Stock Classification from High Frequency Market Data

Sofiane Ezzehi[1] **and Bastien Le Chenadec**[1]

[1]École des Ponts ParisTech, Master MVA

Contribution Statement

1 Introduction

The goal of this challenge is to predict the stock corresponding to a given order book. Each sample is a chronological sequence of 100 events of orders for a given stock. To make this task more challenging, a lot of the data is missing, and some properties have been normalized.

The first part of this challenge was devoluted to correctly dealing with the temporal and categorical aspects of the data. Then we focused on dealing with the change of distribution of the data between the training and the test set.

In this report, we describe the data, the different models we used, the training procedure as well as the results we obtained.

2 Data overview

Each sample in the dataset is constituted of 100 events of orders for a given stock. There are 24 different stocks which are equally distributed in the training, validation, and test sets. There are 160800 samples in the training set and 80600 samples in the test set.

Feature	Type	Description
Venue	Categorical	The venue where the order was placed.
Order id	Integer	A unique identifier, which can be used to retrace updates to the order.
Action	Categorical	The type of action (new, delete, update).
Side	Categorical	The side of the order (buy, sell).
Price	Float	The price of the order.
Bid	Float	The best buying price for the stock.
Ask	Float	The best selling price for the stock.
Bid size	Float	The number of shares available at the best buying price.
Ask size	Float	The number of shares available at the best selling price.
Trade	Categorical	Whether a trade occured or not.
Flux	Integer	The quantity of shares for this order.

Table 1: Data description.

Each event is described by 11 features, as described in Table 1. There are 4 categorical features, 5 continuous features, one integer feature (flux) and the last integer feature (order id) also has some categorical properties as it links the different events of the same order.

2.1 Visualization

To get a better understanding of the data, we performed an in-depth series of visualizations. The idea was to see if some features – original or derived – were sufficient to differentiate between some stocks. To do so, we made use of different types of statistical plots (boxplots, histograms...) where we hoped, at each step, to see if one or several stocks were clearly identifiable.

To test the relevance of the features that we derived, we used them to train a random forest classifier on the training set. Then we used feature importance to see if the features we derived were relevant.

Let's first take a look at some of the interesting features we derived. We need to keep in mind that the goal is to extract an ensemble of features that would be discriminative for different stocks. Also, the following list is not exhaustive, and we will only present the most interesting features.

Bid-ask spread The bid-ask spread (difference between the best buying and selling price) is a natural feature to consider in a financial context. Since it is a measure of the liquidity of the stock, we expected it to be a particularly good indicator of the stock. We plotted, on figure **??**, the boxplot of the bid-ask spread for each stock. More precisely, for each stock, we considered all the bid-ask spreads of all the observations of the stock in the training set. Furthermore, we only collected the points where a limit order was placed, updated, deleted or traded. While we can see that the spread does not seem to completely separate the stocks, it is still a good indicator of the stock with a decent variety of boxplot shapes. As we will see in the results section, the bid-ask spread was actually the most important feature in the random forest classification.

Price outliers (number and price value) An *a priori* good feature we derived was the characteristics of the price outliers. We can see on figure **??** representing the percentage of price outliers per number of ticks, that the stocks have relatively different distributions. We plotted on figure **??** the boxplot of the percentage of price outliers for each stock, over all available observations in the training set. We can see that the stocks do not really separate well. Nevertheless, we get more interesting results when we look at the actual values of the price outliers. We plotted on figures **??** and **??** the boxplot of the bid and ask price outliers for each stock. For example, we can see that stocks 7 and 20, as well as stocks 5, 8, 15, 17, 19 and 23 distinguish themselves from the others by having a higher percentage of outliers. This

type of result is what we are looking for, since it gives additional discriminative information about the stocks.

Volume A very straightforward feature, which is actually given in the data, is the volume of the orders. We plotted on figure **??** the boxplot of Bid and Ask sizes for each stock. We can see that, as in the case of the bid-ask spread, the stocks separate decently well.

Price outliers (flux) Similarly, we looked at the flux of the orders that were price outliers. We distinguished 2×2 types of outliers: Bid/Ask and Addition/Deletion. We plotted on figures **??** and **??** the corresponding boxplots. Here, we can see that a variety of stocks are clearly distinguishable. A striking example is stock 20, which stands out alone in the bid addition outliers. Therefore, if the test set has a similar distribution, we can expect our model to perform almost perfectly on this stock.

Other unconclusive features Among the other features we looked at, we can mention the trade proportion per stock, the trade price and volume per stock, the bid-ask spread distinguished by venue or the proportion of limit orders. None of these features were particularly discriminative.

2.2 Preprocessing

2.3 Graph construction

To deal with the temporal and categorical aspects of the data, one idea is to represent the data as a graph that better represents the relationships between the different events. After a bit of trial and error, we made the following arbitrary choices to construct a graph representing a sample :

1. The graph is undirected.
2. Each event is a node in the graph.
3. Each venue corresponds to a connex component in the graph.
4. If two events happen successively at a venue, there is an edge between them.
5. If two events have the same order id, there is an edge between them.

The separation of the graph into connex components corresponding to the venues is motivated by the high frequency nature of the data. Indeed, very few actors can react with high speed to the events happening in another venue, so it makes sense to assume that the events happening in different venues are somewhat independent. Note that the order id is unique to a venue so there should be no edge between the different venues.

The features that are not encoded in the structure of the graph can be placed on the nodes and the edges.

- Node features: price, bid, ask, bid size, ask size, flux.
- Edge features: action, side, trade.

We also add the venue and a time feature to the nodes, otherwise the model would not be able to distinguish between the different venues, and would not know in which direction the time flows.

3 METHOD

In this section, we describe the different models we used. We experimented with three main types of models: recurrent neural networks which are well suited for sequential data, graph neural networks which exploit the graph representation of the data, and statistical models that exploit features extracted from the sequence of events. From the start we knew that our final prediction would be an ensemble of models, so we explored different types of models to maximize the diversity of the ensemble. Indeed the diversity should improve the robustness of the predictions, especially given the change of distribution between the training and the test set.

3.1 Recurrent Neural Networks

Long-Short Term Memory (LSTM) networks are a type of recurrent neural network that are well suited for sequential data [1]. We used a simple LSTM model to extract features from the sequence of events. We used the following architecture inspired by the baseline :

- An embedding layer to embed the categorical features (venue, action, side, trade) over 8 dimensions. The other features are already continuous and do not need to be embedded.
- A bidirectional LSTM layer with 128 hidden units and two layers.
- A linear layer with one hidden layer of size 128.

This model is simple and fast to train, and it gave us a good baseline to compare with the other models. We quickly reached a validation accuracy of around 50% with this model and some data preprocessing, which was encouraging. However we quickly realized that the model was not robust to the change of distribution between the training and the test set as it only reached around 30% accuracy on the test set.

3.2 Graph Attention Networks

Graph Attention Networks (GAT) [2] have been shown to be effective in many tasks. Like other graph neural networks, GATs aggregate information from the neighbors of each node to compute its embedding. The main difference with other models is that GATs use an attention mechanism to weight the neighbors of each node. Specifically we used the improved version of GATs suggested in [3].

Let G be an undirected graph with N nodes denoted $[\![1, N]\!]$. Let d_1 be the dimension of the node embeddings, and $h_1, \dots, h_N \in \mathbb{R}^{d_1}$ be the said embeddings. Let d_2 be the dimension of the edge embeddings, and $\left\{e_{i,j} \,|1 \leq i, j \leq N\right\}$ be the said embeddings. Let $W_1 \in \mathbb{R}^{d' \times d_1}$, $W_2 \in \mathbb{R}^{d' \times d_2}$ and $a \in \mathbb{R}^{d'}$. The attention weights are :

$$w(h_i, h_j, e_{i,j}) = a^T \,\text{LeakyReLU}(W_1 h_i + W_1 h_j + W_2 e_{i,j}) \quad (1)$$

The attention weights are normalized using the softmax operator :

$$\alpha_{ij} = \frac{\exp(w(h_i, h_j, e_{i,j}))}{\sum_{k \in \mathcal{N}_i} \exp(w(h_i, h_k, e_{i,k}))} \quad (2)$$

where $\mathcal{N}_i$ denotes the set of neighbors of node i in G. The embedding of node i is then computed as :

$$h'_i = \text{LeakyReLU}\left(\sum_{j\in\mathcal{N}_i} \alpha_{ij} W_1 h_j\right) \tag{3}$$

In general we will use multi-head attention, with K heads, $a^{(1)}, \dots, a^{(K)} \in \mathbb{R}^{d'/K}$, $W_1^{(1)}, \dots, W_1^{(K)} \in \mathbb{R}^{d'/K \times d_1}$ and $W_2^{(1)}, \dots, W_2^{(K)} \in \mathbb{R}^{d'/K \times d_2}$:

$$h'_i = \text{LeakyReLU}\left(\frac{1}{K}\sum_{k=1}^{K}\sum_{j\in\mathcal{N}_i} \alpha_{ij}^{(k)} W_1^{(k)} h_j\right) \tag{4}$$

Furthermore, we will stack multiple GAT layers to obtain a deeper model. We may also apply a multi-layer perceptron to the embeddings of the last layer to obtain a more expressive representation. This model is easily parallelizable which is useful for mini-batch training.

3.3 Other graph models

We started out with GAT models because we already had some experience with them, and they were very effective. However, we also trained other graph models to diversify our ensemble (as long as they supported edge features in the graph). Here is an exhaustive list of models that gave us acceptable results and were added to the ensemble (in no particular order) :

- Generalized GNN [4]
- Pathfinder Discovery Network [5]
- Principal Neighbourhood Aggregation [6]
- Graph transformer [7]
- General GNN [8]

None of these models performed as well as the GAT models (although this may be due to more superficial hyperparameter tuning), but they did improve the diversity of the ensemble.

3.4 Statistical models

We used multiple statistical models on simple features extracted from the data to diversify our ensemble. These models are intuitively more robust because they rely on simple features that are less likely to be affected by the change of distribution. Here is a non exhaustive list of the models we used :

- Random Forest Classifier
- Ada Boost Classifier
- Logistic Regression
- K-Nearest Neighbors
- Ridge Classifier
- ...

3.5 Franck Zibi's model

Last year's CFM challenge also involved adapting to a change of distribution between the training and the test set. Inspired by the winning solution of last year's challenge, we devised a similar model. It is a simple model motivated by boosting methods, that learns the residuals of the predictions of a base model.

1. A random forest classifier is trained on the training set. This model is able to output a probability distribution over the classes.
2. For each class, three models are trained to predict the residuals of the random forest classifier (a random forest regressor, a k-nearest neighbors regressor and a linear regressor).
3. For each class, the three models are stacked using a linear regressor.

We can then simply predict the class of a sample by adding the output of the random forest classifier to the output of the stacked models, and taking the class with the highest probability.

4 TRAINING

For all the models we used the same split of the training set into a training and a validation set. We used 90% of the training set for training and 10% for validation. We did not use cross validation because the training set was already very large and well-balanced.

4.1 Loss

Like in most classification tasks, we used the cross-entropy loss to train our models. When it became clear that the distribution of the test set was very different from the training set, we experimented with different loss functions to make the models more robust to this change. We settled on Minimum Class Confusion (MCC) loss [9] which aims at minimizing the confusion between the classes on a target domain (the test set in our case).

Let $(X_n)_{1\leq n\leq N}$ be a batch of testing samples, and $\hat{Y}_n = F(X_n) \in \mathbb{R}^{24}$ be the output of the model for sample X_n. MCC uses temperature scaling to soften the predictions :

$$\tilde{Y}_{n,i} = \frac{\exp\left(\hat{Y}_{n,i}/T\right)}{\sum_{j=1}^{24}\exp\left(\hat{Y}_{n,j}/T\right)} \tag{5}$$

where $T > 0$ is the temperature (we used $T = 2.5$). We then compute an entropy over the predictions :

$$H_n = -\sum_{i=1}^{24} \tilde{Y}_{n,i}\log\left(\tilde{Y}_{n,i}\right) \tag{6}$$

This entropy will be used to give more importance to samples for which the model is more confident. We then use the following weights :

$$W_n = N \times \frac{1+\exp(-H_n)}{\sum_{i=1}^{N} 1+\exp(-H_i)} \tag{7}$$

where N is the batch size. This weight makes use of Laplace smoothing for numerical stability. Finally the class confusion between class i and class j is given by :

$$C_{i,j} = \tilde{Y}_{\cdot,i}^T \text{diag}(W_1, \dots, W_B)\tilde{Y}_{\cdot,j} \tag{8}$$

This class confusion is normalized to account for class imbalance :

$$\tilde{C}_{i,j} = \frac{C_{i,j}}{\sum_{k=1}^{24} C_{i,k}} \tag{9}$$

Which gives us the final loss :

$$\mathcal{L}_{\text{MCC}} = \frac{1}{24} \sum_{i=1}^{24} \sum_{j=1}^{24} \left| \tilde{C}_{i,j} \right| \tag{10}$$

Note that this note does not use labels from the target domain, and is therefore unsupervised. It can simply be added to the supervised loss to make the model more robust to the change of distribution :

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \mu \mathcal{L}_{\text{MCC}} \tag{11}$$

we found that $\mu = 0.1$ was a good value and we could easily go as high as $\mu = 1$ or $\mu = 10$ for models that trained easily. In practice this MCC loss is computed on a batch of the test set every 10 iterations of the training loop. We found that it did raise the accuracy of the models on the test set by around 5%.

4.2 Optimizer

We used the Adam optimizer with a starting learning rate of 5×10^{-3}. We did not use weight decay as it hindered the training of the models.

We used a scheduler to automatically decrease the learning rate. The scheduler decreased the learning rate by a factor of 0.95 every epoch, which yielded a learning rate of around 1×10^{-3} after 30 epochs which is a typical number of epochs for our models.

We generally stopped the training quickly when the accuracy on the validation set stopped increasing significantly. This is motivated by the change of distribution between the training and the test set, which means that we do not want to overfit the training set.

4.3 Other unsuccessful experiments

We tried other tricks to make the models more robust to the change of distribution.

- We experimented with data augmentation on the training set : given the high stochasticity of the data, we could easily generate new samples by adding some noise to the prices, the volumes, the flux, etc.
- Starting with a trained model, we extracted the test samples for which it was the most confident, and submitted them to the leaderboard (with the rest left at random). We could infer that we had around 80% accuracy on these samples. Thus we tried fine-tuning a model on these samples, but it did not improve the accuracy on the test set.

5 RESULTS

5.1 Statistical models

Random Forest Classifier Throughout our experiments, we found the random forest classifier handy for feature selection, as well as for training a model purely based on handcrafted features.

We present in this section the results of 2 random forest classifiers that we trained. The first model was trained on the Bid-Ask spread, the Bid and Ask volume, the number of price outliers, the price outliers, the flux of the price outliers, and the proportion of each venues on which the orders were placed. The second model was only trained on a subset of the features of the first model, which were the Bid-Ask spread, the Bid and Ask volume and the venue proportion.

We obtained a very important result that framed our approach for the rest of the challenge. We found that the first model had a 49% accuracy on the validation set, while the second model had a 28% accuracy. However, both models performed very similarly on the test set, with an accuracy of respectively 22% and 20%. This result is very important because it gives us a clear indication that the "outlier" features, while being very discriminative on the training set, are almost irrelevant on the test set. At this point, we had 2 choices: either we could eliminate the outlier features from our ensemble, or we could try to find a model that would be able to detect how the "outlier" features distribution changed in the test set.

REFERENCES

[1] Sepp Hochreiter and Jürgen Schmidhuber. Long Short-Term memory. *Neural Computation*, 9(8):1735–1780, 11 1997. doi: 10.1162/neco.1997.9.8.1735. URL https://doi.org/10.1162/neco.1997.9.8.1735.

[2] Petar Velikovi, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Píetro Lió, and Yoshua Bengio. Graph attention networks. *arXiv (Cornell University)*, 2 2018. doi: 10.17863/cam.48429. URL https://arxiv.org/pdf/1710.10903.pdf.

[3] Shaked Brody, Uri Alon, and Eran Yahav. How Attentive are Graph Attention Networks? *arXiv (Cornell University)*, 5 2021. doi: 10.48550/arxiv.2105.14491. URL https://arxiv.org/abs/2105.14491.

[4] Guohao Li, Chenxin Xiong, Ali Thabet, and Bernard Ghanem. DeeperGCN: All you need to train Deeper GCNs, 6 2020. URL https://arxiv.org/abs/2006.07739.

[5] Benedek Rozemberczki, Peter Englert, Amol Kapoor, Martin Blais, Bryan Perozzi, and Google Research. Pathfinder discovery networks for neural message passing. page 12, 2021. URL https://arxiv.org/pdf/2010.12878.pdf.

[6] Gabriele Corso, Luca Cavalleri, Dominique Beaini, Pietro Liò, and Petar Velikovi. Principal neighbourhood Aggregation for Graph Nets, 4 2020. URL https://arxiv.org/abs/2004.05718.

[7] Yunsheng Shi, Zhengjie Huang, Shikun Feng, Hui Zhong, Wenjin Wang, and Yu Sun. Masked Label Prediction: Unified Message Passing Model for Semi-Supervised Classification, 9 2020. URL https://arxiv.org/abs/2009.03509.

[8] Jiaxuan You, Rex Ying, and Jure Leskovec. Design space for graph neural networks, 11 2020. URL https://arxiv.org/abs/2011.08843.

[9] Ying Jin, Ximei Wang, Mingsheng Long, and Jianmin Wang. *Minimum class confusion for versatile domain adaptation*. 1 2020. doi: 10.1007/978-3-030-58589-1\{_}28. URL https://doi.org/10.1007/978-3-030-58589-1_28.

Appendix