2019-0703 IST 707 Data Analytics

Homework Assignment 1 (week 1)

Ryan Timbrook
NetID: RTIMBROO
Course: IST 707 Data Analytics
Term: Summer, 2019

Table of Contents

1 Introduction

1.1 Task 1: review data mining concepts and tasks

1.1.1

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.

1.2 Task 2: practice your critical thinking and writing

1.2.1

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/

2 Task 1: Exercise Questions 1 - 3

2.1 Question 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.
   a) No, this is a trivial grouping problem. It could be a subtask performed in the data preprocessing phase of KDD, Data Subsetting. Has no decision-making output.

b) Dividing the customers of a company according to their profitability.
   a) No, this is a trivial grouping problem. It could be a subtask performed in the data preprocessing phase of KDD, Data Subsetting. Has no decision-making output.

c) Computing the total sales of a company.
   a) No, this is a trivial summation task. It could be a subtask performed in the Postprocessing phase of KDD. Has no decision-making output.

d) Sorting a student database based on student identification numbers.
   a) No, this is a trivial function task. Has no decision-making output.

e) Predicting the outcomes of tossing a (fair) pair of dice.
   a) No, if it is a fair pair of dice (6,6), each role of the dice has the same 36 distinct possible outcomes.

f) Predicting the future stock price of a company using historical records.
   a) Yes, this is a regression, Predictive task.

g) Monitoring the heart rate of a patient for abnormalities.
   a) Yes, if the task implies building the software used in the heart rate monitoring system, building the model (i.e., real-time data/processing required for the monitoring). Then this is an anomalies detection, Descriptive task. Otherwise no, since this requires real-time data.

h) Monitoring seismic waves for earthquake activities.
   a) Yes, if the task implies building the software used in the seismic wave sensor system, building the model (i.e., real-time data/processing required for the monitoring). Then this is a classification, Predictive task. Otherwise no, since this requires real-time data.

i) Extracting the frequencies of sound wave.
   a) No, this is a trivial task with no decision-making output. It could be a subtask performed in the data preprocessing phase of KDD, Data Subsetting.

2.2 Question 2:

Suppose that you are employed as a data mining consultant for a Telecommunication Company. Describe how data mining can help that company by giving specific examples of how techniques, such as clustering, classification, association rule mining, and anomaly detection can be applied.

Telecommunication companies maintain huge volumes of data about their customers and their call details. This information is used to profile customers which is then used for marketing and forecasting purposes.

Customer Churn is a major problem and for most companies it is their most important concern. Finding factors that increase customer churn is a necessary action to reduce churn. Data mining techniques, such as clustering and classification, are used to model customer life-time-value which estimate's how long he/she will remain with their current network. Using feature engineering and predictive classification techniques such as Decision Tree, Random Forest, or Gradient Boosted Machine Tree, the company can identify potential churn customers. Applying association rule mining techniques to these new profiles, the company can then provide targeted marketing campaigns engineered to retain the customer.

Fraud Detection is a very serious issue with any company but has become more of an issue in recent months for this telco company. Specifically, superimposition fraud, which occurs when a perpetrator gains illegal access to the account of a legitimate customer has been on the rise. Data mining techniques, such as Anomaly detection can help detect the fraudulent caller by predicting a relatively rare event where the class distributions involved is highly twisted.

2.3 Question 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.

   a. No, the census bureau has policies to ensure the data collected is protected.

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

   a. Yes, this is personal profiling information that could be used maliciously.

c) Images from Earth-orbiting satellites.

   a. No, this is public data and has no connection with peoples privacy.

d) Names and addresses of people from the telephone book.

   a. No, the telephone book is a publicly distributed data source.

e) Names and email addresses collected from the Web.

a. No, this is public data, however in conjunction with other data mining and social engineering (hacking) techniques, it could be used maliciously.

3 Task 2: Critical Thinking & Writing

Google Flu Trend

Criticized:
http://bits.blogs.nytimes.com/2014/03/28/google-flu-trends-the-limits-of-big-data/

Defended:
http://www.theatlantic.com/technology/archive/2014/03/in-defense-of-google-flu-trends/359688/

Summary of: **"The Parable of Google Flu: Traps in Big Data Analysis**", by David Lazer, Ryan Kennedy, Gary King, Alessandro Vespignani

Four researchers found that Google Flue Trends (GFT), which in their opinion is recognized as an exemplary use of big data, grossly miss lead the public with its highly elevated predictions of people's doctor visits for influenza-like illness (ILI). GFT overestimated the prevalence of flu in 2012-2013 session and overshot the actual level in 2011-2012 by more than 50%. And from August 2011 to September 2013, it overshot flu prevalence 100 out of 108 weeks. These researchers recognize and believe in the tremendous value-add big data can have on scientific research, however, they caution their readers of "Big data hubris", where it's implicitly assumed that big data are a substitute for, rather than a supplement to, traditional data collection and analysis.

The authors don't completely discredit the value of GFT. They show that by combining the CDC lagged data and GFT, as well as dynamically recalibrating GFT, they can substantially improve on the performance of GFT or the CDC than alone. Rather they believe the issue is twofold. First, Big Data Hubris (BDH) is at play. And secondly, Google is undermining its own work on the GFT algorithms as it modifies its search product capabilities that have evolved to include advancements in Natural Language Technology, predictive text suggestions (i.e., recommended searches and automated answers to questions) that improve business and provide more relevant results. Whereas GFT, on the other hand, assumes only external events-like more flu virus in the world - are affecting users searches.

A key point these researchers address is the use of the GFT parable as an important case study for learning a critical lesson as big data analysis evolves. The authors ask Google to be more transparent with the research community about how it analyzes its data, taking into privacy concerns into consideration. This would then allow a more empirical approach with other experts building on and learning from their work.

Summary of: **"In Defense of Google Flu Trends"**, by Alexis C. Madrigal Mar 27, 2014

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 it's 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 thoughts:

Taken at face value, focusing exclusively on the stand-alone use of GFT as a predictor of influenza outbreaks and the need for companies such as Google to be more transparent with their data sets and methodologies, the author's of "The Parable of Google Flu: Traps in Big Data Analysis" have valid critical arguments that bring to light core challenge's not only in Big Data, but in how we as consumers of the information need to view it with one-eye-open.

As well, it appears to be subjectively motivated covering one side of the data. On the surface it reads as a major blunder on Googles part and the engineer's who designed the system. Taking into consideration the year, 2008, when the GFT was first published, the algorithms would have been much simpler than today. The authors also never mention the GFT creators work with the CDC in the early design and creation of the algorithms and how they intentionally modeled the algorithm based on their input and need to keep the data sets separate.

Pealing back the layers, the author of "In Defense of Google Flu Trends", does a good job highlighting the original intent of GFT, as stated by its creators, that it was never intended to be a replacement of the CDC reports. Understanding how new technology should be used and living within reasonable expectations of it is at the core of this defense, which makes perfect sense. Technology is not magic, however much it may feel that way, and like all things, there is a natural order to its evolution and how it should be used.

Other considerations when dealing with big data pit falls include understanding the nature of the data. If we're using the collective intelligence of social media channels to predict population wide human patterns, we must recognize how challenging human behavior is. People's habits shift regularly, often without recognizable or published reasons. We are a chaotic, irrational species where social media will only capture a fraction of the populations thoughts and when captured it shouldn't be taken as solely as original thoughts. Group influence plays to heavily in how people react where more behavior psychology understand need to be incorporated into how algorithms are tuned.

4 References

1. *The Parable of Google Flu: Traps in Big Data Analysis, by: David Lazer, Ryan Kennedy, Gary King, Alessandro Vespignani,*

2. *In Defense of Google Flu Trends, by: Alexis C. Madrigal, Mar 27th 2014*