IST 707 Week – I Assignment

Thulasiram Ruppa Krishnan

**Task 1: review data mining concepts and tasks**

1. Discuss whether or not each of the following activities is a data mining task.
2. Dividing the customers of a company according to their gender.

This task is just grouping a dataset based on gender variable. Gender is a predefined variable with low cardinality and in most of the case it takes two values (male or female). This can be accomplished by writing a simple SQL with where clause that divides the dataset into either male/female, if the source is a relational database system. If it is a file system like csv or excel, a simple filter can be applied using excel functionality to divide the data set.

In case of very large dataset, you can use appropriate tool and use the filter technique to divide the dataset based on gender. No complex algorithm or any data mining technique is required to do this exercise because there is no hidden knowledge that we are trying to uncover here.

1. Dividing the customers of a company according to their profitability.

This seems to be a simple categorization based on the profitability of the customer. To decide whether the customer is profitable or not, we need to look at the cost associated for servicing a customer and the revenue incurred by the customer. By finding the above, one can classify if the customer is profitable or not. No data mining technique is required as they are classified based on a simple output of a variable like the above gender classification. There might be some processing and transformation is involved in finding out the cost and revenue, but these are not hidden knowledge which requires exploratory data analysis.

1. Computing the total sales of a company

The total sales of a company is the information that can be calculated by applying aggregation technique of all the available data values. This can be referred as information retrieval problem and not necessarily a data mining task. If the ask is to predict the company sales for next year, then this will become a data mining task.

1. Sorting a student database based on student identification numbers.

Sorting is not a datamining task. It is an orderly information retrieval. No hidden knowledge or prediction is involved in doing this exercise.

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

This is a data mining task. Here, the goal is to predict the future occurrence of outcomes by tossing a pair of dice. This requires applying statistical technique along with the previous outcomes to create a model which can predict the outcome of the next occurrence.

1. Predicting the future stock price of a company using historical records.

This is again a data mining task. There are several variables that will contribute in predicting the stock price of a company. This requires exploratory data analysis to find out the variables and its correlation to stock price. Also, it requires to create a model using all the independent variable to predict the dependent variable which is stock price of the company.

1. Monitoring the heart rate of a patient for abnormalities.

Detecting abnormalities is a data mining to task. This requires classifying the input data from a monitoring device as normal and abnormal. It involves analyzing various parameters and their behavior with respect to each patient and the historical heart rate records to decide whether the new heart rate is normal or abnormal

1. Monitoring seismic waves for earthquake activities.

Monitoring waves for earthquake activities is a data mining task. There are numerous parameters that contributes an earthquake. In order to find the interaction between these variables to predict earthquake is challenging and need exploratory analysis.

1. Extracting the frequencies of a sound wave.

This is related to information retrieval and not a data mining acitivity, extracting frequency from a sound wave is a predefined context and may require some series of step to extract the information.

1. 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.

Here is some example where clustering, classification, association rule mining and anomaly detection can be applied in healthcare.

**Classification:**

Given the blood sample categorize patients as diabetic and non-diabetic is a classification problem.

**Clustering:**

Given the patients’ medical history and diagnosis information, determining high and low risk patients for a specific disease is where clustering technique can be applied. We do not have predefined labels for medical history and diagnosis to determine the risk factor and hence clustering technique should be used

**Association rule mining**

This technique can be used for identifying strong association between drug ingredients and the allergies developed by the patients. Using information of variety of drugs that are administered to patients and the number of cases that developed allergies, association between the ingredients on the drug and the allergies can be tied together.

**Anomaly detection**

Using this technique, a cancer cell which is different from normal cells can be detected and analyzed. Also, any gene mutation can be revealed by comparing against the normal gene dataset.

1. For each of the following data sets, explain whether or not data privacy is an important issue.
2. Census data collected from 1900-1950.

Data privacy is not an importance issue with Census data as it deals with numbers and aggregated information. There is no personal information included here which makes this type of data as low risks.

1. IP addresses and visit times of Web users who visit your Website.

Data privacy is not an important issue for IP address and visit times, it is not possible to target a person with this information and hence it is a low risk for data privacy

1. Images from Earth-orbiting satellites.

Images from Earth-orbiting satellites needs to be protected as these images with very high resolution can reveal an individual’s home, car and other properties which can be monitored and followed closely. Also, for a country it is important to protect army bases and its activities during war.

1. Names and addresses of people from the telephone book.

Name and addresses of individuals need to be protected otherwise it can be misused by criminals. Data privacy is an important issue here.

1. Names and email addresses collected from the Web.

Name and email address are also important for an individual that need to be protected otherwise it can be misused by criminals. Data privacy is an important issue here.

**Task 2: practice your critical thinking and writing**

1. **Summary of the article “Google Flu Trends: The Limits of Big Data”**

Criticism is well-known and most welcomed for any new technology that will challenge or treated as a threat to the existing traditional approach. In this context, Google the so-called “Big Data Pioneer” is criticized heavily for its work on flu trend prediction using big data. An important issue discussed here is about the accuracy of Google’s flu trend when compared with CDC’s (Centers for Disease Control and Prevention) flu trend report. Google’s forecast is blamed for its overestimation (high in 100 out of 108 cases), misses in September 2013 flu season, more than 50% variance when compared with CDC’s report in 2011-12 flu season and overshot by 30% in 2013-14 flu season. As a note from experts and professors from social science, usage of a broader array of data analysis tools will improve the algorithmic smartness. Data mashup with CDC will result in accurate predictions. In the view of these experts, Google’s prediction can be used as a parable notion for flu trends and the skepticism for using it as a stand-alone flu monitor system is highly considerable.

1. **Summary of the article “In Defense of Google Flu Trends (GFT)”**

Despite many criticisms, the article “In Defense of Google Flu Trends (GFT)” shows Google’s flu trend is welcomed by many including chief of the epidemiology and prevention branch in the CDC's influenza division. This new tool is celebrated by media and wall street journals for its approach in building a more real-time solution by using big data (user’s search engine queries). Google flu trend fills the gap in lagged CDC data thereby adding greater value for reporting near real-time flu trends. In an attempt made in 2013 to predict and build a better influenza model by using emergency-room clinical data along with GFT meteorological data (temperature, change in temperature, and relative humidity) and temporal variables resulted in statistically significant forecast improvements by just using GFT. CDC is not available in every part of the world, and GFT gives more value in predicting disease outbreak for a place like Africa and Cambodia as explained by Larry Brilliant. Even though it created disappointment in view of public opinion, researchers both in and outside epidemiology have found Google Flu Trends and its methods useful and relevant.

1. **Thoughts on Google Flu Trend and big data applications**

It is very common that every technological initiative will have a different perspective from people who get benefit and people who are affected either directly or indirectly. In the realm of big data, an effort created by Google which is “Google Flu Trend (GFT)” has got more criticism than its potential benefits. As a result, GFT is projected as not useful in the view of the public. Defining the purpose and usefulness of GFT is way more important than the actual effort spent on building this tool. GFT is not designed to be a replacement for traditional surveillance networks or supplant the need for laboratory-based diagnoses and surveillance which are currently done by the Center for Disease Control and Prevention (CDC). It is a low-cost technique for building a framework when compared with the lagged traditional approach on performing test and capture results.

GFT is not a quantitative science to represent accurate numbers of people affected by the flu. It is an alarming system at real-time to track the flu trend which can proactively alert the public and healthcare system to get prepared for the next wave of epidemics and outbreaks. It is also free from any human bias and hypothesis as they are an accurate representation of collective intelligence. When it comes to the traditional approach, we are also not sure whether it is an accurate description of people infected by the flu. CDC report numbers on people who established a connection with the healthcare system to get treated and this number could vary from the actual count of infected people who does not reach out to any healthcare system. When it comes to Google’s approach in handling this problem using query searches has resulted in overestimation because people who search on flu and flu-related symptoms are not actually who are infected by the same. We have pros and cons in both the approach and one is no better than other as it has its value on a different context. Countries who do not have a system like CDC can leverage the usefulness of GFT and get benefitted. It will be a better approach if GFT can show the trend on a scale of 1 to 10 instead of sticking with real numbers. This solution will gain more attraction and focus on trends instead of comparing the actual number of flu cases. All big data applications will encounter a similar issue and by carefully designing the solution which does not give more details to the public especially if there is a question on validity. Both criticism and defense help us to understand the public reaction, and it will help us to build a better solution by keeping in mind its limitations and usefulness.