John Fields

Dr. Ami Gates

IST707 - Assignment #1

7/11/19

Introduction

The field of data mining has grown rapidly with the advancement of technology and the ability to gather and store large quantities of information and retrieve it quickly from computers. Utilizing this "big data" to provide insight and make predictions has benefits across a wide variety of societal and business areas such as medicine, science and engineering. Chapter 1 of *Introduction to Data Mining* (Tan et al.) provides an excellent overview of the topic of data mining that includes information on the benefits, background, definitions and applications of data mining.

Data mining combines traditional analysis methods from statistics, operations, and other fields with new algorithms to provide new insights to help people make better decisions. The authors define data mining as "…the process of automatically discovering useful information in large data repositories." The authors also emphasize that not all tasks related to information discovery are considered to be "data mining" and the exercises at the end of this assignment will provide examples of data mining compared to other information related activities.

The development of data mining was driven by five specific challenges related to data and information:

1. Scalability (for massive data sets)
2. High Dimensionality (for data with many different attributes)
3. Heterogeneous and Complex Data (for different types of data)

4. Data Ownership and Distribution (for data in different locations)
5. Non-traditional Analysis (for the automation of hypothesis generation and evaluation)

To overcome these challenges, data mining has evolved from being a part of Knowledge Discovery in Databases (see Figure 1 below) to focusing on "…data preprocessing, mining, and postprocessing" by using techniques from statistics, artificial intelligence, machine learning, optimization, visualization, and many other areas. The task performed as part of data mining typically involve Predictive and Descriptive tasks. Predictive tasks involve the prediction of a target or dependent variable (Y) based on the explanatory or independent variables (X's). The first chapter concludes with examples of Classification (iris flower data set) and Association analysis (market basket data) to show some typical data mining tasks and a discussion on the scope and organization of the book.

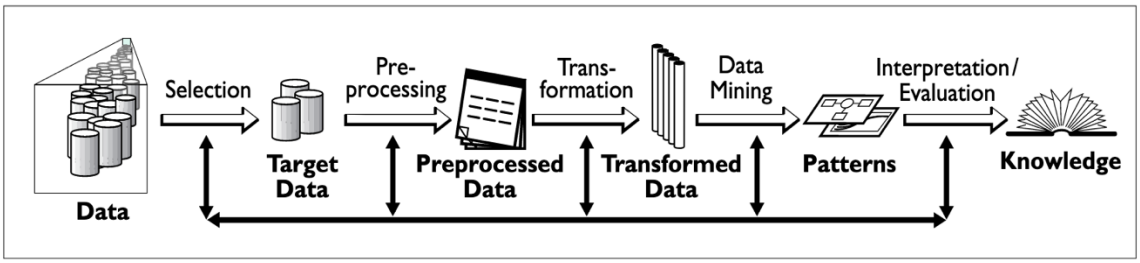

Figure 1. Overview of the steps in the Knowledge Discovery in Databases (KDD) process

Analysis and Models

Task 1 - 1.7 Exercises - review data mining concepts and task

1. Discuss whether or not each of the following activities is a data mining task.

(a) Dividing the customers of a company according to their gender.

This is not a data mining task. A simple database query is required to complete this task.

(b) Dividing the customers of a company according to their profitability.

This is not a data mining task. A simple database query would provide the desired information and then a finance/accounting task is required to determine the groupings to use to divide the customers.

(c) Computing the total sales of a company.

This is not a data mining task. This is a finance/accounting task to sum the sales of the company for financial reporting.

(d) Sorting a student database based on student identification numbers.

This is not a data mining task. A database query could be used to collect this information and sort by student ID numbers.

(e) Predicting the outcomes of tossing a (fair) pair of dice.

This is not a data mining task since you could use statistical methods (sampling, means and central limit theorem) to show that there is a 1 out of 6 chance of rolling one of the numbers on the die.

(f) Predicting the future stock price of a company using historical records.

This is a data mining task. A model could be built that includes the various factors that influence the price of a stock. Time-series analysis could then be performed to predict the expected future stock price.

(g) Monitoring the heart rate of a patient for abnormalities.

This is a data mining task. Data could be collected on normal and abnormal heart conditions to determine the threshold where a notification

should be triggered. This could be accomplished with an algorithm for outlier detection or by classifying the condition as normal or abnormal.

(h) Monitoring seismic waves for earthquake activities.

This is a data mining task. As in (g) above, data could be collected on normal and abnormal earthquake activity. A classification could then be performed to label the activity as normal or abnormal.

(i) Extracting the frequencies of a sound wave.

This is not a data mining task. This information could be queried from the device/system that is collecting the sound wave information.

Task	Data Mining?	Data Mining Technique
(a) Divide customers based on gender	No	
(b) Divide customers by profitability	No	
(c) Compute total sales to customer	No	
(d) Sort students by ID number	No	
(e) Predict outcome of fair dice	No	
(f) Predict stock price	Yes	Time-series forecasting
(g) Monitor heart rate	Yes	Classification and outlier-detection
(h) Monitor seismic waves for earthquakes	Yes	Classification and outlier-detection
(i) Extract sound wave frequencies	No	

Table 1. Comparison of data related tasks

2. Suppose that you are employed as a data mining consultant for an Internet search engine company. Describe how data mining can help the company by giving specific examples of how techniques, such as clustering, classification, association rule mining, and anomaly detection can be applied.

Internet search engine companies (e.g. Google) have pioneered the use of data mining to provide a "free" service in exchange for users supplying behavioral data such as the information that they type in a search engine. This data can be utilized in a variety of ways to understand the

behaviors of users which can be utilized for revenue generating activities such as advertising, up-selling and cross-selling. The chart below shows how different data mining techniques can be used to improve the user experience and provide additional value-added data to an internet search company.

Data Mining Technique	**Benefit to Company**	**Benefit to Users**
Clustering	Group together similar users to target for advertising	More relevant search results
Classification	Predict characteristics such as age and interests (e.g. sports, news)	More relevant search results
Association Rule Mining	Information for advertising to up-sell or cross-sell	Receive information on complementary products and services
Anomaly Detection	Determine if the user is really a malicious bot or potential threat	Protects confidential information and provides user confidence in the security of their data

Table 2. Benefits of Data Mining Techniques for Internet Search Engines

3. For each of the following data sets, explain whether or not data privacy is an important issue.

(a) Census data collected from 1900-1950.

Yes, census information could be a data privacy issue since someone born in 1950 would now be 69 years old. The census information contains information on race, ethnicity, medical conditions, and financial information which some people might be uncomfortable sharing via websites such as Ancestry.com. Other data in the census, such as your mother's maiden name, could be used to circumvent security questions.

(b) IP addresses and visit times of web users who visit your website.

Yes, this information could present data privacy concerns. There are numerous recent examples of web sites being hacked and user information being released that can be utilized by hackers to impersonate others or black mail users.

(c) Images from Earth-orbiting satellites.

No, this does not present a concern today with the current technology. However, if more detailed data is available in the future this could be a cause for concern. An example is Google Maps street view which must manually remove or obscure images which show license plates, faces, etc. If satellites can provide this level of resolution in the future, then it could cause data privacy concerns.

(d) Names and addresses of people from the telephone book.

Names and addresses in the telephone book don't present data privacy issues. This information has been available publicly and the downside is the automated use of robo-calling which has made telemarketing a bigger issue.

(e) Names and email addresses collected from the Web.

Names and email addresses on the web are also not a data privacy issue. Similar to (d) above, it is more of a nuisance now that this information is more readily available and can be utilized by unscrupulous marketers, but it doesn't pose an increased data privacy risk.

Task 2 - Google Flu Trends - practice your critical thinking and writing

In the NY Times article from 2014, *Google Flu Trends: The Limits of Big Data*, the authors review the challenges of using internet searches related to tracking potential outbreaks of the flu. The Centers for Disease Control (CDC) has used manual methods in the past to collect

this information from health care providers which caused several weeks delay in receiving the results. This article discusses the issue where flu cases were overstated 30-50% by Google Flu Trends in the period from 2012-2014 using the faster method of Google search data to predict flu outbreaks.. As the title suggests, there are limits to the use of big data and the NY Times is is critical of Google for not having been more careful about how they used this new technology.

The second article from the Atlantic, *In Defense of Google Flu Trends* (also published in 2014), has a more positive opinion of the role of Google Flu Trends. The author provides additional information on how combining Google Flu Trends with CDC data provides better predictions than each of these independent sources. This article also quotes Google sources who refer back to documents that show that Google Flu Trends made warnings earlier about using their data as a stand-alone source for predictions.

A review of both articles provides interesting insights on the potential "hype" around data mining and the responsibility of companies and data scientists to be cautious with how these new predictive capabilities are utilized in a responsible way. Although Google provided a defense which pointed back to earlier documentation on the use of Google Flu Trends, the company appeared to be more interested in utilizing the early positive press than warning about the potential misleading results. Many of the high-tech companies like Google, Facebook, etc. are learning that they must be very careful about how they deploy new technology to insure it is fully tested and secure. Companies need to be more open about disclosing the risks and potential issues of new data mining capabilities such as Google Flu Trends or the recent government and public outcry could create new laws and regulations if companies are not responsible in regulating how they operate.

Results

The first topic in the analysis section above reviewed the different types of data mining activities and a summary of the different techniques shown in Table 1. This information can be used as a technical reference for determine the types of techniques to apply to different tasks. Similarly, Table 2 provides a review of how data mining can be used for internet searches and the benefits for the company and users. In future assignments, additional technical details will be provided in the Results section of these assignments to provide a comprehensive review of the techniques used.

The second topic in the analysis section was a review of various articles related to the capabilities and challenges with Google Flu trends. These articles highlight the promise and challenges of applying data mining techniques to global health problems such as fighting the annual battle with flu outbreaks. There are lessons here for tech companies like Google and the data scientists who develop and deploy these new capabilities. As the great super-hero Spider Man stated, "With great power comes great responsibility."

Conclusion

The book, Introduction to Data Mining (Tan et al.), is an excellent introduction to the topic of data mining and explains the concepts in concise language and easy to understand examples. Exercise Question 1 in section 1.7, also provided many thought-provoking questions which helps the reader to understand the concepts behind data mining (what it is and what it is not). Question 2 helps the reader to see the value of data science in a fictitious internet search company to explore how a consultant would use different techniques to provide more value to the company and to users. Finally, Question 3 explores the very relevant topic of data privacy

which continues to be a major issue for data science as information continues to become easier to collect and more readily available.

The articles on Google Flu Trends review the very relevant topic of the use of big data/data mining in important areas such as health care. These stories offer different viewpoints on the value and challenges of utilizing these powerful new techniques and available information to help us better predict issues such as flu trends. However, as with any new technologies, there is a learning curve and companies/data scientists need to be very cautious to ensure these capabilities are deployed ethically and responsibly.

Works Cited

Tan, Pang-Ning, et al. *Introduction to Data Mining*. Pearson Education, Inc., 2019.

Fayyad, Usama, et al. "The KDD Process for Extracting Useful Knowledge from Volumes of Data." *Communications of the ACM*, vol. 39, no. 11, 1996, pp. 27–34., doi:10.1145/240455.240464.

"Google Flu Trends: The Limits of Big Data." *New York Times*, 28 Mar. 2014.

"In Defense of Google Flu Trends." *The Atlantic*, 27 Mar. 2014.