 is business understanding. Understanding what question to ask allows you to understand what data you will need to solve it. This is the second phase, data understanding. Understanding the data means selecting data that has the variables you will need to answer your question. For these phases, I will use stock data of five retail companies. Phase 1, I want to ask the question, did small rural retailers perform better than large, centralized retailers over the past 22 years? Phase 2, I know that the best metric to judge a companies’ success is their adjusted stock price. Phase 3 begins the data preparation phase. This is where the data is extracted, transformed, and loaded. For my phase 3, I need to go to Yahoo Finance and extract the stock data for 5 companies, TGT, WMT, DTR, DG, BIG, for the last 22 years to get an adequate sample size for time and number of retailers. This is downloaded and transformed into data that is all in the same formal and free of null values and duplicates. My stock prices don’t have any duplicates, but I must standardize the dates across all the data. Next, my stock prices are combined to one file and loaded into a database for further use. Phase 4 is data modeling. This is where a modeling technique is chosen based on the business objectives. The five most common modeling techniques are classification analysis, association rule learning, anomaly detection, clustering analysis, and regression analysis. *Classification analysis* is good for learning how attributes of data reached a conclusion. *Association rule learning* is a technique for discovering relationships between variables on their way to conclusions. *Anomaly detection* is a technique for discovering data that doesn’t follow the pattern of the other data. This is often used for fraudulent activity in financial data. *Clustering analysis* finds object data like other objects and groups them based on these similarities. Clustering is especially good when data is combined, and objects aren’t defined by characteristics yet. This model is the first model I want to run on my stock history data. The last is *regression analysis*. This model technique is used when a trend is recognized but needs to be predicted and quantified. This is often used for inclement weather and grocery store product predictions. (**Rutgers Bootcamps, 2022) Phase 5 is the model evaluation, where the person asking the business question determines if the model answered that question. My clustering model DLTR and DG together in a group and TGT, WMT, and BIG clustered in another group. DLTR and DG stock price grew by over 750% in 22 years and the other three did not. This answers my question on whether rural retailers are performing better economically than centralized retailers. Phase 6 then begins for me with deployment. **Deployment** is a visual presentation, a corporate report, simply an action for a new business strategy. Phase 6 for stock history data will be writing this report and performing a Tidy Tuesday report in RPubs. Actioning the data mining results can also lead to new insights and questions, so deployment can be used to feed into the CRISP-DM model again.

**Data Mining Solutions**

What are the data mining solutions/software packages available, and which one is best? There is no easy answer for which data mining solution is best. Of the top 10 lists I found though, There are a few that always pop up, such as RapidMiner, Oracle Data Mining, and SPSS. To find the best data mining tool for your use, it is best to decide what you are using it for. Generally, data mining platforms work in two programming languages, R and Python. R is most useful for statistical and data analysis while Python is good for visualization and deployment. Here is a list of process suitability I created for different data mining software. **Statistical analysis** relying heavily on R and can also use SPSS, SAS, Oracle Data Mining, and R Studios. Free, open-source solutions include KNIME, RapidMiner, and Orange. Big data mining can try Apache Spark, Hadoop MapReduce, and QLIK. Academic and small-scale institutions should start with Scikit-learn, Rattle, Pandas, and H3O. Enterprise and cloud solutions should be run with Amazon EMR, Azure ML, or Google AI Platform. Neural Network software solutions include PyTorch and TensorFlow. Finally, visualization heavy projects should use Matplotlib and ggplot2. (Berga, Coelho & Ochman, 2021) Of the listed data mining software, I have personally tried SQL Server Data Tool, ggplot2, Pandas, and SSPS. I used SSPS to visualize my stock history data and tried to run predictions, but the analysis choices are very in-depth, statistically speaking. SQL Server Data Tool has a very robust library of tutorials but can rely on other programs for parts of the data mining process, which makes it prone to connection errors. Connection errors were also inherent in Pandas, though the ETL and visualization process were easy to understand. Ggplot2 was my favorite because of the amount of code examples available on RPubs and YouTube. Visualization was difficult but very customizable. Overall, I would recommend ggplot2 in R Studio to try first with downloadable datasets and customizable code to polish your knowledge and data mining skills.

**Conclusion**

**The purpose of my research paper was to better understand data mining. I defined data mining as** the process of extracting relevant data from a huge collection of digital data repositories. (Chakrabarti et al., 2020) The goal of data mining is to extract textual patterns from a dataset, transform that information into relevant data, and draw conclusions from that data. To better understand why data mining was needed I explored the historical background of events that led to its formal industry acceptance. I outlined the industry specific guidelines of the formal steps in the data mining process and performed those steps on my own stock history dataset as an example. I then listed the data mining software specific for the types of business questions being asked. Last, I relayed my own experience with several data mining solutions to convey a better understanding of their practical usefulness. After writing this report, I have a much better understanding of what data mining is and how it is used. After reading this, it is my hope you will have a better understanding of data mining as well.

**References**

1. Chakrabarti, S., Ester, M., Fayyad, U., Gehrke, J., Han, J., & Morishita, S. et al. (2020). Data Mining Curriculum: A Proposal. Retrieved 28 March 2022, from <http://www.kdd.org/curriculum/index.html>
2. Wikipedia. (2022). Data mining - Wikipedia. Retrieved 28 March 2022, from <https://en.wikipedia.org/wiki/Data_mining#cite_note-acm-1>
3. What Is Data Mining? A Beginner's Guide (2022) | Rutgers Bootcamps. (2022). Retrieved 11 April 2022, from <https://bootcamp.rutgers.edu/blog/what-is-data-mining/>
4. Hayes, A. (2022). Bayes' Theorem Definition. Retrieved 11 April 2022, from <https://www.investopedia.com/terms/b/bayes-theorem.asp>
5. Kenton, W. (2022). Understanding the Least Squares Method. Retrieved 12 April 2022, from <https://www.investopedia.com/terms/l/least-squares-method.asp#:~:text=The%20%22least%20squares%22%20method%20is,relationship%20between%20the%20data%20points>.
6. Copeland, J., & Proudfoot, D. Turing, Father of the Modern Computer. Retrieved 12 April 2022, from <http://www.rutherfordjournal.org/article040101.html>
7. Kay, A. (2001). Artificial Neural Networks. Retrieved 12 April 2022, from <https://www.computerworld.com/article/2591759/artificial-neural-networks.html#:~:text=The%20first%20artificial%20neural%20network,ANNs%20to%20study%20human%20cognition>.
8. Hosch, W. genetic algorithm | computer science. Retrieved 12 April 2022, from <https://www.britannica.com/technology/genetic-algorithm>
9. Haigh, T. (2016). How Charles Bachman Invented the DBMS, a Foundation of Our Digital World. Retrieved 12 April 2022, from <https://cacm.acm.org/magazines/2016/7/204036-how-charles-bachman-invented-the-dbms-a-foundation-of-our-digital-world/fulltext#body-3>
10. What is Knowledge Discovery in Databases (KDD)? - Definition from Techopedia. (2017). Retrieved 12 April 2022, from <https://www.techopedia.com/definition/25827/knowledge-discovery-in-databases-kdd>
11. Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0 Step-by-step data mining guide. Retrieved 12 April 2022, from <https://www.semanticscholar.org/paper/CRISP-DM-1.0%3A-Step-by-step-data-mining-guide-Chapman-Clinton/54bad20bbc7938991bf34f86dde0babfbd2d5a72>
12. Berga, M., Coelho, P., & Ochman, A. (2021). Top 21 Data Mining Tools. Retrieved 13 April 2022, from <https://www.imaginarycloud.com/blog/data-mining-tools/#RvsPython>