Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 4, pp. 2851–2861, DOI 10.12694/scpe.v25i4.2853

APPLICATION OF HETEROGENEOUS DATA ANALYSIS BASED ON SEA GRID IN USER INVESTMENT ANALYSIS

AQIAN LIU[*]AND SHUKE HUANG[†]

Abstract. To solve the problem of yield calculation in complex investment scenarios, a time-cost double-weighted rate of return calculation method based on SEA grid is proposed, and its effectiveness is verified by comparing with traditional methods and researching the quantitative evaluation and analysis method of user investment based on structured data. To solve the above problems, a data-heterogeneous federated learning method based on user investment analysis FedPSG is proposed, which changes the data form transmitted from the client to the server from model parameters to model scores, and only a small number of clients need to upload model parameters to the server in each round of training, thereby reducing communication costs. At the same time, a model retraining strategy is proposed, which uses server data to train the global model for second iteration, and further improves the model performance by alleviating the impact of data heterogeneity on federated learning. The method of event dimension analysis of user investment is designed, and a credibility index is proposed to evaluate the analysis results. Experiments show that by combining event data, it can effectively provide users with event factors in the fluctuation of investment profit and loss, and help users better analyze their own investments.

Key words: SEA grid, algorithm, FedPSG, Heterogeneous data, Event Research Method, Structured data

1. Intrdocution. With the continuous development of China's financial market, investors' enthusiasm is increasing, but the individual investment user group tends to be non-professional, and it is difficult to make systematic evaluation and analysis of their own investment. Investment analysis for individual users mainly focuses on some securities investment software, which often only gives simple statistical data and does not carry out systematic analysis. In the face of complex investment scenarios, the existing yield calculation methods have various problems such as yield jump, inconsistency between income and yield, which in turn affects the accuracy of investment analysis, so it is of great value to analyze user investment from a more comprehensive dimension and give intuitive analysis results.

There is currently a lot of research on the application of investment analysis methods. (Fama and French et al.,1992,2015) [1, 2] constructed three-factor and five-factor models to concretize the influencing factors of excess returns, adding the company's market capitalization factor and the ratio factor of book value to market capitalization in addition to the market portfolio return factor of CAPM (Capital Asset Pricing Model), making the model more perfect. The Brinson model (Brinson et al.,1986) [3]is based on the fund's position data to decompose investment income, and only requires the fund's position data disclosed quarterly. With the advancement of technology, methods based on machine learning have developed rapidly. (Yu et al) [4]. proposed an event extraction model combining tree-structured long short-term memory network (Tree-LSTM) and gated recurrent unit network (GRU), and added the dependent syntactic information in the text to the model through the Tree-LSTM structure, which was improved compared with other models. (Li et al.,2020) [5] proposed to treat event extraction as a multi-round question answering task, and use the good performance of the pre-trained model BERT on the machine reading comprehension task to improve the performance of the event extraction task, and (Du et al.2020) [6] designed a new question answering strategy on this basis to further improve the performance. In addition, there are event extraction methods based on graph convolutional neural networks (Nguyen et al.,2018. Yang et al.,2018) [7, 9], and event extraction methods based on adversarial

[*] School of Finance and Accounting, Henan Industry and Trade Vocational College, Zheng Zhou, 451191, China (Aqian_Liu@outlook.com)

[†] School of Finance and Accounting, Henan Vocational College of Agriculture, ZhengZhou, 451450, China (20210600480@hnca.edu.cn)

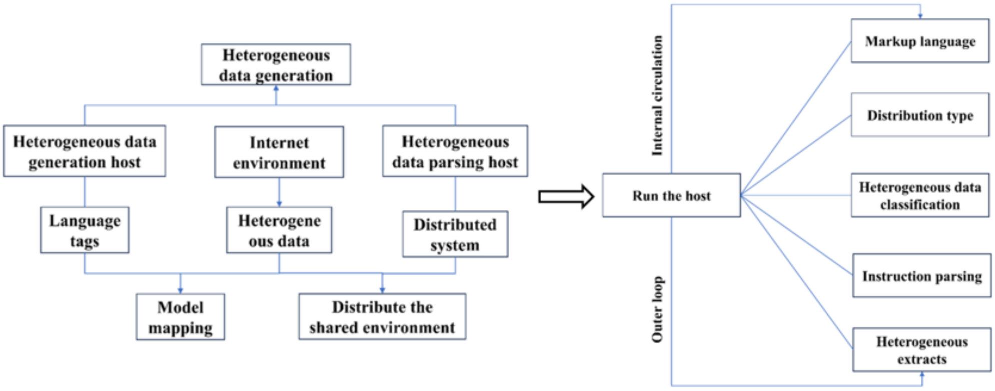


Fig. 2.1: Generate module diagrams

training (Wang X et al.,2019) [10].

One is a method based on stock price prediction, which predicts the change of stock price after the news release by building a model, so as to obtain the impact of events in the news on the stock price. (Schumaker et al.,2009) [11] extracted features from news texts from the perspective of text expression, and used a support vector machine model to predict stock price trends after news releases. (Ding et al.2015) [12] used Open IE technology (Ding et al.2014) [14] and dependency syntactic analysis to extract structured events from unstructured news text, which are composed of actuators, actions, objects and time 4-tuples [13], and then used neural tensor networks to characterize events, and then used convolutional neural networks combined with short-term and long-term events to predict stock prices, and experimental results show that event representation can effectively solve the problem of event sparseness in large-scale financial news data [14], and convolutional neural network models can be used The long-term impact of events to improve predictive performance. Based on the above analysis, this paper hopes to make use of heterogeneous data composed of structured data such as user investment transaction records, stock index quotes, stock industry, and unstructured data such as financial news texts to systematically and comprehensively evaluate and analyze user investment. Based on SEA grid, a new algorithm is added to the user investment system to analyze heterogeneous data more effectively and efficiently

2. Heterogeneous data algorithm models.

2.1. Heterogeneous data algorithm generation module. In order to realize the heterogeneous data synchronization system, the design of model generation/analysis module fully follows the model markup language, and can control the distributed sharing environment of information transmission under the action of host elements [14], so as to generate completely independent data analysis and query statements. The specific connection form is shown in Figure 2.1. The transmission direction of heterogeneous Internet information can only be generated by the data generation host pointing to the external application processing structure, which does not violate the algorithm mapping relationship. The stronger the connection stability of the distributed system, the higher the application level of data sharing service in the synchronous system, and vice versa [15].

2.2. Heterogeneous pattern extraction module. As the key application structure of the distributed heterogeneous data synchronization system, the heterogeneous mode extraction module can synchronously execute the inner loop and outer loop instructions of the information parameters under the action of XML markup language [17], and finally store the information parameters that meet the distributed discrimination requirements directly in the system database host, and the specific connection principle is shown in Figure 1. In the actual application process, due to the different execution links of markup language, the classification and extraction requirements of heterogeneous data will be different, and the discrimination criteria of the running host for distributed nodes will become the only condition to determine the application stability level

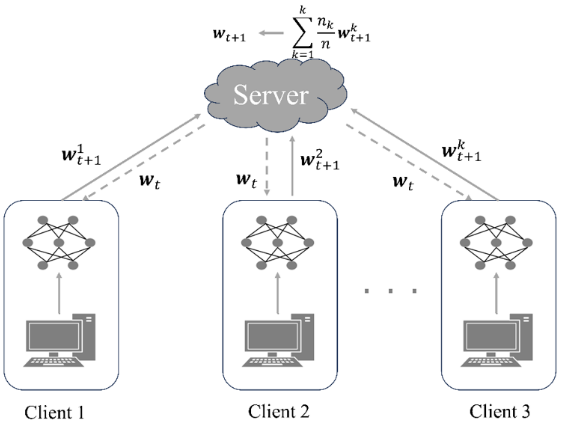


Fig. 2.2: Algorithmic models

of heterogeneous information synchronization and sharing environment.

2.3. Heterogeneous data models and algorithms. Heterogeneous data algorithm systems typically consist of a single server and multiple clients [18]. The idea of this FedPSG (Research on federated learning and particle swarm optimization) algorithm is to combine individual clients running stochastic gradient descent with servers running model averaging calculations. As shown in Figure 2.2, in order to fully apply heterogeneous data to the model, in each round of training, the client downloads the global model and trains the model with its local data, and then uploals the local model parameters to the server. The server coordinates the joint training of the clients and updates the global model by aggregating local model parameters. A smaller value of K means that the server receives fewer client models per round of training, and the less communication costs are required. However, the selection of some client models will cause other clients not to participate in the training of the global model, and some clients that perform better on the scoring dataset will dominate the training of the global model, which will cause unfairness in federated learning. When K is small, such as K equals 1, the server accepts only one client model, but the data of a single client is not representative of all clients, especially in a non-IID data environment. If a client model is simply set as a global model, the data distribution used by federated learning to train the global model is inconsistent with the real data distribution, resulting in a decrease in model accuracy.

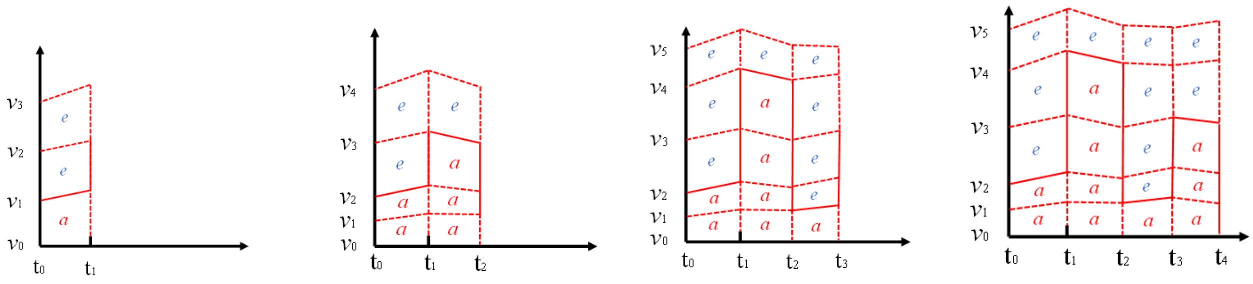
3. SEA grid.

3.1. SEA Mesh and Build. The process of building a SEA (Let all portfolios of an investor be S, in which A portfolio is A, except for A in S). The other parts are taken as combination E, and it can be seen that S is composed of A and E. network is as follows: Step 1 [19]: Initialize i to 0. Step 2: Forward scaling of the SEA mesh. When $0 \leq i < n$, in the t_i to t_{i+1} range, extend the market value line in the grid in the positive direction of the timeline T according to the proportion of the respective market value changes of S, E, and A. Mesh belonging to combination A is marked with a and mesh belonging to combination E is marked with e. Step 3: Reverse divide the SEA mesh. When $0 \leq i < n$, in the investment range from t_0 to t_i, starting from the result of A's capital change at t_i, divide the portfolio grid cells along the opposite direction of the T axis according to the changes in the respective market capitalizations of S, E, and A. The new grid is also marked with a and e. Step 4: Update $i = i + 1$. If $i < n$, go to step 2 to iterate, otherwise the SEA grid has been built. Figure 3.2 shows the SEA grid construction process at $n = 4$ and $m = 5$, which is forward-expanded and reverse-divided according to the capital change at each time point, and finally the entire market capitalization change graph is divided into $4 * 5$ for a total of 20 grid cells.

3.2. A double-weighted rate of return on time cost based on the SEA grid. The idea of double-weighted rate of return on time cost based on SEA grid is to split complex investments into simple investments

Algorithm 1 Global Model Update

```
Enter: Number of clients N; Top-K optimal strategy parameters K; Number of communications T; Number of local iterations E; $\eta$ learning rate; Inertia weights ; Acceleration factors c1, c2
Output: Global model $w_T$
Initialize $w_0$, $gbest$, $pbest_i$, $L_{gid}$
for $t = 1$ to $T$ do
    for each round $t$ from 1 to T do do
        for for each client n from 1 to N in parallel do do
            pbestk ← ClientUpdate(n,wt)
        end for
        L_pbest $[pbest_1, \ldots, pbest_N]$
    end for
    sort(L_pbest)
    pbest ← L_pbest [1, K]
    L_gid ← gid of pbest
    L_wt ← model of the client L_gid
    wt ← the mean of all numbers in L_wt
    if gbest < score(wt) then
        gbest ← score(wt)
    end if
    wt ← wt - $\eta \bigtriangledown l$(wt,Dm)
    initialize $w, \vee, wpbest$
    ClientUpdate(n,wt):
    $\beta$ (split Dn into batches of size B)
    for each weight layer $\vee l \in \vee$ do
        $\vee l \leftarrow \alpha - \vee l + c1r1(wpbest - \vee l) + c1r1(wt - \vee l)$
    end for
    $w \leftarrow w + \vee$
end for
for each client epoch i from 1 to E do
    for batch $b \in \beta$ do $w \leftarrow w - \eta \bigtriangledown l(w, b)$
    end for
end for
return pbestn to server
```



(a) (b) (c) (d)

Fig. 3.2: SEA Mesh building process

by constructing the SEA network method proposed [20] in this paper, and then weighting the cost of each simple investment, weighted into two parts:

1. Time weighting, considering that the duration of each simple investment is different, the corresponding cost has different action time, and then the impact on the overall rate of return is not the same, so the cost needs to be time-weighted;
2. Cost weighting, considering the different effects of funds on the rate of return at different points in time, it is necessary to consider the source of invested funds and the destination of the thrown funds for cost weighting. After obtaining the yield and double-weighted cost of each simple investment, it is finally combined into the yield of the complex investment.

3.3. Examples and Analysis of Results. There are various problems existing in the traditional rate of return calculation methods, which are summarized into three constraints in this paper: smooth transition constraint, that is, the return rate of the portfolio maintains a smooth transition without jumping when the capital flows in and out; The return consistency constraint, that is, the return rate and the return keep positive and negative consistency; Comparability constraint, that is, the return rate between various varieties in the same portfolio can be compared, and can be compared with the external return rate such as the return rate of the broad market index. Considering the shortcomings of the current yield calculation method, a time-cost double-weighted yield calculation method based on SEA is proposed:

1. Time-weighted, considering that the duration of each simple investment is different, the corresponding cost has different action time, and the impact on the overall yield is also different, so the cost needs to be time-weighted;
2. Cost weighting, considering that funds at different time points have different influences on the rate of return, it is necessary to consider the source of invested funds and the destination of sold funds for cost weighting. After obtaining the return and double-weighted cost of each simple investment, they are finally combined into the return rate of the complex investment.

To prove the effectiveness of the time-cost double weighted rate of return calculation method proposed, the traditional cost offset rate of return, cost accumulation rate of return, time weighted rate of return, and internal rate of return calculation method are compared below [21]. The comparison results of each return of Portfolio A are shown in Table 3.1, and the comparison chart is shown in Figures 3.2 and 3.3. The δ is a small value, minus δ indicates before the inflow and outflow of funds occur on the day, and the addition of δ indicates that after the inflow and outflow of funds occur on the day, it is used to show whether there will be a rate jump problem when the inflow and outflow of funds occur. As can be seen from Figure 3.3, the cost offset yield and the cost accumulation rate of return do not consider time weighting, that is, different costs have different action times, so there is a yield jump at the time of t1 and t3, and the cost weighting is not considered, that is, the cost at different points in time has different effects on the rate of return, so the calculated results also have errors. The time-weighted rate of return has a positive and negative inconsistency between the yield and the yield at both the t3 and t4 moments, because the cost weighting is not considered. The internal rate of return takes into account cost weighting, but when calculating cost weighting, it is assumed that the interest rate before the inflow and the interest rate after the capital outflow are the same as the internal rate of return to be calculated [22]. Only double-weighted yields satisfy all constraints. The following analyzes the double-weighted rate of return to meet the constraints. Because it is time-weighted, the time weight of the sudden inflow and outflow of funds is 0, so it does not cause the yield to jump and satisfy the smooth transition constraint. Because of the cost weighting, the positive and negative consistency of earnings and returns can be maintained, and the income consistency constraint can be satisfied. Satisfying the comparability constraint is discussed separately in two parts: internal comparison and external comparison: for internal comparison, the yield calculation method of portfolio market capitalization S, individual or combination A, and other combination E is a unified double-weighted calculation method, so the yield calculation between S, A, and E can be directly compared; For external comparison, first of all, considering that there is no external capital inflow and outflow of the portfolio market value S, the yield of S calculated by the double-weighted rate of return calculation method is consistent with the results calculated by the time-weighted method, and its double-weighted return for the entire period of time satisfies the time-weighted constraint, which is consistent with the time-weighted return calculation method used by indices, funds, etc., and they can be directly compared on this benchmark. If S

Table 3.1: Comparison of the results of the calculation of each yield in Portfolio A

Time	Time label	Cost offset rate (%)	Cost-incremental return (%)	Time-weighted yield (%)	Internal rate of return (%)	Double weighted return (%)
2015-01-01	t0	0.00	0.00	0.00	0.00	0.00
	t1-δ	100.00	100.00	100.00	100.00	100.00
2015-06-01	t1	33.33	33.33	100.00	100.00	100.00
	t1+δ	33.33	33.33	100.00	100.00	100.00
	t2-δ	-33.33	-33.33	0.00	-48.67	-62.66
2015-10-01	t2	-55.56	-33.33	0.00	-48.67	-62.66
	t2+δ	-55.56	-33.33	0.00	-48.67	-62.66
	t3-δ	-46.67	28.00	20.00	-42.27	-55.46
	t3	-36.84	-24.14	20.00	-42.27	-55.46
	t3+δ	-36.84	-24.14	20.00	-42.27	-55.46
	t4	-26.32	-17.24	40.00	-29.98	-40.14

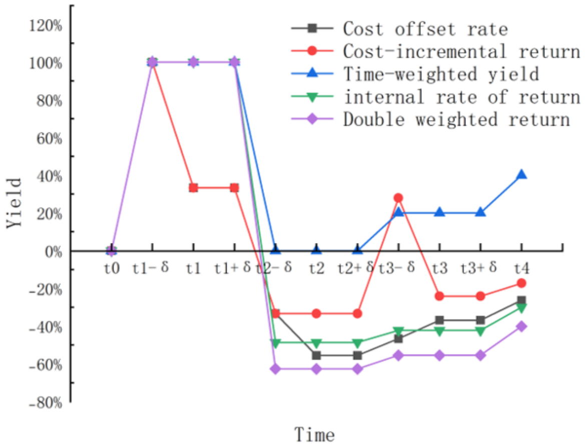


Fig. 3.3: Comparison chart of the calculation results of each yield of Portfolio A

has external capital inflows and outflows, at this time it is not known the returns before and after the inflow of these funds, and it can be assumed that the yield of these funds is 0 or equal to the market yield of the broader market, which can still be calculated normally. If you want to compare the yield of portfolio A with the outside, such as with the market index of the large market, you can replace E with the market index and then through internal comparison, A can be compared with the market index of the large market index.

4. User investment analysis combined with SEA grid heterogeneous data algorithm model.

4.1. Quantify the impact of algorithmic models on user investment. When conducting event research, the event window used is $(5, 10)$ for a total of 16 days, which means that from 5 days before the event date to the 10th day after the event date, each day corresponds to an average abnormal rate of return, a total of 16 average abnormal returns, which is not convenient to show the impact of the event on the stock price [23]. Therefore, this paper divides the time into three phases, namely the early stage of the event, corresponding to the time interval $(-5, -2)$, the middle of the event, corresponding to $(-1,1)$, and the late stage of the event, corresponding to (2,5). By calculating the average cumulative abnormal rate of return in each time interval, the weight of the impact of the corresponding stage event on the stock price is obtained. Like the average rate of return, a positive average cumulative abnormal rate of return indicates that the event at that stage has a positive impact on the stock price, and a negative one is a negative impact, and the greater the absolute value, the greater the degree of impact. We refer to the average cumulative abnormal rate of return of these 3 stages as the weight factor of the impact of event types on stock prices, which is used to express the impact of event types on stock prices.

4.2. Analysis of user investment event dimensions. Take the weight factor vector in the opposite direction 20 vector represents 0 points, as long as it performs better than the vector, it will be greater than 0 points, the better the performance, the higher the score [24]. Define the weighted logarithmic Manhattan distance for 0 points as Dmax. The confidence score is calculated as shown in Equation 4.1.

$$\text{trust\_score} = \begin{cases} \frac{D_b - D_s}{D_b - D_{\min}} \times 50 + 50, & \text{if } D_{\min} \leq D_s < D_b \\ \frac{D_{\max} - D_s}{D_{\max} - D_b} \times 50, & \text{if } D_b \leq D_s \leq D_{\max} \end{cases} \tag{4.1}$$

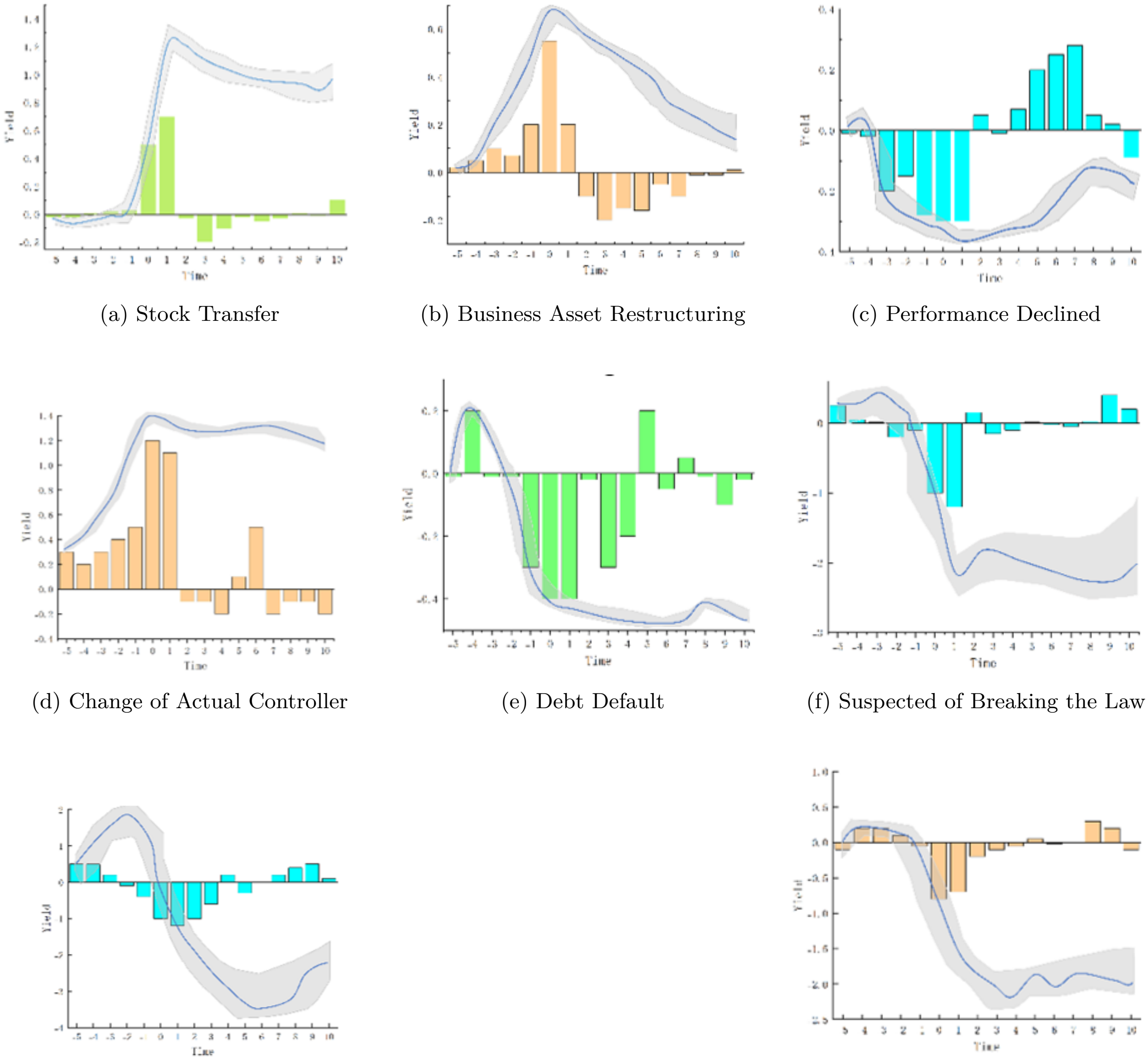
4.3. Experimental results and analysis.

4.3.1. Experimental heterogeneous data. To test the performance of FedPSG under heterogeneous client data conditions, IID, Non-IID (1) and non-IID (2) environments were set up experimentally, and the accuracy of FedPSG, FedAvg, FedShare and FedPS algorithms was compared on MNIST, FashionMNIST, and CIFAR-10 datasets. Among them, the FedPS algorithm is a simplified version of FedPSG after removing the model retraining strategy, and sets the ratio of the data volume of server data Dm to the data volume of client training data D $\gamma = 0.2$. IID data is independently and homogeneously, and there is no heterogeneity of client data. Non-IID (1) and NonIID (2) data are non-independently homogeneous, but Non-IID (1) data are more heterogeneous. The experimental results are shown in Figure 2 and Table 3.1. It can be observed that as the degree of client data heterogeneity increases, the accuracy of FedAvg algorithm on MNIST, FashionMNIST and CIFAR-10 datasets decreases significantly. After 100 rounds of training, when the data heterogeneity changed from IID to NonIID (2), the accuracy of FedAvg algorithm in MNIST and CIFAR-FashionMNIST datasets decreased by 15%, 20.08% and 28.65%, respectively. When the data heterogeneity changed from IID to Non-IID (1), the accuracy of FedAvg's algorithm on MNIST, CIFAR-10, and FashionMNIST datasets decreased by 57.53%, 35.89%, and 32.76%, respectively. It can be seen that the accuracy of the FedAvg model decreases with the deepening of data heterogeneity, which verifies that the client-side data heterogeneity has a great impact on user investment.

For the user investment event dimension analysis method that combines structured data and financial event data, it involves two parts of data, one of which is market data, which mainly includes the daily price data of all A-share companies listed on the SSE and SZSE, as well as the daily price data of the SSE Index and the SZSE Composite Index. The time period is from January 1, 2017 to October 1, 2020. The other part is the financial event data extracted from Chapter 3. Since ”change of beneficial owner” and ”change of beneficial shareholder” represent similar events, they are combined into one event and the event name is unified as ”change of beneficial owner”. In order to obtain more accurate statistical results, 11 event types with a sample size of less than 40 were removed, and 3 event types with the least significant results were removed according to the t-test results, and finally 15 event types were retained, which were: stock transfer/equity transfers, business asset reorganization, performance decline, change of actual controller, debt default, suspected violation of law, trading violation, financial constraints, violation of credit approval, falsification of financial information, change of actual controller, and so on. The event types are: stock transfer/equity transfer, performance decline, change of actual controller, debt default, suspected violation of law, transaction violation, fund tension, credit approval violation, financial information falsification, actual controller involved in litigation and arbitration, negative assets, negative executives, fund account risk, and reorganization failure. The total number of event samples is 18,047, covering 3,315 listed companies.

4.4. The results of the analysis of the impact of financial events on stock prices. This chapter uses the event research method to analyze the impact of 15 financial events extracted from Internet financial news on individual stock prices. The results of the analysis are described in detail below.

Figure 4.3 shows the AAR (Average Abnormal Rate of Return) and CAAR (Average Cumulative Abnormal Return) change chart for each event type in the event window, in which the ordinate represents the rate of return, and the abscissa represents the event window with a time range of $(5, 10)$, for a total of 16 trading days. The histogram represents AAR, the line chart represents CAAR, the gray area represents the 95% confidence



(a) Stock Transfer

(b) Business Asset Restructuring

(c) Performance Declined

(d) Change of Actual Controller

(e) Debt Default

(f) Suspected of Breaking the Law

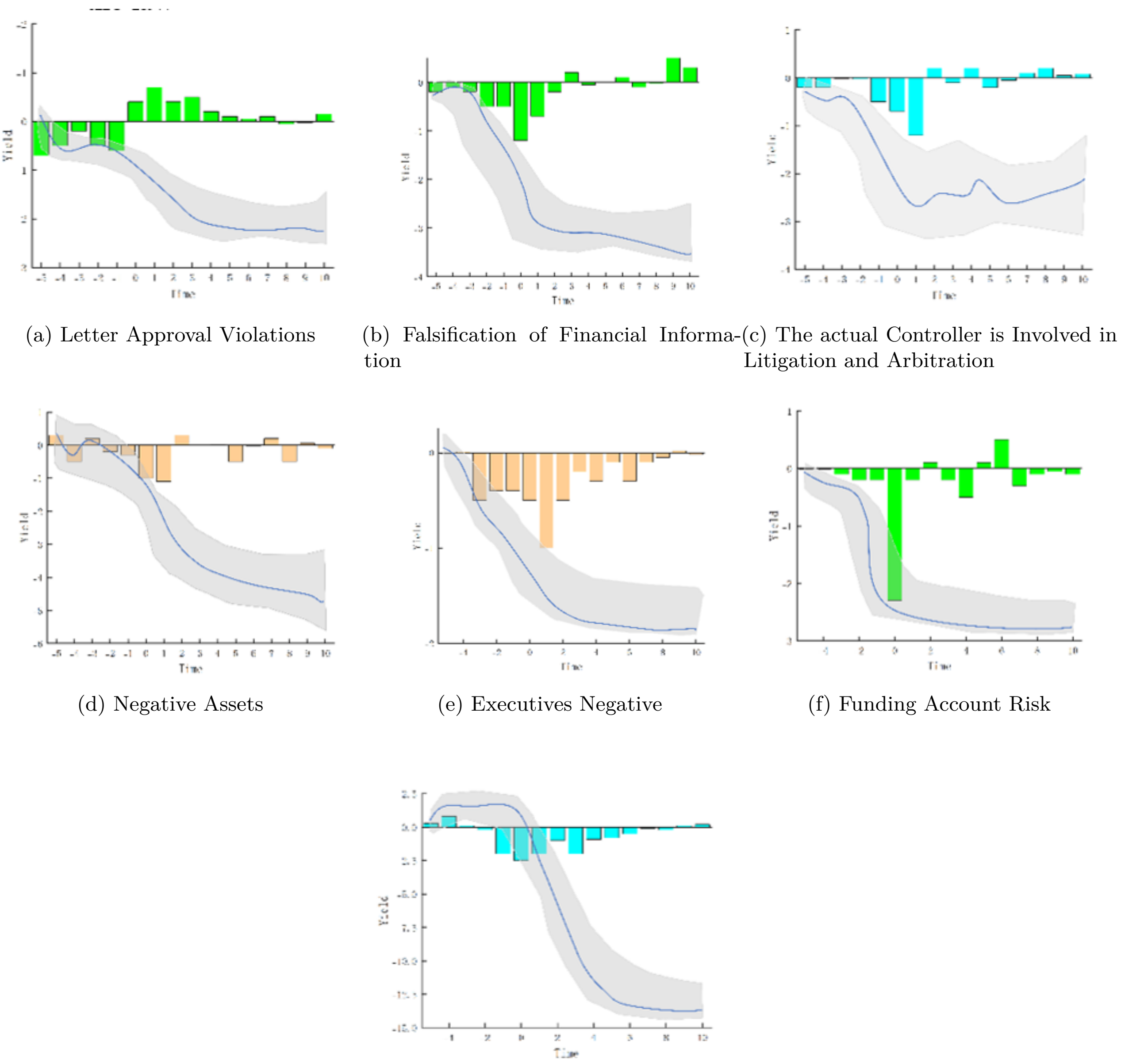
(g)

(h) Trading Violations

Fig. 4.2: AAR and CAAR change charts for each event type in the Events window (AAR for bar charts and line charts CAAR, gray area is CAAR 95% confidence interval)

interval of CAAR, the narrower the area indicates that the smaller the margin of error of the result, the more credible the result, and vice versa, the less trustworthy the result. A negative AAR indicates that the actual yield is lower than the expected yield, indicating that the event had a negative impact. Conversely, it indicates a positive impact. CAAR is the accumulation of AAR, with an upside indicating a positive impact and a downside indicating a negative impact. The results for each event type are analyzed below [24].

After 15 rounds of training, when the data heterogeneous condition changed from IID to Non-IID(1), the accuracy of the model ($\gamma = 0.20$) on the MNIST data set decreased by 0.08% from 97.33% to 97.25%, while



(a) Letter Approval Violations

(b) Falsification of Financial Information

(c) The actual Controller is Involved in Litigation and Arbitration

(d) Negative Assets

(e) Executives Negative

(f) Funding Account Risk

(g) Reorganization Failed

Fig. 4.3: Continued: AAR and CAAR change charts for each event type in the Events window (AAR for bar charts and line charts CAAR, gray area is CAAR 95% confidence interval)

the accuracy of the model ($\gamma = 0.02$) decreased from 96.54% to 92.03%, a decrease of 4.51%, and it can be seen that the accuracy of the model ($\gamma = 0.02$) on the MNIST dataset decreased more than the model ($\gamma =$).0.20) The accuracy of the model ($\gamma = 0.20$) on the CIFAR-10 dataset decreased by 1.1% from 51.16% to 50.06%, while the accuracy of the model ($\gamma = 0.02$) decreased by 8.22% from 42.79% to 34.57%, which shows that the accuracy of the model ($\gamma = 0.02$) decreased more than that of the model ($\gamma = 0.20$) on the CIFAR-10 dataset. Similarly, the accuracy of the model ($\gamma = 0.20$) on the FashionMNIST dataset decreased by 0.45% from 85.71%

Table 4.1: The accuracy of FedPSG, FedShare, FedAvg and FedPS in MNIST, CIFAR-10 and FashionMNIST after 15 rounds of training under three heterogeneous conditions of client data

	MNIST			CIFAR-10			FashionMNIST		
	IID	Non-IID (1)	Non-IID (2)	IID	Non-IID (1)	Non-IID (2)	IID	Non-IID (1)	Non-IID (2)
FedPSG	97.33	97.25	97.54	51.16	50.01	49.76	85.71	85.26	85.55
FedShare	97.94	96.02	96.38	48.39	36.59	40.82	85.7	80.52	81.89
FedAvg	97.73	40.20	82.23	52.28	16.39	32.17	86.46	53.7	57.81
FedPS	95.52	9.8	39.64	43.28	13.3	16.85	81.7	10	46.11

to 85.26%, while the accuracy of the model ($\gamma = 0.02$) decreased by 3.27% from 80.21% to 76.94%, i.e., the accuracy of the model ($\gamma = 0.02$) decreased more on the FashionMNIST dataset than the model ($\gamma = 0.20$). It can be seen that with the decrease of , the robustness of FedPSG's model accuracy when the degree of data heterogeneity changes. Therefore, in order to maintain the high quality of FedPSG training results in the case of data heterogeneity, the data amount of server data DM should be increased as much as possible [25].

Compared to FedAvg, FedShare and FedPSG performed more consistently under NonIID data conditions [25]. Especially on MNIST datasets, when the data heterogeneity conditions change from IID to NonIID (1) or non-IID (2), neither FedShare nor FedPSG accuracy changes by more than 2%. However, FedPSG's accuracy on CIFAR-10 and FashionMNIST datasets is significantly higher than FedShare's, and FedPSG accuracy decreases less than FedShare as client data heterogeneity increases. As shown in Figure 6 and Table 4.1, on the CIFAR-10 and FashionMNIST datasets, the FedPSG accuracy varies by no more than 2% as the data heterogeneity deepens [26]. FedShare's accuracy on CIFAR-10 and FashionMNIST datasets decreased by 11.8% and 5.18%, respectively, when the data heterogeneous conditions changed from IID to Non-IID (1). It can be seen that the FedPSG proposed in this paper is effective in dealing with the problem of data heterogeneity, and the effect is better than the general improvement algorithm.

5. Conclusion. This paper proposes a time-cost double-weighted rate of return calculation method based on SEA grid to solve the problem of calculating the rate of return in complex investment scenarios of users.

1. A user investment analysis method based on structured data is proposed, and six dimensional indicators of profitability, capital liquidity, timing ability, stock selection ability, action choice ability and risk control ability are designed, and multi-dimensional analysis of user investment is carried out.
2. FedPSG needs to select fewer client models when using the Top-K optimal strategy, so the training results of FedPSG are very dependent on the quality of server data. How to build high-quality server data while ensuring data privacy is one of the future research directions.
3. Combined with heterogeneous data, the dimensional analysis method of user investment events is studied. The influence of various events on the stock price is analyzed by using the event research method, the weight factor of event types on the stock price is designed to quantify the influence degree, the corresponding event dimension analysis is carried out on the stocks invested by users, and the result credibility index is calculated. The experiment shows that the introduction of financial event data can effectively provide the event factors in the fluctuation of investment profit and loss, and help users better analyze and reflect on their own investment behavior.
4. However, the heterogeneous data processing model based on SEA grid studied in this paper still has shortcomings. In order to achieve the goal of reducing communication costs, FedPSG needs to select fewer client models when using the Top-K optimal strategy, so the training results of FedPSG are very dependent on the quality of server data. How to build high quality server data while ensuring data privacy is one of the future research directions.

REFERENCES

[1] Fama, E. & French, K. The Cross-section of Expected Stock Returns. *The Journal Of Finance.* **42**, 427-465 (1992)

[2] Fama, E. & French, K. Five-Factor Asset Pricing Model. *Journal Of Financial Economics.* **116**, 1-22 (2015)

[3] Brinson, G. Bee bower G L. Determinants of Portfolio Performance. *Financial Analysts Journal.* **42**, 39-44 (1986)

[4] Yu, W., Yi, M., Huang, X. & Others Make It Directly: Event Extraction Based on Treelet and Bi-GRU. *IEEE Access.* **8** pp. 14344-14354 (2020)

[5] Li, X., Yin, F., Sun, Z. & Others . *Entity-relation Extraction As Multi-turn Question Answering, Proceedings Of The Annual Meeting Of The Association For Computational Linguistics.* pp. 1340-1350 (2019)

[6] Du, X. & Cardie, C. . *Event Extraction By Answering (Almost) Natural Questions//Proceedings Of The Conference On Empirical Methods In Natural Language Processing P.* pp. 671-683 (2020)

[7] Nguyen, T. & Grishman, R. . *Graph Convolutional Networks With Argument-Aware Pooling For Event Detection, Proceedings Of The Association For The Advance Of Artificial Intelligence.* **18** pp. 5900-5907 (2018)

[8] Rodríguez-Barroso, N., Jiménez-López, D., Luzón, M. & Others Survey on Federated Learning Threats: concepts, taxonomy on attacks and defences, experimental study and challenges. *Information Fusion.* **90** pp. 148-173 (2023)

[9] Yang, H., Chen, Y., Liu, K. & Others . *DCFEE: A Document-level Chinese Financial Event Extraction System Based On Automatically Labeled Training Data, Proceedings Of The Association For Computational Linguistics: System Demonstrations.* pp. 50-55 (2018)

[10] Wang, X., Han, X., Liu, Z. & Others Adversarial Training for Weakly Supervised Event Detection, Proceedings of the Association for Computational Linguistics: Human Language Technologies. (0)

[11] Schumaker, R. & Chen, H. Textual Analysis of Stock Market Prediction Using Breaking Financial News: The AZFin Text System. *ACM Transactions On Information Systems.* **27**, 1-19 (2009)

[12] Ding, X., Zhang, Y., Liu, T. & Others . *Deep Learning For Event-driven Stock Prediction, Proceedings Of The International Joint Conference On Artificial Intelligence P.* pp. 2327-2333 (2015)

[13] Park, S., Suh, Y. & FedPSO, L. Federated Learning Using Particle Swarm Optimization to Reduce Communication Costs. *Sensors.* **21** pp. 2 (2021)

[14] Ding, X., Zhang, Y., Liu, T. & Others . *Using Structured Events To Predict Stock Price Movement: An Empirical Investigation, Proceedings Of The Conference On Empirical Methods In Natural Language Processing P.* pp. 1415-1425 (2014)

[15] Zeng, Y., Yang, H., Feng, Y. & Others A Convolution BiLSTM Neural Network Model for Chinese Event Extraction[M]//Natural Language Understanding and Intelligent Applications. (Springer,2016)

[16] Huang, X. & Others Make It Directly: Event Extraction Based on TreeLSTM and Bi-GRU. *IEEE Access.* **8** pp. 14344-14354 (2020)

[17] Liu, S., Li, Y., Zhang, F. & Others . *Event Detection Without Triggers, Proceedings Of The Conference Of The North American Chapter Of The Association For Computational Linguistics: Human Language Technologies.* pp. 735-744 (2019)

[18] Almanifi, O., C-o, C., M-l, T. & Others Communication and computation efficiency in Federated Learning: A survey. *Internet Of Things.* **22**, 2 (2023)

[19] Li, X., Feng, J., Meng, Y. & Others A Unified Mrc Framework for Named Entity Recognition, Proceedings of the Annual Meeting of the Association for Computational Linguistics. (0)

[20] Devlin, J., M-w, C., Lee, K. & Others Bert: Pre-Training of Deep Bidirectional Transformers for Language Understanding, Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. (0)

[21] Ma, X., Zhu, J., Lin, Z. & Others A state-of-the-art survey on solving non-IID data in Federated Learning. *Future Generation Computer Systems.* **135** pp. 244-258 (2022)

[22] Vaswani, A., Shazeer, N., Parmar, N. & Others Attention Is All You Need, Proceedings of the Advance in Neural Information Processing Systems. (0)

[23] Sundermeyer, M., Schl"uter, R. & Lstm, N. Neural Networks for Language Modeling, Proceedings of Annual Conference of the International Speech Communication Association. (0)

[24] Huang, Z., Xu, W. & Yu, K. Bidirectional LSTM-CRF Models for Sequence Tagging. arXiv. (arXiv1508.01991 p,2015)

[25] Antweiler, W. & Frank, M. Do US Stock Markets Typically Overreact to Corporate News Stories. *SSRN Electronic Journal.* **10** pp. 2139 (2006)

[26] Mohammadtalebi, B., Jahangiri, M. & Eshghiaraghi, M. Investigating the Effect of Internal Rate of Return on Cash Recycling on the Abnormal Returns of Companies Accepted in Tehran Stock Exchange. *Advances In Mathematical Finance And Applications.* **3**, 1-10 (2018)

[27] Magni, C. & Martin, J. The Reinvestment Rate Assumption Fallacy for IRR and NPV. *SSRN Electronic Journal.* **10** pp. 2139 (2017)

Edited by: Zhengyi Chai
Special issue on: Data-Driven Optimization Algorithms for Sustainable and Smart City
Received: Nov 15, 2023
Accepted: Jan 8, 2024