COMP4432: Final Part 2

ZHANG Caiqi 18085481d

December 17, 2021

1

1.1 a)

Several methods can be applied to handle the missing data for Stock 857. Here we assume the missing data refers to the data before the IPO day of the stock 857. The simplest way is to ignore the missing data and make the full use of the existing data. No matter which algorithm we want to use to predict the trend of the stock, there is enough data to train. Alternatively, we may fill in the missing data using the IPO value if needed. Or in general cases, we can fill in the null values using regression according to the non-null value around it. However, in terms of stock 857, there are a great number of missing values. Other methods are acceptable but applying the data reduction seems to be the simplest way.

1.2 b)

We should apply the data reduction in this case. We can ignore the missing data before the IPO day of stock 857 and use the data of all six stock after the IPO day. The reason is that no matter what we fill in to the missing days of stock 857, we will change the pattern frequencies. Filling incorrect data will leads to incorrect rules which is useless for the decision maker.

1.3 c)

Since the stock market is always in the process of small-scale changes, such fluctuations can affect our understanding of the overall trend. So we can take some methods to smooth the data appropriately. On the one hand, some denoising methods can be applied, such as wavelet domain denoising, Kalman filter, SVD denoising and so on. On the other hand, we can also apply a sliding window strategy to do the regression within a small window. If we

find any value is too far from regression curve, we can substitute it using the mean of the adjacent values.

1.4 d)

In non-sequential association rule mining task, we will change the numeric value into symbolic value to find the potential rules. Therefore, on the one hand, by properly defining the numeric-to-symbolic rules, we do not need to deal with the noise in the data. On the other hand, we can still use the above mentioned denoising methods such as wavelet domain denoising, Kalman filter to preprocess the data.

1.5 e)

For simplicity, we can adopt the Min-Max Normalization.

$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

However, some scholars also argues that for stock data, the global normalization is probably not the best solution [1], so they proposed a piecewise normalization which is also worth trying in this case.

2

- Hypothetical dimension table: Table 1.

Period Code	Year	Quarter	Month
001	2001	1	4
002	2001	1	5
003	2001	1	6
004	2001	2	7

Table 1: Hypothetical dimension table for *time*

- Fact table containing hypothetical measures: Table 2.
- Hypothetical OLAP result: First, lets suppose that we have a data cude with the following dimensions:
  - x-axis: Day1, Day2, Day3, Day4...
  - y-axis: Stock 001, Stock 011, Stock 857...

Stock ID	Period Code	Total Volume Sold
001	001	10000000
001	002	465465469
293	001	16876161
011	001	544912644

Table 2: Fact table containing hypothetical measures

– z-axis: Close, High, Low, Open

With the above data cube, after applying the Slice operation, we may get the following results for a certain day:

	Close	High	...
Stock 001	4	6	...
Stock 011	3	7	...
Stock 857	5	8	...
...	...	...	...

Table 3: OLAP result

3

- Customer: each stock can be treated as a customer.
- Item: the price movement of the stock can be treated as items. For example: Up, Down, Level.
- Transaction: in traditional sequential ARM, the transaction is a list of un-ordered items. However, as to the stock data, what we care are the sequence patterns, so the order does mater. Therefore, the transaction should be the sequence (with order) of the items (Up, Level, Down).
- Hypothetical data records: For example, if we choose the Stock 001 to be our target to find the sequential patterns:

Stock ID	Date	Items
001	Jan 01-04 2001	Up-Up-Up-Down
001	Jan 02-05 2001	Up-Up-Down-Level
001	Jan 03-06 2001	Up-Down-Level-Up

Table 4: Hypothetical data records for sequential association rule mining

- Hypothetical mining results: If given some certain min support and confidence, we may have some patterns look like the following.

- Up-Up-Up-Down
- Down-Down-Up-Down
- Level-Up-Level-Down

With the sequences above, if we find the stock if of "Up-Up-Up" movement in the past three days, it is very likely that it will go "Down" in next day. In other words, the rules represent the frequently appearing patterns of a stock. These findings can be applied to predict the future trend of a stock. It can also help people to select the stock according to the observation of its recent price movement.

4

4.1 a)

We can enhance the decision tree in supervised learning tasks in the following way: 1) apply k-means to cluster the training set into several clusters, 2) for each cluster, we can train a decision tree. When a new object comes, we can first see which cluster it belongs to, then apply the decision tree of that cluster.

Discussion: The advantage of the new model is that the classification can be done in a more detailed way in different clusters. [2] also proved the improvement in accuracy of this method by testing it in a heart disease dataset. The potential disadvantage is that it may increase the computational cost and make the runtime increase. Also, if the result is not good in the first clustering step, it may make the classification even worse later.

4.2 b)

Feature selection is a very crucial step to remove the redundant and irrelevant features. However, in unsupervised learning, the feature selection difficulty is greater than supervised learning. Therefore, we can enhance the k-means clustering to help select the features for future supervised learning. Proposed by [3], Feature Important Factor (FIF) can be used to indicate the significance of the features. The basic idea is that: 1) we do a first time clustering to generate class labels; 2) set up decision tree to calculate the FIF; 3) do the cluster algorithm again with the FIF to modify the similarity measure and then get the modified clustering result. Note that the first step can be done by the ULAC model introduced in [4]. The overall pipeline is shown in Figure 1.

Discussion: The advantage of the new model is that it can help to remove the redundant or irrelevant features in k-means clustering. The FIF will be close to 0 if one feature is regarded as not important by the decision tree. With more significant features, the clustering results can be improved. [3] also proved the improvement in accuracy of this method by testing it in 3 different datasets. The potential disadvantage is that it may also increase the computational

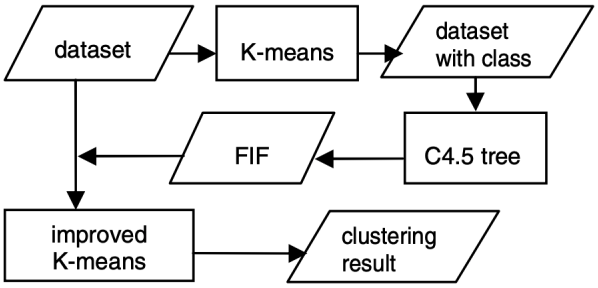

Figure 1: The structure of how to use decision tree to enhance k-means introduced in [3].

cost and make the runtime increase. Meanwhile, if FIF computed by decision tree cannot reflect the real weighting of a feature, it may make the clustering even worse later.

5

Algorithm 1: Progressive DBSCAN

```
Draw the k-dist graph;
Partition the graph by the sharp changes;
Use all the cut-points as ε_1, ε_2, ...ε_n (i = 1, 2, ...n)
for i in (1, 2, ...n) do
    ε = ε_i, MinPts = k;
    Adopt DBSCAN algorithm for points that are not marked;
    Mark all the clustered points.
end
```

To solve the problem, we can leverage the k-dist graph and do the clustering progressively. The k-dist graph can be drew in the following way: 1) compute the distance of each node to its k^{th} nearest neighbor, which is the so called k-dist. 2) sort all the nodes' k-dist in an ascending order and plot the k-dist graph. Figure 2 gives an example of varied density and its k-dist graph (k = 3).

Then, with the k-dist graph, we can find several sharp changes from the graph. For example, in Figure 2 (2), two sharp changes happen in dist = 1.2, and dist = 2.2. With $\varepsilon_1 = 1.2$ and $\varepsilon_2 = 2.2$, we can continue to cluster the nodes in 2 steps. Figure 3 (2)(3) show the clustering results with $\varepsilon_1 = 1.2$ and $\varepsilon_2 = 2.2$. Note that the nodes clustered in previous steps will be marked and not join the following clustering. Therefore, Figure 3 (3), it will not cluster the nodes which are already clustered in (2). Figure 3 (1) demonstrates the results using original DBSCAN. All the nodes are in the same cluster.

Discussion: The benefits of this model is that it can have better performance with different densities. The potential shortcoming is that it still requires a user-specified parameter k.

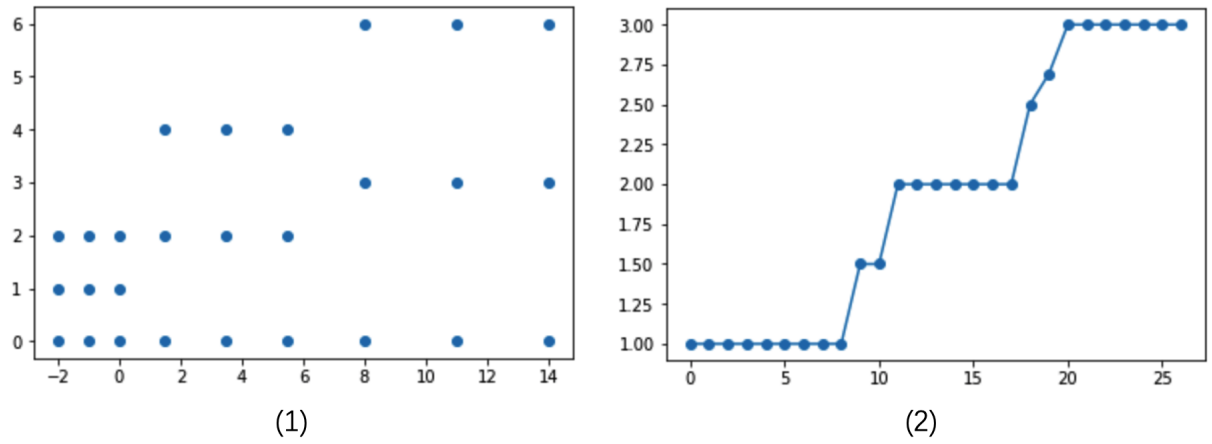

Figure 2: (1) An example of the nodes with various density. (2) the k-dist graph.

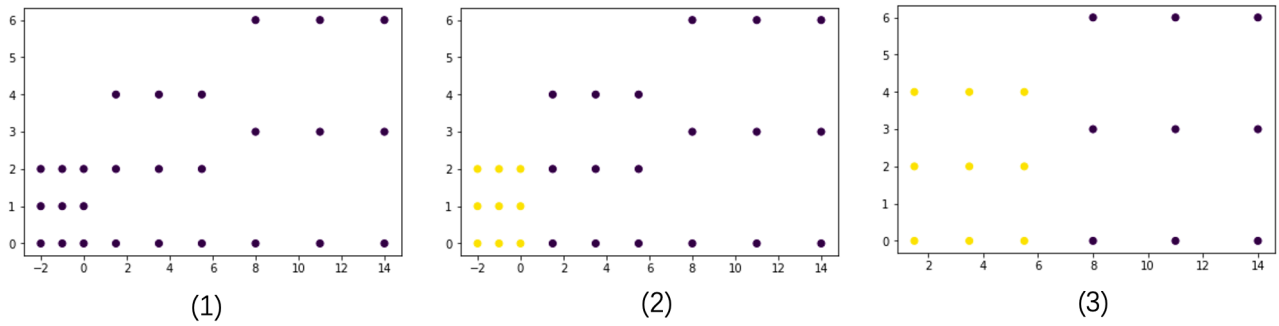

Figure 3: (1): Using original DBSCAN; (2)(3): Using improved progressive DBSCA.

It will mainly influence the results by $MinPts$. Although value of ε_i also depends on k, it does not change dramatically as k changes. One can find almost the same cut-points when k changes.

References

[1] M. Gavrilov, D. Anguelov, P. Indyk, and R. Motwani, "Mining the stock market (extended abstract) which measure is best?" in *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2000, pp. 487–496.

[2] M. Shouman, T. Turner, and R. Stocker, "Integrating decision tree and k-means clustering with different initial centroid selection methods in the diagnosis of heart disease patients," in *Proceedings of the International Conference on Data Science (ICDATA)*. The Steering Committee of The World Congress in Computer Science, Computer ..., 2012, p. 1.

[3] X. Li, J. Yang, Q. Wang, J. Fan, and P. Liu, "Research and application of improved k-means algorithm based on fuzzy feature selection," in *2008 Fifth International Conference on Fuzzy Systems and Knowledge Discovery*, vol. 1. IEEE, 2008, pp. 401–405.

[4] P. Liu, J. Zhu, L. Liu, Y. Li, and X. Zhang, "Data mining application in prosecution committee for unsupervised learning," in *Proceedings of ICSSSM'05. 2005 International Conference on Services Systems and Services Management, 2005.*, vol. 2. IEEE, 2005, pp. 1061–1064.