Matthew L. Pergolski

IST 707

1/12/2022

Homework 1

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.
2. Dividing the customers of a company according to their gender.

**Within the process of Knowledge Discovery in Databases (KDD), as referred to in Figure 1.1 of the book, a task such as this would most likely fit into the ‘Data Preprocessing’ section. More specifically, we may associate this task with ‘Data Subsetting.’**

**Because of this observation, this task would not be one that directly relates to Data Mining. Instead, it would fit into Data Preprocessing. One aspect of the preprocessing step is to select records and/or features relevant to the data task at hand, which could be the case in this instance.**

Diagram

Description automatically generated

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

**Like the first task, this would also not be directly associated to Data Mining – instead, we see that it would also fit under the ‘Data Preprocessing’ header. The grouping could be relevant to a data mining task that comes after this stage.**

1. Computing the total sales of a company.

**Similarly to the previous tasks, computing the total sales of a company would also not be directly related to Data Mining. Instead, this would be considered a general calculation from financial documents such as a Balance Sheet – which may correspond as a feature or record being selected for the upcoming data mining task at hand.**

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

**Sorting would not equate to a task of Data Mining; it is a general preprocessing technique that may serve useful for the upcoming Data Mining task at hand.**

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

**Although this is prediction, the category for this type of task (i.e., fair dice roll prediction) would more closely relate to probability and statistics versus true Data Mining.**

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

**Predicting the future stock price of a company would be considered a Data Mining task. Regression is a term that, different from Classification (i.e., use of discrete target variables), is used to determine continuous target variables.**

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

**Abnormality detection would be considered a data mining task.**

1. Monitoring seismic waves for earthquake activities.

**Monitoring seismic waves for earthquake activities would be considered forecasting (i.e., predicting) and would thus be a data mining task.**

1. Extracting the frequencies of a sound wave.

**Extracting frequencies would most likely be considered a preprocessing task and would thus not be considered true ‘Data Mining.’**

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.

**In this example, I will replace ‘Internet Search Company’ for a company I work for in the defense industry: Lockheed Martin.**

**Within Lockheed Martin, the following techniques are relevant and significant throughout the company for data miners, data engineers, data scientists, etc.: Clustering, Classification, Association Rule Mining, Anomaly Detection.**

1. **Clustering**
   1. **The technique of ‘clustering’ can be used by Lockheed Martin, and others in the defense industry, to group together similar items that are used for various sub-assemblies. For instance, engineering documents for particular assemblies can be grouped together based on the part numbers they use. All subassemblies that use a particular resistor or capacitor may closely relate to an effort where the assembly will be used for ‘Development’ activities with the internal engineering team or ‘Production’ efforts that will result in eventual shipment to the customer.**
2. **Classification**
   1. **An example of classification within the Lockheed Martin sphere is involvement with determining build locations based on component. Furthermore, any item requiring ‘reeling’ or fixed within a ‘matrix tray’ would considered an Ocala, FL build location, while any wire or epoxy usage would be classified as both an Orlando, FL and Troy, AL item. By collecting various Bill of Materials dataset, data miners could utilize classification to determine which build locations will be needed for a particular end-item purchased by the customer.**
3. **Association Rule Mining**
   1. **For Association Rule Mining, the defense industry would be very interested in learning more about how this topic could help catch potential product failures before they occur. For instance, within a flight test for a particular asset/item, specific reading measurements could strongly correlate or associate with a particular type of failure down the line. Particular components or suppliers that infrequently provide parts could be associated with specific quality issues. This would provide the LM quality team a chance to review, in more detail, possible root causes for why this item or supplier could be a specific risk to the mission.**
4. **Anomaly Detection**
   1. **Anomaly detection could very well be the most important aspect of the defense industry – at least for Lockheed Martin. Data Miners that are involved in anomaly detection in terms of quality of product are responsible for ensuring the product shipped to the customer will perform as intended, each time, and without failure. Low quality standards of these items could result in serious injury or death to the user (i.e., those in armed forces), or a failed mission. Detecting anomalies in, for example, a missile detection system is important to ensure the safety of those using it. Data from a series of tests before shipment to the customer could allow data miners to ensure the product is performing to the Lockheed Martin standard (i.e., similar shipped units that were tested in the past).**

3. For each of the following data sets, explain whether or not data privacy is an

important issue.

* 1. Census data collected from 1900-1950.
     1. **Although census data is public information, it is supposed to be anonymous; therefore, consequently, data privacy would be an issue when collecting information regarding the national census**
  2. IP addresses and visit times of Web users who visit your Website.
     1. **Data privacy would also be a concern here. There are an unfathomable amount of hackers lurking on the internet that are interested in obtaining your information. Google also utilizes information to potentially sell it to 3rd party services for their monetary gain.**
  3. Images from Earth-orbiting satellites.
     1. **Whenever one utilizes Google Earth and ‘zooms in’ on a particular street, it is noticeable that all faces are blurred. It’s important to hide people’s identify in these instances where consent is not formally given. Although ‘data privacy’ may not be an argued issue – the general privacy of people’s identity through technology would be an issue here. Data can assist in automatically detecting and blurry faces, lessening the manual burden on any given worker or employee attempting to perform the task.**
  4. Names and addresses of people from the telephone book.
     1. **Generally, phone numbers and addresses are public information.**
  5. Names and email addresses collected from the Web.
     1. **Activity related to collecting names and email addresses is important to data privacy. A substantial number of phishing/hacking takes place with email activity. Apple’s recent iOS update introduced a private relay system to ‘Hide My Email’ from websites to prevent tracking and phishing/scamming. If spam is sent to the relayed email address, it can be deactivated, allowing the user to no longer be susceptible to these risks.**

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/

**Within the critical Big Data article by Steve Lohr, a few resounding statistics were provided to convey overall accuracy. The article states, “Google Flu Trends’ estimate for the 2011-12 flu season was more than 50 percent higher than the cases reported by the Centers for Disease Control and Prevention. And, they wrote, for a period of more than two years ending in September 2013, the Google estimates were high in 100 out of 108 weeks.” Needless to say, the numbers are not accurate when comparing to the data provided by the CDC – even if Google’s is about two weeks sooner (i.e., close to real-time). Towards the end, the article also provides insight into how the tool was never meant to replace CDC data reporting or any other system, but rather serve as a complementary tool – adding value in providing warnings of potential spikes in real time (even if the numbers are ‘rough’ estimates). Overall, the article suggests hesitation over adoption of ‘big data’ techniques and appears to prefer ‘conventional methods’ that are conducted by the CDC – despite ‘slower’ results.**

**While the ‘criticism’ article takes (major) note of Google’s inaccurate limitations of their Flu Trends algorithm, the ‘defense’ article frames the situation in a way such that the effort is ongoing and a work in progress – collaborating with outside sources, such as the CDC, to improve and shape the tool into a more accurate reporting resource. While both articles mentioned the “big data hubris” as first mentioned in the ‘The Parable of Google Flue: Traps in Big Data Analysis’ paper, the ‘defense’ article doubles down on the overall goals of the tool. Instead of an all-out replacement, the Google product is meant to supplement the CDC’s traditional methods of generating flu data.**

**To summarize a different perspective on the tool, a direct quote from The Atlantic says it best: “The goal was to build a complementary signal to other signals.” This can suggest that, to address an overall issue or problem, one tool or method may not be the ‘end all, be all.’ Instead, it may take a collaborative approach, combining multiple ideologies, methods, and models, to turn data into information that ultimately assists in the process of making important decisions. Many things in life are how they are framed. If open-mindedness can stay alive in well in society, a more collaborative approach to finding solutions to problems will come easier – and faster – in time.**

**All in all, Google should continue to work with 3rd party services – and the CDC – to keep improving the predictive models. Concurrently, most of the ‘outside’ world have should open minds and attempt to understand the value of the current iteration of Google’s product offering. The defense of the article is keen in pointing out its value through real-time data and general spike signals, even if the detailed data values are off. The main, overall problem seen with Big Data – and future technologies – is public perception and acceptance. Humans tend to be very ‘stuck in their own ways,’ and may not be open to change/evolvement; this is especially true if initial results do not ‘go as planned.’ Open-mindedness and a motivation to improve will prove to be the strategy needed to innovative as quickly and accurately as possible.**

**References**

**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/**