IST 707 – Data Analytics/Machine Learning

Name: Akshay Chaurasia **Email: ahchaura@syr.edu**
SUID: 309898873 **Date: 09/02/2019**

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.

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.
➔ No

(b) Dividing the customers of a company according to their profitability.
➔ No

(c) Computing the total sales of the company.
➔ No

(d) Sorting a student database based on student identification number.
➔ No

(e) Predicting the outcomes of tossing a (fair) pair of dice.
➔ No

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

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

(h) Monitoring seismic waves for earthquake activities.
➔ Yes

(i) Extracting the frequencies of a sound wave.
➔ No

IST 707 – Data Analytics/Machine Learning

Name: Akshay Chaurasia **Email: ahchaura@syr.edu**
SUID: 309898873 **Date: 09/02/2019**

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.**

➔ First off, Data Mining is the process of collecting useful and previously unknown information which can be used to discover patterns in a dataset or to make future predictions. So, it can be a very useful technique to increase the customer base by analyzing the customer needs.

Internet search engine company: Google

Clustering: Clustering is a form of unsupervised learning that analyzes data objects without known class labels. It is a set of points that are more similar to each other as compared to the points in another cluster.

Suppose a user is looking for movie suggestions (say comedy movies) on google. They will simply search for it on Google and Google will display all the results.

Initially, the company (Google) kept all the movies in an open vector. But over the years, Google started collecting the details of different movies and then based on some similarities (keywords), it started adding the movie to a genre (horror, comedy, romance, action, etc.), forming a cluster.

Classification: Classification is a form of supervised learning that analyzes the data using predefined class labels.

Continuing with the previous example… Now, as soon as a new movie comes into the market, Google tests the data of this movie against the earlier defined labels (#romantic or #horror) and classify it accordingly.

Association Rule Mining: If a user is into romantic-comedy movies, Google can use the concept of Association rule mining to suggest them some romantic movies too.

Anomaly Detection: Google can check the payment history of user to detect abnormalities like making a huge payment online, payment made from different devices/locations or buying multiple movies at once.

IST 707 – Data Analytics/Machine Learning

Name: Akshay Chaurasia **Email: ahchaura@syr.edu**
SUID: 309898873 **Date: 09/02/2019**

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.
➔ No

(b) IP addresses and visit times of Web users who visit your Website.
➔ Yes

(c) mages from Earth-orbiting satellites.
➔ No

(d) Names and addresses of people from the telephone book.
➔ No

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

IST 707 – Data Analytics/Machine Learning

Name: Akshay Chaurasia **Email: ahchaura@syr.edu**
SUID: 309898873 **Date: 09/02/2019**

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:

During the onset of big data, Google released an experimental project called, "Google Flu Trends". Initially, Google Flu Trends was considered as a poster child for the power of big-data analysis. However, in the first half of 2014, Google Flu Trends was under the attack of the critics for "wildly overestimating" the number of flu cases in the United States in the 2012-2013 flu season. It continuously over-estimated the flu cases. In 2011-2012 flu season, Google Flu Trends' estimate was more than 50 percent higher than the cases reported by the Centers for Disease Control and Prevention. It was also reported that for a period of more than two years ending in September 2013, the Google estimates were high in 100 out of 108 weeks. Critics also blamed Google for over-estimating the Big Data Analysis by declaring that Google was guilty of "big data hubris". Critics even said that simply using the recent trend of C.D.C. reports from doctors on influenza-like illness would have been a more accurate predictor than Google Flu Trends, even though it lagged by two weeks. The basic idea of criticism was that the Google Flu Trends was not using a broader array of data analysis tools and that its algorithm was not accurate due to which the service overshot by about 30 percent.

Defense:

In their defense, the Google Flu Trends team stated that the service was always intended as a "complementary signal" rather than a stand-alone forecasting tool. They said that the system was not designed to replace the traditional surveillance networks or displace the laboratory-based diagnoses and surveillance. In fact, the goal behind Google Flu Trends was to build a complementary signal to other signals. In the 2007-2008 flu season, Google Flu Trends gave the indication of flu outbreak in advance. It gave near real-time signal and it had proven that it does add value. Few respected authors and academics also pointed that the Google Flu Trends is a proof of the triumph of the big data approach. Researchers both in and outside epidemiology have found Google Flu Trends and its methods useful and relevant.

IST 707 – Data Analytics/Machine Learning

Name: Akshay Chaurasia **Email: ahchaura@syr.edu**
SUID: 309898873 **Date: 09/02/2019**

My Opinion:

Considering the fact that, Google Flu Trends overestimated the cases, the criticism that it faced was valid. But the fact that, Google Flu Trends was always intended as a "complementary signal" rather than a stand-alone forecasting tool, backs the defense. According to me, this criticism was due to over-estimation of the newly arrived Big Data technology. Since, it was a new technology, people did not know its capabilities or how to use this technology. Also, since it was a product of Google, everyone had huge expectations form the Google Flu Trends. This led to the backlash from the people. The major problem with the Google Flu Trend or similar big data applications is that Big Data is an emerging technology. No one is fully aware of its capabilities or even how to use it. So, expecting a solution for a problem using something that we do not yet understand properly is a big problem.