IST 707

Hana Kim

HW 1

Task 1.

1.7

1.

a) Not a data mining task

b) A data mining task

c) Not a data mining task

d) Not a data mining task

e) Not a data mining task

f) A data mining task

g) Not a data mining task

h) Not a data mining task

i) A data mining task

2. When a security clearance company conducts background check for employee clearance determination, a data mining technique such as clustering, classification, association rule mining, and anomaly detection can be applied. Clustering technique can be applied in such way that employees are grouped together based on previous records of previous employment, length of employment, residence locations, previous and current positions, foreign places visited, etc. Classification example can be measurement applied to criteria such as less than one-year employment low, one to three years employment medium, more than three years high. Association rule mining example can be that a strong relation is found in that employees with many records of less than one-year employment tend to leave current company in less than two years. Anomaly detection example can be excessive financial status or excessive or abnormal foreign places visited.

3.

a) Not important data privacy issue because Census data from 1900 – 1950 is a publicly available information that does not pertain to any specific individual

b) This can be an important data privacy issue because it traces personal footprint of individual and information can be induced.

c) This can be an important issue if the satellite images can clearly identify a person or a car with location and time.

d) Even though names and addresses of people are already exposed in the telephone book, it can be an important data privacy issue depending on how the information is used

e) Same as the telephone book, depending on the use of this information, it can pose an important data privacy issue.

Task 2.

The first article, Google Flu Trends: The Limits of Big Data criticizes Google Flu Trends in that the model greatly misses the prediction by over 50 percent higher than the cases reported. It also criticizes that the prediction model of Google Flu Trends continuously over-estimated flu cases 100 out of 108 weeks. The author points out the danger of blindly looking into big data and drawing conclusions from big data only.

The second article, In Defense of Google Flu Trends, the author defenses Google Flu Trends by taking a position that the fault is in people’s high expectation that the model will forecast everything by itself. The author points out that the intention of Google Flu Trends is not to make an accurate future prediction but to assist and supplement professional organization with prediction and it does well of prediction when two data sets are meshed-up.

In my opinion, the criticism of Google Flu Trends should have been written to point out the public’s hyped opinion of blindly trusting big data that it can make all prediction correctly by itself. I think the defense of Google Flu Trends article points out the intention of the tool and the use of big data prediction well. That is, the model is not intended to be used to make a prediction solely by itself, but when combined with other available data, it can enhance the outcome more accurately. In this global age where anywhere in the world can be physically reached, the use of tool such as Google Flu Trends is a necessary tool to monitor and predict global epidemics that can reach anywhere and anytime. In fact, I have witnessed the use case of prediction model such as Google Flu Trends when MERS-coronavirus, aka ‘camel flu’ hit South Korea and almost brought the economy down to its knees in 2015. Only after the effect, the government and public health community started closely track and monitor MERS flu cases in the middle east region. So it is unavoidable to adapt and use a prediction model from big data such as Google Flu Trends.