IST 707: Applied Machine Learning

Name: Hrishikesh Mahesh Telang
SUID: 889489533

HW1

Task 1: Review data mining concepts and tasks

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 it is not, as we are just segregating customers on the basis of their gender. If we had predicted the gender of a customer based on its buying and spending habits, then it could have potentially been a data mining task.
    b. **Dividing the customers of a company according to their profitability**
    No, because this is an accounting calculation by comparing it with a threshold value to determine profitability.
    c. **Computing the total sales of a company**
    No, because this is mere computation of calculating the sum of total sales. There is no meaningful insight or analysis derived from the data.
    d. **Sorting a student database based on student identification numbers**
    No, because we are just sorting the data. We are not deriving meaningful insights behind that data.
    e. **Predicting the outcomes of tossing a (fair) pair of dice**
    No, because we are deriving the probabilistic distribution of tossing a fair pair of dice. It involves more of statistical computations than an actual prediction or classification of a feature vector.
    f. **Predicting the future stock price of a company using historical records**
    Yes, because we are performing predictive analytics to determine the value of an output using past records, which is one of the major definitions of data mining.
    g. **Monitoring the heart rate of a patient for abnormalities**
    Yes, because in this case, we are performing anomaly detection in which we are monitoring the heart rate and comparing it with the historical records of normal heart ranges. If there is an anomaly detected, an alarm/siren could be raised. This indeed is an example of a data mining task.
    h. **Monitoring seismic waves of earthquake activities**
    Yes, because in this case we are building a model to predict different classes of seismic wave behavior associated with earthquake activities. If an alarm is raised, we are certain that some form of earthquake has been observed. This is essentially what we term this as a classification problem.
    i. **Extracting the frequencies of a sound wave**
    No, since the action of extracting the frequencies associates more to a signal processing task than a data mining one.

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 such as Google was one of the pioneers of the applications of data mining in that they have been providing a free service in exchange for user's behavioral data based on the information they type. This vast data collected can be used to improve the search algorithm of the company, provide advertisements based on the keyword search, and to improve overall user experience. The different techniques are as follows:

   Clustering:

   In this technique, a process known as keyword identification is performed to analyze user behavior. Based on the semantic context, we group or cluster the different keywords. In this way, word clusters are made and are added over a period of time as more users perform searching operations. Also, based on the frequency of the words used in the search, the respective keywords tend to start looking bigger i.e.: more "important" than the rest of the words within that cluster. Finally, these word clusters are stored within the search engine database and the keywords are organically compared to identify the cluster it belongs to.

   Classification:

   Based on the action of clustering performed, the classification algorithm is incorporated in order to understand what the user is searching for. The keywords are matched with the different other keywords within the clusters. Each of these keywords are assigned a binary value. These binary values are passed as feature vectors to a classification algorithm to provide several search classifications such as "news", "weather", "sports", "TV", etc. This ensures in providing relevant search results.

   Association rule mining:

   In this technique, we compare the keywords entered in the search result with the historical record of searches made by the users, and provide the strongest correlation of items which may otherwise not have been explicitly typed in the search box. For example, when we type the word "hello" in the Google search box, it shows "hello kitty", "hello world", "hello bello" depending on the past searches made by users.

   Anomaly detection:

   In this technique, we try to identify whether the user is an artificial bot or not. This form of anomaly check is performed in order to prevent the search algorithm to predict wrong user behavior while at the same time protecting confidential information.

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 is a subjective concept. If the data is an aggregate measure to determine the demographics of the people, whether it's based on gender, age, religion, sex, political ideologies or race within that state/region/country, then it is not a violation of data privacy. However, if the census data contains confidential information such as the name, maiden name of a woman and age of each individual, then it surely breaches data privacy.

b. **IP addresses and visit times of Web users who visit your website**
Yes, this is a data privacy issue as hackers can use such confidential information to perform spoofing or impersonation or by changing the VPN using that IP address to grab unauthorized content.

c. **Images from Earth-orbiting satellites**
No, as these images are publicly available under a free license over the internet.

d. **Names and addresses of people from the telephone book**
No. They do not present privacy issues as the telephone book is a directory to allow users to access this public information, for example: name and address of an entrepreneur can be public as they are providing a product or a professional based service.

e. **Names and email addresses collected from the Web**
No, as the names and email addresses are information that is collected from social media especially in brochures or other marketing posts.

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

Response:

There was a point in time when the Google Flu Trends (GFT), which used to be a poster child for power of Big Data Analytics, was under criticism for widely overestimating the number of flu cases in the US. It had periodically overestimated the flu cases for two consecutive years by a whooping 50%. These estimates were high in 100 out of 108 weeks which clearly shows the inefficiency of the trends. However, these researchers recognize the value that big data can provide towards the domain of scientific research but is also limited to the issue of what was given to be termed as the "Big data hubris", in which it is assumed that big data are a substitute for traditional data collection and analysis rather than being a supplement to it. Due to this setback, a lot of users found the GFT algorithm questionable. Despite having updated the algorithm, the flu trend yet overestimated the cases by 30%. At the same time, the advent of the CDC reports from the doctors based on influenza-like illnesses which lag by two weeks was a more accurate predictor than GFT.

Criticism:

The creators of GFT never intended for it to replace traditional surveillance networks such as the CDC lagged data reports, rather they meant for it to be used as a complementary signal to other signals. The author of this article defends GFT and its creators by calling upon its readers to go past the GFT reports, that were inaccurate in its predictions, as the authors of "The Parable of Google Flu: Traps in Big Data Analysis", have highlighted, and see the original intent behind the GFT algorithms. As readers, we are also reminded, that in 2008 when the GFT algorithms were first introduced, big data was hardly talked about nor a thing yet. The author shows us how the GFT creators were consciously aware of the need to work closely with the CDC as they designed this algorithm. The tool was to be used by them and therefor needed their input. It was this collaboration that enabled the GFT to be used as a complementary data set in the Johns Hopkins research, of 2013, which lead to a better influenzas' prediction model, of all data sources, showing the only statistically significant forecast improvements over the base model. In that respect, the author suggests that GFT wasn't a frailer, according to the standards laid out in the Nature paper describing it in 2009. But more so a failure of populous imagination of Big Data algorithms and new technology perceived as magic.

My opinion:

It is valid to say that GFT has overestimated the cases that caused its failure. However, when we look at the broader perspective, GFT has laid a foundation for Big Data Analytics. The defense also makes sense. Tracking 45 flu related terms over billions of searches, monitoring trends and, making correlations was a huge win for a big data approach. GFT has set an example for further big data analysis. The system served its purpose of complimenting traditional surveillance system. The problems like over-fitting, finding the right signal in the noise, data integration, speed and scalability are found in big data applications. Whereas, big data applications also set base for future generation, help identifying business levers and measuring priorities.