Design of Intelligent Financial Data Management System Based on Higher-Order Hybrid Clustering Algorithm

Haitao Lu Corresp. 1

1 Department of Accounting, Henan Institute of Economics and Trade, Zhengzhou, China

Corresponding Author: Haitao Lu
Email address: lu001710@163.com

Amid the ever-expanding landscape of financial data, the importance of predicting potential risks through artificial intelligence methodologies has steadily risen. To achieve prudent financial data management, this manuscript delves into the domain of intelligent financial risk forecasting within the scope of system design. It presents a data model based on the variational encoder (VAE) enhanced with an attention mechanism, meticulously tailored for forecasting a company's financial peril. The framework embarks on its journey by encoding and enhancing multidimensional data through VAE. It then employs the attention mechanism to enrich the outputs of the VAE network, thereby demonstrating the apex of the model's clustering capabilities. In the experimentation, we subject the model to a battery of training tests using diverse datasets. The results conspicuously highlight the model's commendable performance in comparison to publicly available datasets, surpassing numerous deep clustering networks at this juncture. In the realm of financial data, the ATT-VAE model, as presented within this treatise, achieves a clustering accuracy index exceeding 0.7, a feat demonstrably superior to its counterparts in the realm of deep clustering networks. The method outlined herein provides algorithmic foundations and serves as a pivotal reference for the prospective domain of intelligent financial data governance and scrutiny.

Design of Intelligent Financial Data Management System Based on Higher-Order Hybrid Clustering Algorithm

Haitao Lu*

Department of Accounting, Henan Institute of Economics and Trade, Zhengdong New District 450018, Zhengzhou, Henan, China.

Corresponding Author:
Haitao Lu
No.58, Hubei Road, Longzi University Park, Longzi Lake, Zhengzhou
Email address: lu001710@163.com

Abstract

Amid the ever-expanding landscape of financial data, the importance of predicting potential risks through artificial intelligence methodologies has steadily risen. To achieve prudent financial data management, this manuscript delves into the domain of intelligent financial risk forecasting within the scope of system design. It presents a data model based on the variational encoder (VAE) enhanced with an attention mechanism, meticulously tailored for forecasting a company's financial peril. The framework embarks on its journey by encoding and enhancing multidimensional data through VAE. It then employs the attention mechanism to enrich the outputs of the VAE network, thereby demonstrating the apex of the model's clustering capabilities. In the experimentation, we subject the model to a battery of training tests using diverse datasets. The results conspicuously highlight the model's commendable performance in comparison to publicly available datasets, surpassing numerous deep clustering networks at this juncture. In the realm of financial data, the ATT-VAE model, as presented within this treatise, achieves a clustering accuracy index exceeding 0.7, a feat demonstrably superior to its counterparts in the realm of deep clustering networks. The method outlined herein provides algorithmic foundations and serves as a pivotal reference for the prospective domain of intelligent financial data governance and scrutiny.

Keywords: CNN; VAE; Attention Mechanism; deep clustering; Finance risk prediction.

1 Introduction

With the continuous development of the global economy, refined quantitative analysis of finance has become an indispensable part of decision-making for enterprises and investors. Financial analysis can help investors and clients evaluate the financial health of a company, thereby completing corresponding tasks from a more scientific perspective. Traditional financial analysis methods include financial ratios, financial statement analysis, trend analysis, and comparative analysis. In the study of financial information, the focus needs to monitor the financial risks of companies. Once financial risks occur, especially for listed companies, it will bring huge losses or even bankruptcy to the enterprise. Financial risks may also lead to deterioration of the company's operating conditions, decline in stock prices, loss of investor confidence, and other issues, thereby affecting the long-term development and interests of the company. Therefore, monitoring the financial condition of a company and evaluating its financial health through more intelligent means has become a research focus in the current financial and financial data management fields. [2] .

The types of financial data are complex, especially for large listed companies, which have a massive amount of financial data, such as the company's profits, cash flow, liabilities, and revenue on the same day. In addition to these numerical types of data, it usually includes text and chart forms of data. Therefore, leveraging the data processing advantages of machine learning and deep learning methods to achieve multimodal data fusion is very important for intelligent management of financial data. The task of financial analysis is to identify financial risks through the clustering method of multimodal data, which is a clustering problem. Traditional machine learning methods typically rely on manually designed features and have limited ability to handle massive amounts of data. Some traditional clustering algorithms, such as K-means clustering, are very sensitive to the selection of initial cluster centers. Different initial values may lead to different clustering results, and it is necessary to run the algorithm multiple times to obtain stable results. If financial analysis involves textual data, traditional machine learning methods may not be as flexible in processing and analyzing textual features as deep learning methods.

Deep learning models are usually more suitable for situations with complex data modalities and diverse properties, such as auto-encoder. Autoencoder is an unsupervised learning model commonly used for feature learning and data dimensionality reduction [4]. Deep models such as CNN, LSTM, and other modeling methods that excel in processing time series data [5] can capture implicit features in time series, thereby completing tasks such as data classification, regression, and prediction. These methods are used in fields such as stock prediction and futures analysis in the financial field. For the design of an intelligent financial system data system, in addition to ensuring basic data storage and visualization functions, it also requires a certain level of intelligent decision-making ability. Using deep learning methods to identify financial risks based on multimodal data is crucial for the construction of financial systems. This article proposes a clustering model based on deep networks for intelligent management of financial data, aiming to achieve intelligent clustering of financial data risks. The specific contributions are as follows::

(1) Synergizing Multiple Techniques: Our method amalgamates the Variational Autoencoder (VAE) and the attention mechanism, facilitating multi-dimensional data

clustering and enhancing financial data analysis. The combination of these techniques maximizes the performance of each method and improves data clustering capabilities.

(2) Holistic Financial Data Analysis: Our approach not only enables high-quality clustering analysis of financial data but also supports intelligent financial system management, substantially reducing human intervention for sustainable development.

(3) Model Validation: We validate our approach against established clustering databases and benchmark it against common deep clustering network methods. The experimental results demonstrate the superior performance of the proposed ATT-VAE method in analyzing financial risk data.

The subsequent sections of this paper are organized as follows: Section 2 presents related work, while Section 3 introduces the proposed methods, VAE, and the Attention mechanism. Section 4 provides details about the experiments, and Section 5 offers conclusions..

2 Related works

2.1 Intelligent management of financial data

The realm of financial risk early warning holds a pivotal standing within the purview of financial management and investment decision-making. As the sands of time have drifted by, the focus of inquiry and the arsenal of analytical tools have undergone a progressive evolution. It has transitioned from the examination of a scant number of company samples in its nascent stages to the comprehensive scrutiny of a multitude of companies replete with complete datasets. Concurrently, the means of analysis has advanced from the rudimentary utilization of financial indicator ratios to the construction of secondary indicators. It has further embraced the infusion of mathematical and statistical methodologies to give rise to multivariate discriminant analysis models, epitomized by the Z models. Presently, the landscape resonates with the widespread adoption of diverse machine learning and deep learning models [6]. Ohlson's seminal work, expounded in [7], unveiled a Logistic-based early warning model. This model, distilled through the meticulous examination of over 2,000 solvent and insolvent companies, aspires to prognosticate the probability distribution of corporate bankruptcy. Shin and colleagues [8] embarked on a pioneering exploration into the application of support vector machine models within the sphere of machine learning for financial risk prognosis. Empirical evidence underscored the supremacy of the SVM model over traditional multivariate discriminant analysis and Logit models in the realm of financial risk assessment. With the burgeoning computational prowess, the mantle was passed to neural network models and CNN models, both of which garnered resounding success within this domain [9]. Odom et al [10] championed a financial risk early warning model leveraging artificial neural networks. They painstakingly assembled a dataset teeming with an equitable number of precarious and stable companies for in-depth analysis, ultimately yielding commendable predictive outcomes. Marcano et al ventured into the realm of meta plasticity neural networks (AMMLP) and engineered an enhanced ANN model for credit default risk evaluation [11].

Furthermore, Hosaka introduced a novel approach, transmuting the financial indicators of insolvent companies into grayscale images. Subsequently, they harnessed a CNN deep learning model to prognosticate corporate bankruptcy risks with promising outcomes [12].

2.2 Research on Traditional Clustering Methods

Traditional clustering algorithms, often referred to as early clustering algorithms, exhibit commendable performance on small-scale, low-dimensional datasets. Their development has reached a mature stage, owing to their intuitively comprehensible principles and straightforward implementation. These methods find extensive utility across various domains, particularly in the realm of image processing. Prominent among these traditional clustering algorithms are the K-Means algorithm and the spectral clustering algorithm, among others. Typically, these algorithms take as input a data matrix composed of image or text features, employing diverse clustering strategies to gauge the similarity relationships among these features and subsequently generating clustering outcomes [13].The K-Means clustering algorithm halts when it attains a local optimum solution. Notably, it is tailored for numerical data clustering. It boasts the virtues of simplicity and efficiency, characterized by low algorithmic complexity. However, it bears certain drawbacks, such as sensitivity to predetermined values for the number of clusters, vulnerability to noise and outliers, and suboptimal performance on datasets with non-spherical clusters. In an endeavor to enhance the K-Means algorithm's performance, the DIANA split hierarchical clustering algorithm was introduced, treating the provided data as a cluster structure and progressively partitioning the most recently formed cluster into smaller clusters based on cluster diameter or average dissimilarity [14].

Guha et al. introduced the CURE algorithm, an improved hierarchical clustering method that leverages a representative subset of points to depict a cluster, departing from the conventional approach of using all points or a single center of mass. This modification renders it more resilient to isolated points and equips it to identify clusters characterized by complex shapes and varying sizes [15]. The incorporation of fuzzy set theory into hard clustering algorithms, assigning each sample a certain probability of belonging to a particular class, has birthed fuzzy clustering algorithms [16]. Krinidis et al. contributed a robust C-mean clustering algorithm for fuzzy local information, introducing a fuzzy local neighborhood factor to amalgamate local spatial and grayscale information, thereby diminishing the clustering method's sensitivity to noise [17].
Karlekar et al. introduced a fuzzy clustering technique employing nonlinear distances, substituting s-distance for the Euclidean distance metric, resulting in more robust natural clustering outcomes [18]. Beyond K-Means approaches, spectral clustering and its derivatives have garnered significant traction in contemporary clustering methodologies. Wang et al. introduced constrained spectral clustering, augmenting spectral clustering with additional sub-information to bolster clustering

results. By leveraging pairwise constraints, this approach tackles challenging segmentation tasks by determining whether two points are linked based on the introduced edge information [19]. Chen et al. proposed a parallel spectral clustering approach for deployment in distributed systems [20]. They juxtaposed two methods for approximating dense matrices to address concerns related to memory consumption and computational time scalability in spectral clustering. Ultimately, they opted to retain the nearest neighbors to sparsify the matrix, a strategy applicable to solving problems within distributed systems. Traditional clustering methodologies can also harness representation learning techniques for feature extraction, including subspace representation learning and deep network representation learning. This circumvents the limitations of conventional methods when confronted with high-dimensional data.

2.3 Research on clustering algorithms based on deep learning

Deep learning based clustering methods are categorized into generative model based methods and discriminative model based methods based on the nature of the network model and the results of these two types of methods can be subdivided again as shown in Figure 1:

Figure 1 The deep clustering methods

Deep clustering methods encompass a diverse array of techniques, including those predicated on VAE, GAN, intricate deep models, and GNN. The architecture of autoencoder-based clustering methods typically comprises two fundamental components: the autoencoder module and the similarity measure module [21]. Various methods adopt distinct training strategies, with one of the pioneering approaches being the inception of deep embedding clustering [22]. This method disentangles the dimensionality reduction process from the similarity metric computation within the framework. It commences with the acquisition of a proficient encoder model via self-encoder training, subsequently proceeding to joint training of the encoder and similarity metric modules. However, this approach renders the embedded features overly reliant on the initialized encoder model, which can exert an impact on clustering outcomes. Building upon this foundation, a fusion between the traditional K-means method and deep clustering methodology was realized, culminating in the co-optimization of dimensionality reduction and similarity metrics, thereby yielding more optimized results [23]. Presently, deep learning-based clustering methodologies predominantly hinge on similarity metrics. In pursuit of neighborhood relationships, these methods typically employ local constraints during the similarity metric computation process. While local constraints effectively ascertain the similarity of points within clusters, they may falter in precisely distinguishing the class attributes of points positioned at the cluster periphery. Consequently, this can result in indistinct cluster boundaries within the feature space [24]. Moreover, approaches grounded in self-coder models can also be amalgamated with spectral clustering, subspace clustering, and other techniques. The choice of neural networks for encoding can significantly impact outcomes, with convolutional neural networks often outperforming fully connected neural networks [25-26].

The evolution of segment clustering research clearly demonstrates that contemporary deep learning methods offer superior practical utility when juxtaposed with traditional clustering approaches. Furthermore, the amalgamation of VAE with CNN and other techniques augments the robustness of self-supervised and semi-supervised models. Given the sheer volume of financial data, along with the challenges of missing data and the impracticality of manual labeling, the adoption of advanced deep clustering methods assumes paramount significance in the domain of financial system management. Consequently, this manuscript proffers a novel proposition: the enhancement of unsupervised analysis of financial data through the synergistic integration of VAE and existing CNN methodologies.

3 Methodology

CNN is an influential deep learning model extensively deployed in computer vision tasks, encompassing image classification, target detection, and image segmentation. The watershed moment for CNNs occurred in 2012 with the advent of the AlexNet network [27], which solidified the standing of convolutional neural networks in the domain of deep learning. In this study, we opt to employ AlexNet for data analysis. In addition to the convolution operation outlined in equation (1), AlexNet augments network generalization performance by incorporating local normalization (LRN). The LRN is computed as depicted in equation (2):

$$(I * K)(x, y) = \sum_{i=-a}^{a} \sum_{j=-b}^{b} I(x+i, y+j) \cdot K(i, j) \qquad (1)$$

$$R(x, y, k) = \frac{A(x, y, k)}{\left(\kappa + \alpha \sum_{i=\max(0, k-\frac{n}{2})}^{\min(N-1, k+\frac{n}{2})} (A(x, y, i))^2 \right)^{\beta}} \qquad (2)$$

In equation (1), I represents the input data, K is the convolution kernel, i,j are the coordinates of the convolution kernel, and a and b are the radii of the convolution kernel. For Alexnet's special link LRN, R in Eq. (2) represents the output response, A is the original unnormalized response, and the rest of the adjustable parameters are used to control the degree of normalization and parameters such as the window and the number of channels. In addition to this, models such as ResNet, GoogleNet, etc. are widely used methods in CNN-like networks [28]. The clustering ability of the model can be greatly enhanced by augmenting the current information through convolutional neural networks.

3.2 Self-Encoder (AE) Models and Variational Self-Encoders (VAE)

AE and VAE both belong to the realm of unsupervised learning models that find prominent utility in deep learning for uncovering latent data representations. They serve a multitude of

purposes, including feature extraction, data compression, and dimensionality reduction. AE constitutes a neural network architecture comprising two principal components: an encoder and a decoder. The encoder serves to map input data into a lower-dimensional representation, while the decoder endeavors to reconstruct this lower-dimensional representation back into the original input data. The primary objective of AE is twofold: to achieve accurate reconstruction of the input data and to distill essential features of the input data within the low-dimensional representation generated by the encoder. This is illustrated in Figure 2:

Figure 2 Structure of self-encoder

For encoding and decoding of the encoder is done through function mapping, after completion, the model training needs to be realized through the definition of the loss function and objective function, which are defined as shown in equations (3) and (4):

$$L(x,x') = \frac{1}{n}\sum_{i=1}^{n}(x_i - x_i')^2 \qquad (3)$$

Where:.$\mathrm{L}(\mathrm{x},\mathrm{x}')$ denotes the loss function, representing the original input x and reconstructed input x' mean square error between the original and reconstructed inputs.

The goal of this loss function is to minimize the difference between the reconstructed data and the original input, allowing the AE to learn a valid representation of the data. However, depending on the specific task and data type, other loss functions can be chosen, such as the cross-hashing loss. The objective function is then expressed by equation (4).

$$\Theta^* = \arg\min_{\Theta}\frac{1}{N}\sum_{i=1}^{N}L(x^{(i)},x^{(i)}) \qquad (4)$$

Where:Θ^* denotes the model parameters to be optimized to minimize the loss function.N denotes the number of samples in the training dataset.$\mathrm{L}(\mathrm{x}^{(\mathrm{i})},\mathrm{x}'^{(\mathrm{i})})$ denotes the loss function that measures the number of samples in the first i original input of the first training sample $\mathrm{x}^{(\mathrm{i})}$ and reconstructed input $\mathrm{x}'^{(\mathrm{i})}$ between the original and reconstructed inputs of the first training sample, and the mean square error or other loss measures. The optimal output of the model can be obtained by optimizing the objective function and the loss function.

3.3 Overall framework of the attention-based VAE

VAE has the following advantages over ordinary AE: a VAE is a generative model that learns valid data representations and generates new samples; its latent space is continuous and interpretable, allowing operations such as interpolation, sampling, and so on, to generate diverse samples; the latent representations are more easily interpretable, which helps with a variety of downstream tasks; and the generated samples are typically of higher quality because the VAE generates samples by learning the latent distributions, rather than just replicating the training data points. The structure of the VAE is shown in Figure 3:

Figure 3 The framework for the VAE

Figure 3 underscores the key distinction between VAE and AE. In VAE, an additional statistical module comes into play, which corresponds to the encoder component of the VAE, while the generator aligns with the decoder, symmetrically positioned with respect to the encoder. The encoder undertakes the computation of mean and variance for each input, assigning a normal distribution to each input data point. It's essential to ensure that the variance in this normal distribution is not zero, as a zero variance would lead to a loss of randomness, making it challenging for the decoder to effectively reconstruct the samples in the presence of noise. During the sampling process, as sampling itself is a non-differentiable operation, the sampled result is not directly amenable to gradient-based optimization. To circumvent this issue, the re-parameterization technique is employed, allowing for the design of a differentiable sampling operation. This enables the optimization of the mean-variance model in reverse. The probability distribution of the encoder is encapsulated in equation (5).

$$q_{\Theta}(z|x) = \mathcal{N}\left(\mu_{\Theta}(x),\sigma_{\Theta}(x)^2\right) \qquad (5)$$

where Θ denotes the parameters of the encoder, and $\mu_{\Theta}(x)$ and $\sigma_{\Theta}(x)$ denote the mean and standard deviation, respectively. Latent space sampling.

$$z = \mu_{\Theta}(x) + \sigma_{\Theta}(x) \cdot \epsilon \qquad (6)$$

where ϵ is the random noise sampled from the standard normal distribution. The normal distribution feature is also introduced in the decoder section with the conditions shown in equation (7):

$$p_{\Phi}(x|z) \qquad (7)$$

This means that given the potential variables z that generates the data in the case of x of the conditional distribution. The parameters of the decoder are denoted by Φ.

In this paper, in order to enhance the model performance, we add Attention Mechanism, which is a technique widely used in deep learning to enhance the neural network's attention to certain parts of the input data, thus improving the model performance. In Attention Mechanism, there are usually three key components, Query, Key and Value features and the model enhancement is achieved by Attention Score, Attention Weight and Weighted Sum, Q denotes the query and K denotes the key, then the Attention Score can be expressed as (8)-(10):

$$\text{Attention}\ (Q,K) = \frac{Q \cdot K}{\sqrt{d_k}} \qquad (8)$$

$$\text{Attention\_Weights}\ (Q,K) = \text{softmax}\ (\text{Attention}\ (Q,K)) \qquad (9)$$

$$\text{Attention\_Output}\ (Q,K,V) = \sum_i \text{Attention\_Weights}\ (Q,K)_i \cdot V_i \qquad (10)$$

where d_k denotes the dimension of the key. The core idea of the attention mechanism is to

assign the weights of the values based on the relationship between the query and the keys in order to better capture the relevant information of the input data in different tasks and different contexts. This dynamic attention mechanism enables the neural network to better handle sequential data and improves generalization ability of the model. The model based on the attention mechanism established in this paper is shown in Figure 4:

Figure 4 Framework of the proposed ATT-VAE

In Figure 4, after completing the data input, we realized the feature extraction enhancement of the data by VAE, and completed the final model optimization by using the attention mechanism on the right side to realize the clustering judgment of the system.

4 Experiment setup and Result analysis

4.1 Datasets

Considering the characteristics of the clustering method and the characteristics of the data used, the data used in this paper include the following five:

AwA (https://cvml.ist.ac.at/AwA/) has a total of 5814 instances and consists of three modalities, local self-similarity features, SIFT features and SURF features, and contains 10 clusters. Scene-15 [29] has 3000 instances and consists of three modalities consisting of LBP features, GIST features and CENTRIST features containing 15 clusters. CUB (http://www.vision.caltech.edu/visipedia/CUB-200-2011.html) contains 50 clusters totaling 2889 data instances. Two modalities consist of 1024-dimensional image features extracted by GoogleNet and 1024-dimensional corresponding text features [30]. flowers(http://www.robots.ox.ac.uk/~vgg/data/flowers/102/) contains 50 clusters totaling 3235 data instances. The two modalities consist of GoogleNet extracted 1024-dimensional image features and 1024-dimensional corresponding text features. Both image features are removed at the time of input. The specific information of the adopted dataset is shown in Table 1:

Table 1 The specification of the Dataset

4.2 Experiment details

After completing the data collection of the dataset, it is necessary to determine the relevant details of the paper, mainly including: model evaluation indexes, model training process and so on. In order to evaluate the experimental results, we adopt three evaluation indexes: Accuracy (ACC), Normalized Mutual Information (NMI) and Adjusted Rand Index (ARI), all of which are higher to indicate better performance. ARI), all of which are higher for better performance. Clustering accuracy Used to compare the clustering assignment labels with the true labels provided by the data.

$$ACC = \frac{\sum_{i=1}^{N} \delta(s_i, \mathrm{map}(c_i))}{N} \qquad (11)$$

where s_i denotes the true label of the first i true label of the first sample, and c_i denotes the

label assigned by the clustering algorithm to the i the label assigned to the first sample by the clustering algorithm, and N is the total number of data, and $map(c_i)$ Calculate c_i to s_i The mapping between δ is determined by the following formula:

$$\delta(x,y) = \begin{cases} 1 & \text{if } x = y \\ 0 & \text{otherwise} \end{cases} \quad (12)$$

NMI is defined as the following equation.

$$NMI = \frac{MI(S,C)}{\max(H(S),H(C))} \quad (13)$$

where S, C are two different labels of the same sample, i.e., the true label and the cluster assignment label, and $MI(C, C')$ The NMI results do not change depending on the arrangement of the clusters, and they are normalized to the cluster assignment labels. $H(\cdot)$ The results of NMI do not change depending on the arrangement of the clusters, they are normalized to the range of 0 for uncorrelated and 1 for perfectly correlated. $[0, 1]$ The results of NMI do not change according to the arrangement of clusters, and they are normalized to the range of 0 for no correlation and 1 for perfect correlation.

RI (Rand Index) represents the rate of correct decision making and is defined as.

$$RI = \frac{TP + TN}{TP + TN + FP + FN} \quad (14)$$

where TP is true positive ,TN is true negative, FP is the false positive and FN is the false negative.The Rand index has values between $[0, 1]$ The Rand Index (RI) has a value between 1 and 2, and the RI is 1 when the two classifications match.

Unsupervised Deep Embedding clustering (DEC) [31], improved DEC (IDEC) [32] and Deep Neural Networks for Spectral Clustering (SpectralNet) [33], Deep canonically correlated auto-encoders (DCCAE) [34] used for clustering analysis in clustering research are the more widely used methods that are more mature and represent their respective fields. methods, so this paper chooses the above methods for the comparison. After confirming the dataset and related indexes, we trained the model, and the training methods used for different models are similar, and the specific steps are shown in Algorithm1:

Algorithm 1: Training process of ATT-VAE for clustering
Input: AWA dataset, Scene dataset, CUB dataset, Flower dataset
Initialization.
Define the ATT-VAE.
Define the hyperparameters and Initialization.
Define the loss function.
Define the optimizer: Adam optimizer.
Define the number of training epochs.

Feature extraction.
Using the original features in the dataset
Pre-training: Initialize the pre-training step counter.
while pre-training step counter < pre-training steps do
    Sample a batch of data.
    Feed data to the ATT-VAE framework.
    Update model.
    Counter++;
End

Parameters tuning
Tuning counter definition TT.
while TT< Preset iteration **do**
        Feed sample data to the proposed network.
        Loss and gradients calculation.
        Model updated.
        Compute ACC, NMI and RI
        Save the optimal model
end
Output: Trained ATT-VAE

After completing the model building and training of the relevant data, we performed statistics for the model.

4.3 Experiment Result and Analysis

Based on the pertinent metrics and model training procedures elucidated in sections 4.1 and 4.2, we subjected the data to diverse test sets and shall now elucidate the detailed clustering outcomes for each dataset. Table 2 and Figure 5 encapsulate the clustering results for the AWA dataset. Notably, the introduction of the deep network has manifestly enhanced clustering performance, with the proposed method showcased in this study surpassing conventional approaches in the present stage across all three metric categories: ACC, NMI, and RI. Indeed, the proposed method outperforms common techniques across all three metric types at this juncture.

Table 2 The comparison result of three indicators concerning AWA datasets

Figure 5 The comparison result of three indicators concerning AWA datasets

After completing the analysis of the AWA data, we similarly analyzed the data on the three

datasets SCENE, CUB, and Flower, the results of which are shown in figure 6,7,8, and the corresponding data results are given accordingly in Table 3,4,5.

Table 3 The comparison result of three indicators concerning SCECE datasets

Figure 6 The comparison result of three indicators concerning SCECE datasets

Table 4 The comparison result of three indicators concerning CUB datasets

Figure 7 The comparison result of three indicators concerning CUB datasets

Table 5 The comparison result of three indicators concerning Flower datasets

Figure 8 The comparison result of three indicators concerning Flower datasets

After completing the comparison of multiple methods, this paper also carries out batch size comparison experiments of the proposed method under different datasets, which are tested through eight batch sizes ranging from 2 to 128, and the corresponding boxplots obtained are shown in Figure 9:

Figure 9 The ACC for the different datasets using different batch sizes

In Figure 9, it's evident that the variance in clustering results across different batch sizes is relatively modest. This observation underscores the inherent robustness of the method advanced in this paper.

Examining the data presented in the icon, it becomes evident that the ATT-VAE model, as introduced in this paper, boasts commendable generalization prowess and data clustering acumen. This is particularly conspicuous in the case of the CUB and Flower datasets, which feature fewer attributes and categories. In such scenarios, the clustering efficacy is notably pronounced. This attribute bodes well for the application of this method in the financial analysis of low-dimensional data characteristic of financial system analysis. As an extension of this approach, the paper now extends its ambit to the realm of economic and financial research (https://cn.gtadata.com/). Leveraging solvency, operational capacity, and profitability indicators as provided by this source, the model undertakes risk analysis through clustering, stratifying entities into high-risk and low-risk categories. This dataset is denoted as the Finance database. The results pertaining to ACC and NMI under this database are depicted in Figure 10.

Figure 10 The ACC and NMI result with the Finance datasets

Figure 10 illustrates a discernible trend: as the complexity of the deep network progressively intensifies, the clustering accuracy of the model registers a corresponding increase. Notably, the method proposed in this paper exhibits a notable advantage in terms of accuracy.

5 Discussion and Conclusion

In this study, we have introduced an innovative deep clustering paradigm utilizing ATT-VAE, with the objective of facilitating the adaptive clustering of financial perils within the realm of financial data. The model's performance has been exhaustively assessed employing common high-dimensional, low-dimensional clustering public datasets, along with real-world financial risk data distinguished by practical relevance. The experimental findings presented herein substantiate the effectiveness of the approach delineated in this manuscript. By harnessing diverse clustering methodologies and enhancing the VAE model through the integration of attention mechanisms, our method achieves significantly enhanced clustering outcomes. The approach has further demonstrated its prowess in authentic financial data assessments, boasting a clustering accuracy surpassing 70%. This accomplishment provides robust validation for future undertakings in financial analysis and data management

Anticipating forthcoming work, we envision broadening the adaptability of the model to embrace a plethora of financial data formats while refining its data processing capabilities. Furthermore, we intend to delve into advanced optimizations for the clustering network and the augmentation of model capabilities through the incorporation of techniques such as reinforcement learning. Although the VAE model employed in this investigation exhibits robust clustering capabilities, we will explore avenues to render the model more streamlined in the future. Furthermore, for financial data characterized by judiciously selected and processed features, the data analysis may potentially be carried out by a more lightweight clustering model. Thus, the extraction of features from the model proposed in this paper, as well as from related deep clustering models, constitutes a central focus for future research, with the objective of attaining more comprehensive features.=.

Data Availability

The dataset employed in this investigation is made readily available and accessible to interested parties.

Conflicts of Interest

The author declares that there are no conflicts of interest.

Funding Statement

This work received no fundings.

Acknowledgments

I would like to thank the anonymous reviewers whose comments and suggestions helped improve this manuscript.

Reference

[1] Loughran T, McDonald B. Textual analysis in accounting and finance: A survey[J]. Journal of Accounting Research, 2016, 54(4): 1187-1230.

[2] Goodell J W, Kumar S, Lim W M. Artificial intelligence and machine learning in finance: Identifying foundations, themes, and research clusters from bibliometric analysis[J]. Journal of Behavioral and Experimental Finance, 2021, 32: 100577.

[3] Goodell J W, Kumar S, Lim W M. Artificial intelligence and machine learning in finance: Identifying foundations, themes, and research clusters from bibliometric analysis[J]. Journal of Behavioral and Experimental Finance, 2021, 32: 100577.

[4] Karim M R, Beyan O, Zappa A. Deep learning-based clustering approaches for bioinformatics[J]. Briefings in bioinformatics, 2021, 22(1): 393-415.

[5] Shi X, Wang Z, Zhao H. Threshold-free phase segmentation and zero velocity detection for gait analysis using foot-mounted inertial sensors[J]. IEEE Transactions on Human-Machine Systems, 2022, 53(1): 176-186.

[6] Lin W Y, Hu Y H, Tsai C F. Machine learning in financial crisis prediction: a survey[J]. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 2011, 42(4): 421-436.

[7] Ohlson J.Financial Ratios And The Probabilistic Prediction Of Bankruptcy[J].Journal of Accounting Research,1980,18(1):109-131.

[8] Shin K S,Lee T S,Kim H J.An Application of Support Vector Machines in Bankruptcy Prediction Model[J].Expert Systems with Applications,2005,28(1):127-135.

[9] Atiya A F.Bankruptcy prediction for credit risk using neural networks:A survey and new results[J].IEEE Transactions on Neural Networks,2001,12(4):929-929.

[10] M.D.Odom,Sharda R.A neural network model for bankruptcy prediction[C]//Neural Networks,1990.1990 IJCNN International Joint Conference on.IEEE,2012.

[11] Marcano-Cedeno A, Marin-De-La-Barcena A, Jiménez-Trillo J, Pinuela J A, Andina D. Artificial metaplasticity neural network applied to credit scoring[J]. International journal of neural systems, 2011, 21(04): 311-317.

[12] Hosaka T. Bankruptcy prediction using imaged financial ratios and convolutional neural networks[J]. Expert systems with applications, 2019, 117: 287-299.

[13] Ahmed M, Seraj R, Islam S M S. The k-means algorithm: A comprehensive survey and performance evaluation[J]. Electronics, 2020, 9(8): 1295.
[14] Patnaik A K, Bhuyan P K, Rao K V K. Divisive Analysis (DIANA) of hierarchical clustering and GPS data for level of service criteria of urban streets[J]. Alexandria Engineering Journal, 2016, 55(1): 407-418.
[15] Guha S,Rastogi R,Shim K.Cure:An efficient clustering algorithm for large databases[J].ACM Sigmod record,1998,27(2):73–84.
[16] Nayak J, Naik B, Behera H S. Fuzzy C-means (FCM) clustering algorithm: a decade review from 2000 to 2014[C]//Computational Intelligence in Data Mining-Volume 2: Proceedings of the International Conference on CIDM, 20-21 December 2014. Springer India, 2015: 133-149.
[17] Krinidis S, Chatzis V. A robust fuzzy local information C-means clustering algorithm[J]. IEEE transactions on image processing, 2010, 19(5): 1328-1337.
[18] Seal A, Karlekar A, Krejcar O. Fuzzy c-means clustering using Jeffreys-divergence based similarity measure[J]. Applied Soft Computing, 2020, 88: 106016.
[19] Wang X, Qian B, Davidson I. On constrained spectral clustering and its applications[J]. Data Mining and knowledge discovery, 2014, 28: 1-30.
[20] Chen W Y, Song Y, Bai H. Parallel spectral clustering in distributed systems[J]. IEEE transactions on pattern analysis and machine intelligence, 2010, 33(3): 568-586.
[21] Song C, Liu F, Huang Y. Auto-encoder based data clustering[C]//Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications: 18th Iberoamerican Congress, CIARP 2013, Havana, Cuba, November 20-23, 2013, Proceedings, Part I 18. Springer Berlin Heidelberg, 2013: 117-124.
[22] Xie J, Girshick R, Farhadi A. Unsupervised deep embedding for clustering analysis[C]//International conference on machine learning. PMLR, 2016: 478-487.
[23] Yang B, Fu X, Sidiropoulos N D. Towards k-means-friendly spaces: Simultaneous deep learning and clustering[C]//international conference on machine learning. PMLR, 2017: 3861-3870.
[24] Min E, Guo X, Liu Q. A survey of clustering with deep learning: From the perspective of network architecture[J]. IEEE Access, 2018, 6: 39501-39514.
[25] Santhosh K K, Dogra D P, Roy P P. Vehicular trajectory classification and traffic anomaly detection in videos using a hybrid CNN-VAE Architecture[J]. IEEE Transactions on Intelligent Transportation Systems, 2021, 23(8): 11891-11902.
[26] Xu C, Zhang Y, Chen H, Dong L, Wang W. A fairness-aware graph contrastive learning recommender framework for social tagging systems, Information Sciences, 640, 119064, 2023.
[27] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[J]. Advances in neural information processing systems, 2012, 25.
[28] Targ S, Almeida D, Lyman K. Resnet in resnet: Generalizing residual architectures[J]. arXiv preprint arXiv:1603.08029, 2016.
[29] Fei-Fei L, Perona P. A bayesian hierarchical model for learning natural scene

categories[C]//2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05). IEEE, 2005, 2: 524-531.

[30] Reed S, Akata Z, Lee H. Learning deep representations of fine-grained visual descriptions[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 49-58.

[31] Xie J, Girshick R, Farhadi A. Unsupervised deep embedding for clustering analysis[C]//International conference on machine learning. PMLR, 2016: 478-487.

[32] Guo X, Gao L, Liu X. Improved deep embedded clustering with local structure preservation[C]// International Joint Conferences on Artificial Intelligence. 2017: 1753-1759.

[33] Shaham U, Stanton K, Li H. Spectralnet: Spectral clustering using deep neural networks[J]. arXiv preprint arXiv:1801.01587, 2018.

[34] Wang W, Arora R, Livescu K. On deep multi-view representation learning[C]//International conference on machine learning. PMLR, 2015: 1083-1092.

Figure 1

Figure 1 The deep clustering methods

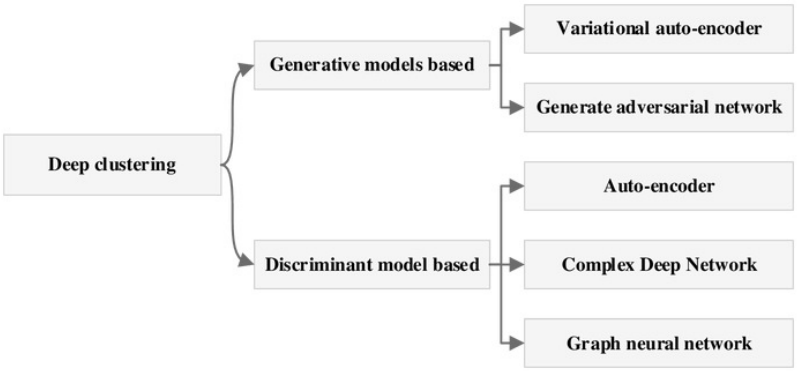
Deep clustering
Generative models based
Variational auto-encoder
Generate adversarial network
Discriminant model based
Auto-encoder
Complex Deep Network
Graph neural network

Figure 2

Figure 2 Structure of self-encoder

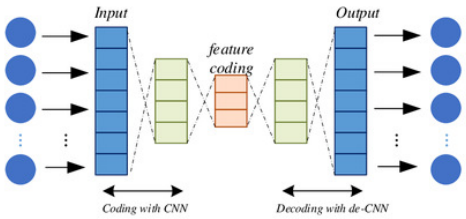
Input
feature coding
Output
Coding with CNN
Decoding with de-CNN

Figure 3

Figure 3 The framework for the VAE

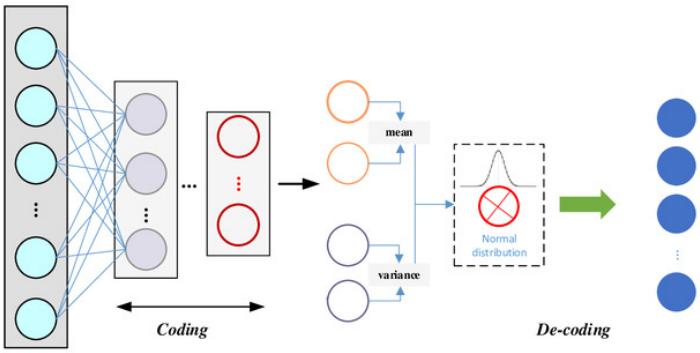
mean
variance
Normal distribution
Coding
De-coding

Figure 4

Figure 4 Framework of the proposed ATT-VAE

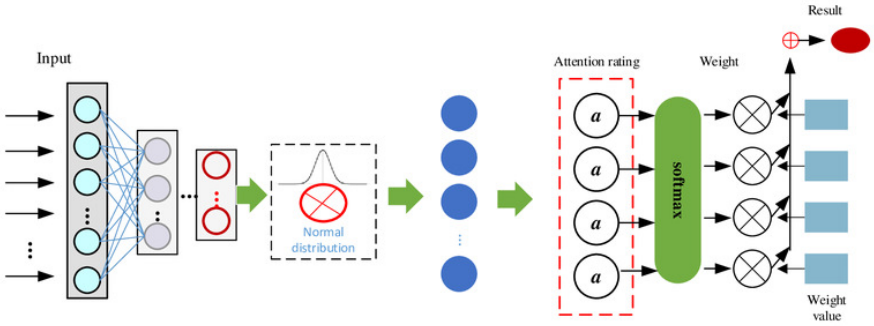
Input
Normal distribution
Attention rating
Weight
Result
a
a
a
a
softmax
Weight value

Figure 5

Figure 5 The comparison result of three indicators concerning AWA datasets

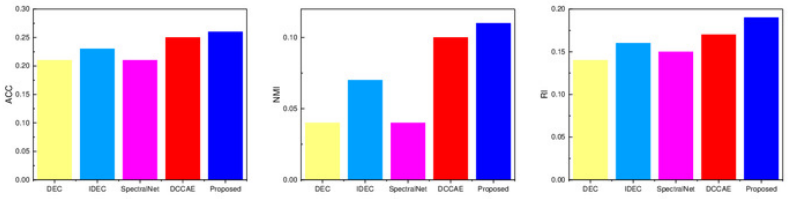
ACC
NMI
RI
DEC
IDEC
SpectralNet
DCCAE
Proposed

Figure 6

Figure 6 The comparison result of three indicators concerning SCECE datasets

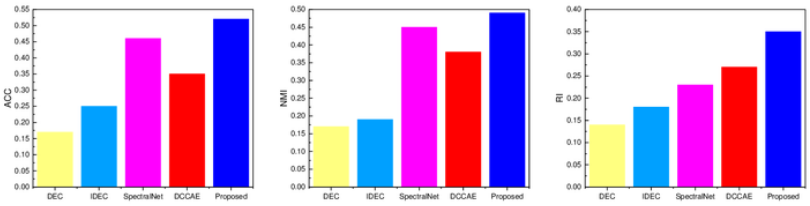
ACC
NMI
RI
DEC
IDEC
SpectralNet
DCCAE
Proposed

Figure 7

Figure 7 The comparison result of three indicators concerning CUB datasets

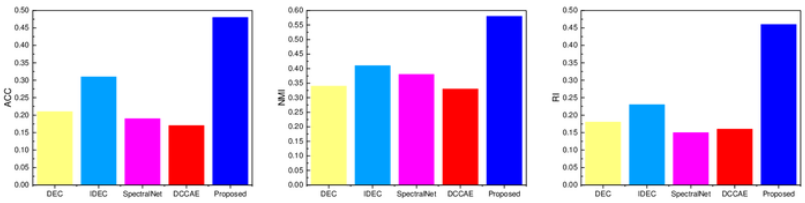
ACC
NMI
RI
DEC
IDEC
SpectralNet
DCCAE
Proposed

Figure 8

Figure 8 The comparison result of three indicators concerning Flower datasets

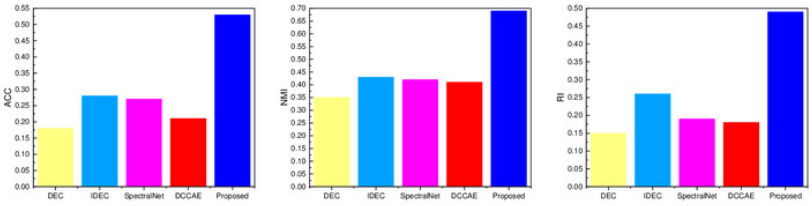
ACC
NMI
RI
DEC
IDEC
SpectralNet
DCCAE
Proposed

Figure 9

Figure 9 The ACC for the different datasets using different batch sizes

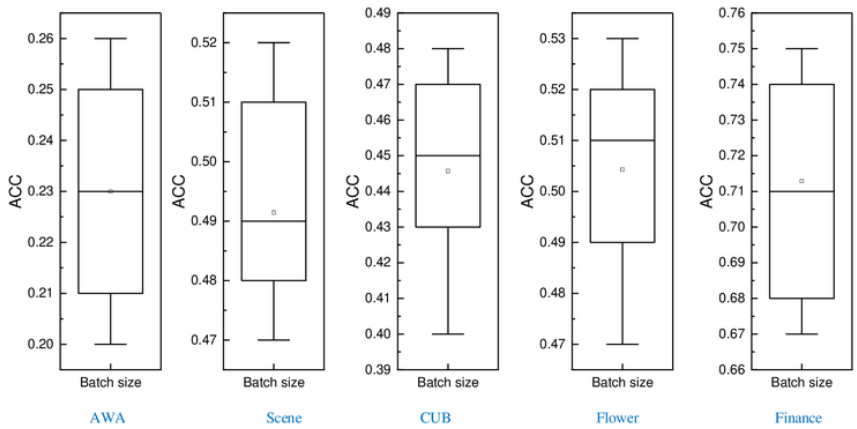
ACC
Batch size
AWA
Scene
CUB
Flower
Finance

Figure 10

Figure 10 The ACC and NMI result with the Finance datasets

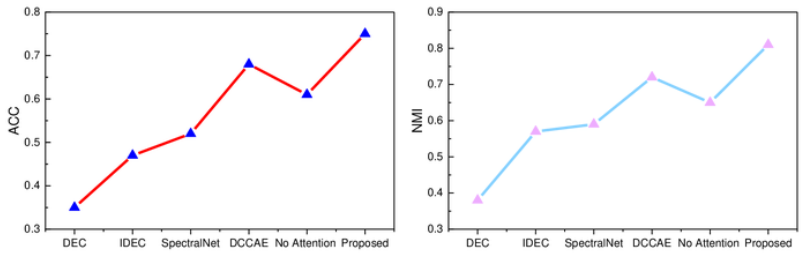
ACC
0.8
0.7
0.6
0.5
0.4
0.3
DEC
IDEC
SpectralNet
DCCAE
No Attention
Proposed
NMI
0.9
0.8
0.7
0.6
0.5
0.4
0.3
DEC
IDEC
SpectralNet
DCCAE
No Attention
Proposed

Table 1(on next page)

Table 1 The specification of the Dataset

Dataset	Modal	Samples	Cluster
AwA	3	5814	10
Scene	3	3000	3
CUB	2	2889	2
Flower	2	3235	2

1

Table 2(on next page)

Table 2 The comparison result of three indicators concerning AWA datasets

	DEC	IDEC	SpectralNet	DCCAE	Proposed
ACC	0.21	0.23	0.21	0.25	0.26
NMI	0.04	0.07	0.04	0.1	0.11
RI	0.14	0.16	0.15	0.17	0.19

1

Table 3(on next page)

Table 2 The comparison result of three indicators concerning AWA datasets

	DEC	IDEC	SpectralNet	DCCAE	Proposed
ACC	0.17	0.25	0.46	0.35	0.52
NMI	0.17	0.19	0.45	0.38	0.49
RI	0.14	0.18	0.23	0.27	0.35

1

Table 4(on next page)

Table 4 The comparison result of three indicators concerning CUB datasets

	DEC	IDEC	SpectralNet	DCCAE	Proposed
ACC	0.21	0.31	0.19	0.17	0.48
NMI	0.34	0.41	0.38	0.33	0.58
RI	0.18	0.23	0.15	0.16	0.46

1

Table 5(on next page)

Table 5 The comparison result of three indicators concerning Flower datasets

	DEC	IDEC	SpectralNet	DCCAE	Proposed
ACC	0.18	0.28	0.27	0.21	0.53
NMI	0.35	0.43	0.42	0.41	0.69
RI	0.15	0.26	0.19	0.18	0.49

1