CS 43105 Data Mining Techniques

Homework 1

Instructor: Xiang Lian
Due Date: September 18th, 2019

1. Based on your understanding, please explain what the data mining is. [10 points]

Data Mining is the process of sifting through large data sets and attempting to extract some kind of new and previously hidden, information. The information that is obtained should not be trivial or useless, instead it should help provide insight into the topic/subject area and help in problems such as decision making and categorization.

2. Please list 3 real applications for the data mining. Explain the data mining problems and solutions in these applications (give citations in reference format, if you use any online resources) [30 points]

- Medical Image Classification - The problem here is to more effectively detect tumors in the body. The solution is to use data mining to find associations that can be used to help train neural networks. For instance, in the paper Antonie et al. describe using the Apriori algorithm to discover association rules that allowed them to categorize each image [1]. This makes the issue a classification problem.

- Financial Fraud Detection - The problem here is to effectively detect instances of financial fraud. The solution is to use data mining on financial transaction data and try to pull associations from the data to more accurately detect anomalies that may lead to financial issue. In the paper by Ngai et al. they talk about the current state of data mining in the field of Fraud Detection. They show that many techniques including classification, clustering, and regression are used. The most common technique they found was the use of a logistics model [2].

- Customer Relationship Management (CRM) - The problem here is finding patterns of effective relationship management. The solution is to mine through data and find correlations between good relationship management and particular business practices. This is an association rule discovery problem. In the paper by Ngai et al., they describe the current state of data mining in the field of CRM. They show that many techniques, including classification, clustering, and sequence discovery, have been used in the past to classify and data mine CRM details. In this paper the found that articles on the application of data mining between 2000 – 2006, most commonly used neural networks as their technique [3].

3. Please discuss whether or not the following problems are data mining tasks. Explain why. [30 points]

3(a). Retrieve students' records from a relational table with grade = "A". [5 points]

No. This is a trivial task and does not yield new knowledge.

3(b). From the table of students' information, check if attributes last name and address have any correlations. [5 points]

Yes. We are gleaming new information by finding potential correlations between a person's last name and their address.

3(c). Find all the documents from the text database containing keywords "data mining". [5 points]

No. This is again a trivial task, a simple SQL statement. We are not finding new knowledge we are simple getting a subset of the data.

3(d). Divide the text database into several groups, each group containing near-duplicate or similar documents. [5 points]

Yes. This is a clustering task. We are potentially finding new information by clustering objects together using a particular attribute and then observing the other attributes of clustered objects.

3(e). Based on historical stock data, as well as other attributes (e.g., gold price, gas price, etc.) for the past few days, predict the trend of a stock tomorrow. [5 points]

Yes. This is a predictive task, specifically regression. The goal is to determine a potential pattern in data to help predict future events.

3(f). Please provide your own example of the data mining. [5 points]

The Disney MagicBand collects a lot of information that could contain interesting correlations. For instances, with the bands we could know what park people go to what rides they ride where they eat, when they leave, when they get back to the hotel. We could then match this with anonymized Guest information like what hotel there at, how big their party is, what tickets they bought and where they travel from. All this data could be used in Association Rule Discovery. There are many possibilities of associations.

4. Please choose a specific data mining problem (e.g., classification, clustering, regression, etc.), and discuss the corresponding solutions to this data mining problem. You can search for these problems via Google from the Internet (e.g., Wikipedia, Web pages, research papers, etc.), and explain the problem definition and solutions. [30 points]

Clustering is the idea of taking a set of data points that all have attributes associated with them and grouping based upon similar attributes. This will let us possibly gleam new information by looking at the clustered data points and looking at attributes they don't have in common and deciding if they may also have these attributes. An example of this problem might be medical image processing. For instance, you could have patient condition details associated with their images. You could then run through all the images and cluster them on the similarity of the images. Finally, you could go through the clusters and find associations between non-image attributes thus giving you new information about correlations between clustered objects.

REFERENCES

[1] Antonie, M., Zaiane, O. and Coman, A. (2001). Application of Data Mining Techniques for Medical Image Classification. In: Second International Workshop on Multimedia Data Mining. Springer-Verlag Berlin, Heidelberg.
[2] Ngai, E., Hu, Y., Wong, Y., Chen, Y. and Sun, X. (2011). The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. Decision Support Systems, 50(3), pp.559-569.
[3] Ngai, E., Xiu, L. and Chau, D. (2009). Application of data mining techniques in customer relationship management: A literature review and classification. Expert Systems with Applications, 36(2), pp.2592-2602.