COMP4433 Data Mining and Data Warehousing

Final Assessment: Part II (Due 11:59pm, 17 December 2021)

Instructions:
- Answer all questions.
- **Interpret the questions logically, show your steps and write down your assumption(s) when necessary**.
- Please submit your answer to L@PU before the due date.
- Late Submission Policy
  - 3-hour "grace period" is given.
  - After that, no late submission
- Plagiarism Policy
  - Both giver and receiver subject to the same penalty below
  - All the students involved not only will receive 0 marks for this assessment, but also will have an additional 50% penalty applied, e.g., 5 marks for a 10-mark assessment.

Time Series Data Mining and Data Warehousing (covering Q.1-Q.3)

Given the following plot of six stocks' closing prices covering the period from Jan 2000 to Jan 2007. They correspond to stocks with code 001, 011, 293, 857, 13 and 23. The real data can be found from Part I of the final assessment.

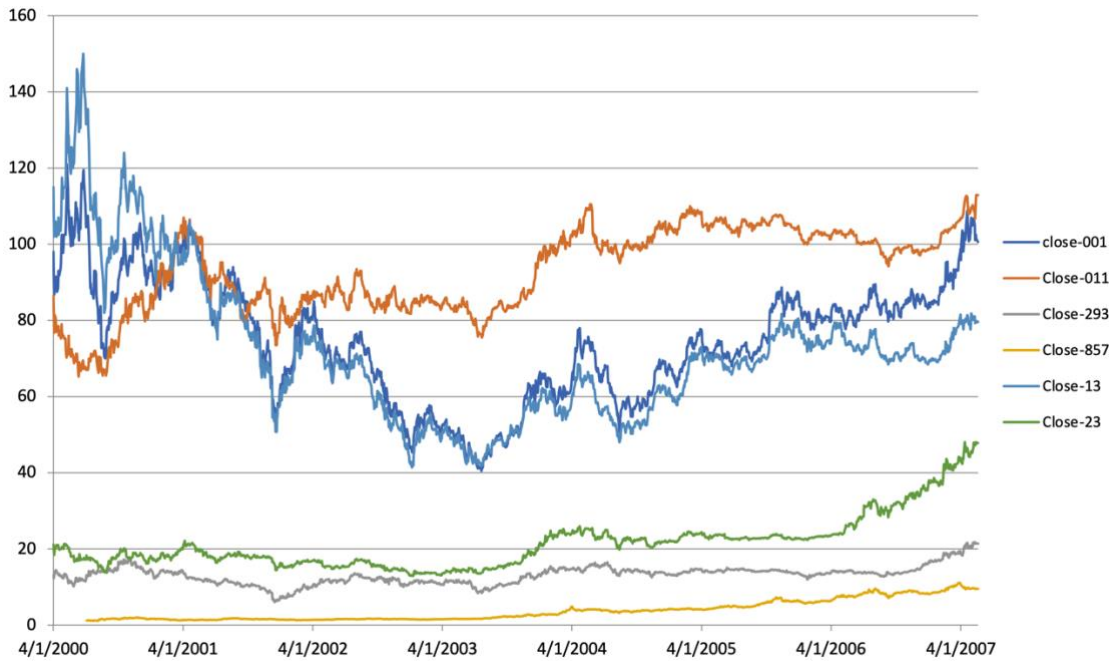

Fig.1 Closing price time series plot of six HK stocks

1. [25 marks] Based on the given stock time series plot, what data preprocessing tasks should be carried out for possible data mining attempts. Elaborate your answers with appropriate examples taken from Fig.1 for the following scenarios:

   a) Missing data handling for predicting stock price trend of stock 857
   b) Missing data handling for non-sequential association rule mining from all six stocks
   c) Noisy data handling for predicting stock price trend of stock 857
   d) Noisy data handling for non-sequential association rule mining from all six stocks
   e) Normalization of six stocks' data for similarity measures

2. [15 marks] Assuming that the given stock data in Fig.1 have been properly preprocessed, think about how data warehousing can be used to obtain useful analytical knowledge. Describe your solution by
   - giving 1 hypothetical dimension table
   - giving 1 fact table containing 1 hypothetical measure
   - listing 1 hypothetical OLAP result.

   This is an open question and you can refer to other stocks and/or other stock data related information to answer this question.

3. [20 marks] In view of the temporal nature of stock time series, you are asked to apply **sequential** association rule mining to discover sequential rules/patterns. By referring to the data in Fig.1, describe how you formulate the problem by
   - defining what a "customer" is
   - defining what a "transaction" is
   - defining what an "item" is
   - listing a few hypothetical data records for **sequential** association rule mining;
   - outlining some hypothetical mining results and discussing how they can be used in practice.

Algorithmic Development for Data Mining (Covering Q.4-Q.5)

4. [20 marks] *k*-means clustering and decision tree are popular models for unsupervised learning and supervised learning respectively. Think about how they can benefit each other in the following settings.

   a) Empowering decision tree by *k*-means clustering for supervised learning tasks. Describe your idea and present a conceptually sound model. What are the expected benefits of your new model? What are the potential shortcomings?
   b) Empowering *k*-means clustering by decision tree model for unsupervised learning tasks. Describe your idea and present a conceptually sound model. What are the expected benefits of your new model? What are the potential shortcomings?

   Note that if any of the ideas above is unfeasible, no solution is also acceptable with elaborated justifications.

5. [20 marks] DBSCAN clustering is sensitive to parameters (MinPt and epsilon ε) used and it does not work well when the data contains varying densities. Given the idea to use varying parameter values to adapt to regions with different densities, think about how DBSCAN algorithm should be modified so that it can be made more effective for data with varying densities. Describe your idea and present a conceptually sound model. What are the expected benefits of your new model? What are the potential shortcomings?

- E N D -