IST 707 Applied Machine Learning

HW1

Dixitha Kasturi
dkasturi@syr.edu

Task 1: review data mining concepts and tasks

Answer the exercise questions 1-3 in Textbook 1.7. For Question 2, feel free to change the question scenario from "an Internet search engine company" to any organization that you would like to think of. It can be a company, government office, NGO, etc.

See questions 1-3 as below:

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
    Not a Data mining task as it is simple filtering which can be done through a query or if the data isn't huge excel as well.
    b. Dividing the customers of a company according to their profitability
    Not a Data mining task, it is just filtering and grouping which can be done through a query or if the data isn't huge excel as well.
    c. Computing the total sales of a company
    Not a Data mining task, as it is simple addition/summation
    d. Sorting a student database based on student identification numbers
    Not a Data mining task, simple sorting
    e. Predicting the outcomes of tossing a (fair) pair of dice
    Not a Data mining task despite being a predictive one because of the constant set of possible outcomes. 36 combinations but all having the same probability.
    f. Predicting the future stock price of a company using historical records
    Data mining task as it involves generating a value. Typically regression is used in these cases.
    g. Monitoring the heart rate of a patient for abnormalities
    Data mining task. It involves observing patterns through real-time data processing.
    h. Monitoring seismic waves of earthquake activities
    Data mining task. This too involves observing patterns through real-time data processing and possibly use historical records of seismic waves that caused earthquakes to predict/identify potential earthquake activities. This can be both a
    i. Extracting the frequencies of a sound wave
    Not a Data mining task. It is just getting data. Can be considered signal processing.

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. A very simple and popular example is amazon

Clustering – Grouping of search results which are related, eg: if we search for a mobile, all related/similar items and relative purchases by others(recommendations) are shown for example phone accessories that were purchased, like case, tampered glass, pop socket etc. find high-, medium-, and low-spending customers from their transaction histories are clustered to help send promotional offers to low spending customers to increase their purchases

Classification - Predict if an item belongs to a category. Based on the products that are added, classification can be performed into their respective categories. A good or bad review can be classified through textual classification and sentiment analysis. multi-label tags can be provided for a product that falls into multiple categories. Based on regions of purchases, customers can be classified to understand trends/patterns in purchases.

Association Rule Mining – If the search pattern of the user is understood, it would be easier to give correct suggestions for their next search. Market basket analysis if the items in the basket of a user will be purchased or not can be predicted, what the user might buy next can also be predicted by association rules.

Anomaly detection – This helps in understanding any unusual patterns or events that are occurring. For example, if there are transactions are fraudulent, that user and potentially their IP can be noted for further fraudulent activities. If a product that has low sales and is relatively less known, is shown higher in searches, it is an anomaly. Suppose, there is a surge in purchase from one customer but also a lot of reports for missing packages are being placed, it shows abnormal behavior and the case can be further looked into if it is a fraudulent customer or if there is an issue with deliveries in the area the orders are being placed from.

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
    Census data contains a lot of important and personal information, which on falling into the wrong hands, could lead to huge problems., so in general privacy is important but considering the timeframe, old data is irrelevant for the most part. So, privacy is not very important
    b. IP addresses and visit times of Web users who visit your website
    From a user point of view, IP addresses and web activities being known is a privacy issue as one can fall prey of unnecessary marketing by certain companies. On the other hand, from a company's perspective, knowing information about IP addresses and visit times can help in enhancing the business through targeted marketing.
    c. Images from Earth-orbiting satellites
    Privacy is not really a concern here because there's not a lot that can be exploited from aerial images

d. Names and addresses of people from the telephone book
   Despite telephone books being outdated or less used now, it still is a privacy concern to have our information out accessible, as we can have random people call us or show up at our doorstep.
e. Names and email addresses collected from the Web
   From a user point of view, it is concerning, as we are not exactly aware of how our information is being stored and used. From a company's point of view, having names and email addresses would help in enhancing the business by sending out promotional emails to potential/loyal customers

Task 2: practice your critical thinking and writing

Read the following two news articles. One criticized Google Flu Trend, and the other defended it. Write one paragraph to summarize the criticism, and another paragraph for the defense. Write the third paragraph to offer your own thought, e.g. is the criticism valid? Does the defense make sense? What other problems or benefit do you see in Google Flu Trend or similar big data applications?

http://bits.blogs.nytimes.com/2014/03/28/google-flu-trends-the-limits-of-big-data/
http://www.theatlantic.com/technology/archive/2014/03/in-defense-of-google-flu-trends/359688/

Criticism:
Google flu trends, once the poster child for power of Big Data Analytics was under fire for widely overestimating the number of flu cases in the United States of America in 2012-13. Wrong/over estimation by 50% in the year 2011-12. So, for 2 years in a row, over-estimation was done. It showed poor accuracy in identifying trends. The google estimates were high in 100 out of 108 weeks. People were skeptical of Google's Flu trend algorithm, which was used for prediction, even after updates it overestimated by 30%. The comparative value of the algorithm as a stand-alone flu monitor was questionable. The 2 week lagged CDC reports were considered more accurate than google flu trends.

Defense:
40 flu related queries were used to predict the flu prevalence. Despite the over-estimation in cases, a combination of CDC monitoring and Google flu trends, proved to provide better results than either would individually. Google Flu Trend system was designed to supplement rather than replace established surveillance systems. Later, when a team was trying to figure out how to make a better influenza model, the google flu trend model came in handy. It served as a base model to develop anything that can be used in predictions.

In my opinion, the criticism was valid because of the failure/over-estimation. But given how GFT system was developed to aide the CDC monitoring system and not overtake it at that point, the criticism was harsh. On the brighter side, GFT laid a foundation for monitoring trends and searches. It showed how important it is to understand and view data in the right context. A lot of misinterpretations can be done with wrong analysis. But a lot of big data analysis and trend mining approaches were given importance after GFT system. When we deal with huge data, knowing the context and purpose behind it would help in avoiding any misanalysis or wrong interpretations. Using GFT as a reference, more sophisticated approaches have been used and designed to analyze and understand trends and patterns in different fields using big data.