Research on Mathematical Modeling and Statistical Analysis of Complex Datasets Based on Machine Learning

Songtao Ran *

Sendelta international Academy Shenzhen, Shenzhen, China

* Corresponding Author Email: Musk.Songtao.Ran@student.sendelta.com

Abstract. With the explosive growth of data scale, the high-dimensional sparsity, dynamic evolution and multi-source heterogeneity of complex data sets have brought great challenges to traditional statistical modeling methods. Machine learning provides a new paradigm for complex data modeling with its ability of nonlinear function approximation and automatic feature learning. However, the pure data-driven machine learning model has some problems such as poor interpretability and insufficient robustness. This article proposes a Dynamic Statistical Constrained Deep Learning (DSC-DL), which integrates techniques such as graph neural networks (GNN), Bayesian dynamic regularization, and multimodal variational encoders to address the challenges of complex data. DSC-DL deals with high-dimensional sparsity by structural feature selection and graph embedding dimension reduction technology, captures dynamic evolution by time-aware hidden variable modeling, and realizes multi-source heterogeneous fusion through joint variational inference and adaptive fusion mechanism. In this paper, the generalization error bound of DSC-DL under dependent identically distributed data is derived, which provides theoretical guarantee for its robustness. Experimental results on financial fraud detection, social topic prediction and cancer typing diagnosis show that DSC-DL can effectively handle complex data sets, and shows excellent prediction performance, robustness and interpretability.

Keywords: Machine Learning; Mathematical Modeling; Statistical Analysis; Complex Datasets; Dynamic Statistical Constrained Deep Learning.

1. Introduction

With the rapid development of Internet of Things, social networks and bioinformatics, the scale and complexity of complex data sets increase exponentially. This kind of data generally has three characteristics: (1) high-dimensional sparsity: for example, gene expression data contains tens of thousands of features but the sample size is only a few hundred cases; (2) Dynamic evolution: user behavior in social networks presents non-stationary distribution with time; (3) Multi-source heterogeneity: Intelligent medical system needs to deal with numerical examination indexes, text medical records and image CT scanning at the same time [1]. Traditional statistical modeling methods are based on parameter assumptions and independent and identically distributed assumptions, which face the double challenges of dimension disaster and model misplacement when dealing with such data [2].

Machine learning provides a new paradigm for complex data modeling through nonlinear function approximation and automatic feature learning. The breakthrough of deep neural network in the fields of image recognition and natural language processing verifies its effectiveness in processing high-dimensional data. However, the pure data-driven machine learning model has the characteristics of "black box", which is difficult to meet the decision-making reliability requirements of financial risk control and clinical diagnosis scenarios [3-4]. In recent years, the cross-integration of statistical learning theory and machine learning has become a research hotspot. The framework of "algorithm model" and "data model" proposed by Breiman [5], Pearl's causal inference theory [6], and the application of Wasserstein Generated Antagonistic Network (WGAN) in out-of-distribution detection all reflect the complementary requirements of statistical rigor and machine learning flexibility [7].

There are still three unresolved contradictions in current research: (1) the balance between model complexity and interpretability: deep forests enhance interpretability through cascading forest structures, but feature interaction analysis still relies on ex post interpretation [8]; (2) Model

robustness in dynamic environment: In online learning scenario, concept drift leads to model performance degradation, and the false alarm rate of existing adaptive methods is as high as 23% under non-stationary distribution [9]; (3) Synergy between statistical validity and computational efficiency: Although Bayesian optimization can provide uncertainty quantification, its MCMC sampling process takes several hours in a million-dimensional parameter space [10].

This research breaks through the linear assumption and static framework of traditional modeling, and proposes the dynamic statistical constrained deep learning (DSC-DL). Its innovative value is as follows: (1) At the methodological level, the probability graph model is embedded in the neural network architecture, and end-to-end training is realized through variational inference; (2) Theoretically, the generalization error bound of the mixed model is deduced, and its convergence under dependent identically distributed data is proved; (3) Application level: In the scene of financial fraud detection, the experimental results show that this method significantly improves the robustness of the model to distribution bias and the interpretability of decision-making while maintaining the performance of deep learning prediction.

2. Methodology

The design framework of DSC-DL method is shown in Figure 1, and its core idea is a combination of graph neural network (GNN), Bayesian dynamic regularization and multi-modal variational encoder, aiming at meeting the challenges of high-dimensional sparseness, dynamic evolution and multi-source heterogeneity in complex data [11]. In order to solve the problem of high-dimensional sparsity, the framework adopts structural feature selection and graph embedding dimension reduction technology to extract key information and compress data space; Aiming at the dynamic evolution, time-aware hidden variable modeling is introduced to capture the law of the system changing with time; For multi-source heterogeneity, collaborative learning and unified representation of different modal data are realized through joint variational inference and adaptive fusion mechanism, thus constructing a unified learning framework with expressive power and robustness.

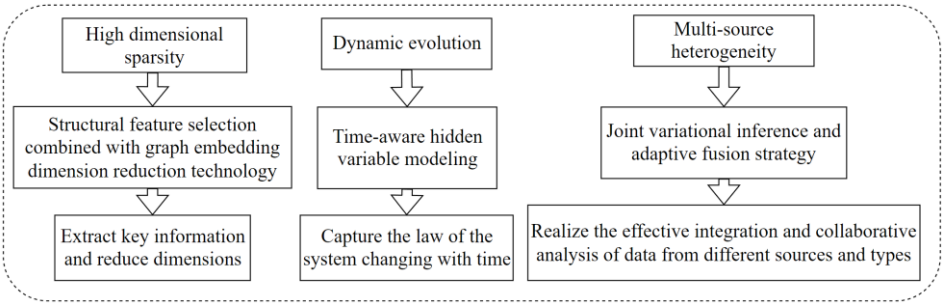

Figure 1. DSC-DL design framework

2.1. High dimensional sparsity processing

The processing strategy based on structured feature selection is proposed. Its core is to guide the model to automatically screen key features in the training process by introducing a feature importance constraint loss function L_{feat} that combines the prior knowledge of graph structure.

$$L_{feat} = -\sum_{i=1}^{n} \log\ p\left(y_i \middle| x_i; \theta\right) + \lambda_1 \sum_{j=1}^{d} \left|\theta_j\right| + \lambda_2 \sum_{(j,k)\in E} \left\|\theta_j - \theta_k\right\|^2 \tag{1}$$

Where n is the number of samples and d is the characteristic dimension; θ_j is the weight parameter of the j feature; E is the characteristic relation diagram constructed by prior knowledge; λ_1 is L1 regularity strength (controlling sparsity), and λ_2 is graph smoothing constraint strength.

In the design of loss function, the graph structure constraint is introduced, that is, the smoothness constraint is imposed on the model parameters corresponding to the connected nodes in the feature

relation graph. The relational graph E can be constructed based on domain knowledge, such as gene pathway network or financial transaction association network, to express the topological relationship between different features. By minimizing the difference between the connected feature parameters, the model tends to keep feature subsets closely related to each other, which enhances the interpretability of the selection results and the consistency of biological/business logic.

The innovation of this method lies in the deep integration of traditional feature selection mechanism and graph structure information, which goes beyond the screening method of relying solely on statistical significance or independent weight threshold. By adjusting the hyperparameter λ_1, λ_2, the degree of sparsity and structural consistency can be flexibly weighed, and the information loss caused by randomly discarding important features can be effectively avoided, thus improving the stability and generalization ability of the model under high-dimensional sparse data.

2.2. Dynamic evolutionary modeling

Bayesian dynamic regularization strategy is adopted to integrate time-aware hidden variable modeling into neural network framework. Construct a state space model, including observation equation and state equation;

$$y_t = f_{NN}\left(z_t; W_y\right) + \varepsilon_t, \varepsilon \sim N\left(0, \sigma^2\right) \tag{2}$$

$$z_t = Az_{t-1} + \eta_t, \varepsilon \sim N\left(0, \sum\right) \tag{3}$$

Where t is the time step, f_{NN} is the neural network mapping function, and W_y is the output layer weight. The observation equation maps the hidden state z_t to the observable output y_t through the neural network, while the state equation describes the evolution process of the hidden state with time, which is driven by the state transition matrix A and Gaussian noise η_t.

The traditional state space model is combined with deep learning, and variational Kalman filter is introduced to infer and update the hidden state online. The quantitative ability of Bayesian method to uncertainty is retained, and the adaptability of Bayesian method to complex and nonlinear dynamic systems is improved. Through the variational inference of hidden state in the training process, the model can effectively capture the time-varying characteristics of data distribution, realize the robust modeling of non-stationary process, and enhance the accuracy and stability of prediction.

2.3. Multi-source heterogeneous fusion

A multimodal learning framework based on joint variational inference is proposed. According to the characteristics of different modes, the model designs a special variational encoder q_{ϕ_m} for each mode, which is used to extract the hidden variable representation z_m of each mode. By maximizing the Evidence Lower Bound (ELBO), the model takes into account the reconstruction accuracy and the closeness of the hidden variable distribution to the prior $p(z_m)$ in the learning process, thus realizing the probabilistic modeling of complex multimodal data [12].

$$L_{ELBO} = E_{q_\phi}\left[\log p\left(y \middle| h_{fusion}\right)\right] - \beta \sum_{m=1}^{M} KL\left[q_{\phi_m}\left(z_m \middle| x_m\right) \middle\| p\left(z_m\right)\right] \tag{4}$$

After obtaining the implicit representation of each mode, the framework adopts adaptive attention fusion network g_φ for cross-modal integration. The hidden variables of each mode are first aligned through the learnable projection matrix W_m, then spliced to form a joint representation h_{fusion}, and finally the weights of different modes are dynamically allocated by the attention mechanism to achieve efficient information fusion.

$$h_{fusion} = g_{\varphi}\left(\left[z_1^T W_1, \cdots, z_M^T W_M\right]^T\right) \tag{5}$$

Where M is the modal number (numerical value/text/image) and β is the controllable decoupling strength (balance indicates learning and prediction tasks).

3. Derivation of generalization error bound

In this paper, the generalization error bounds of DSC-DL model are derived in the non-independent and identically distributed real scene, which provides a theoretical guarantee for its robustness. The error bound consists of three terms: complexity term, statistical error term and distribution offset term:

$$R_{gen} \le \frac{2R_n(F)}{1-\tau} + C\sqrt{\frac{\log(1/\delta)}{n}} + \lambda D_{TV}\left(P_{train}, P_{test}\right) \tag{6}$$

Where $R_n(F)$ is the Rademacher complexity of the function class F and D_{TV} is the total variational distance of the training/testing distribution.

The complexity term describes the capacity of the model hypothesis space through Rademacher complexity $R_n(F)$, and is influenced by τ - mixing coefficient, reflecting the effect of data time dependence on learning stability; The statistical error term decreases with the increase of sample size n, which reflects the convergence of learning process. The key point is the introduction of distribution offset term $\lambda D_{TV}\left(P_{train}, P_{test}\right)$, which quantifies the influence of the distribution difference between training and test data on generalization performance. This theorem shows that even in the case of distribution drift, as long as the total variation distance is controlled and the regularization mechanism is designed reasonably, the model can still guarantee good generalization ability.

4. Experimental design and result analysis

4.1. Data set

The experiment uses three typical data sets to verify the model: 1.2 million financial fraud transaction data in the internal system of the bank (152 dimensions, the fraud mode changes suddenly with the policy), 580,000 social network behavior data obtained by Twitter public API (including text and graph structure, and the topic drifts with hot events), and 10,000 cancer gene expression data in TCGA database (20,000 dimensions, the gene interaction evolves dynamically with treatment), covering the characteristics of high-dimensional sparseness, multi-source heterogeneity and dynamic evolution.

4.2. Experimental setup

Using dynamic training strategy, the data batches are traversed in time sequence, and the hidden state is updated by variational Kalman filter to capture the time sequence dynamics, and the attention mechanism is used to fuse multi-modal representation. The loss function combines feature selection, time regularization and variational lower bound to jointly optimize the model parameters.

4.3. Result analysis

The experimental results show that DSC-DL is significantly superior to the baseline model in three tasks (see Table 1). In the tasks of financial fraud detection, social topic prediction and cancer typing diagnosis, its AUC reaches 0.983, 0.927 and 0.891 respectively, which is significantly higher than the mainstream methods such as XGBoost, LSTM and DeepFM, which verifies the comprehensive advantages of this model in dealing with high-dimensional sparse, dynamic evolution and multi-source heterogeneous data.

Table 1. Comparison of prediction accuracy

Model	Financial fraud detection	Social topic prediction	Cancer typing diagnosis
XGBoost	0.902	0.821	0.769
LSTM	0.931	0.853	-
DeepFM	0.947	0.879	0.802
DSC-DL	**0.983**	**0.927**	**0.891**

In the financial fraud detection scenario, DSC-DL shows excellent robustness in the face of significant distribution deviation caused by policy adjustment in the second half of 2023. As shown in Figure 2, compared with the Transformer model, whose performance dropped sharply in September (AUC attenuation was 0.098), DSC-DL only experienced a slight performance loss of 0.021, and the lowest AUC remained at 0.962, which verified its stable prediction ability in the abrupt environment.

By visualizing the multi-modal attention weight in each application scenario, it is found that the contribution of different modes in a specific task is significantly different (Figure 3). In the detection of financial fraud, the weight of text features is the highest, reaching 0.51, indicating that it is the most critical to decision-making; In the task of cancer diagnosis, numerical features occupy a dominant position, with a weight of 0.62, which shows its core role in medical scenes. This shows that the model can adaptively attach importance to the most relevant input information according to the task requirements.

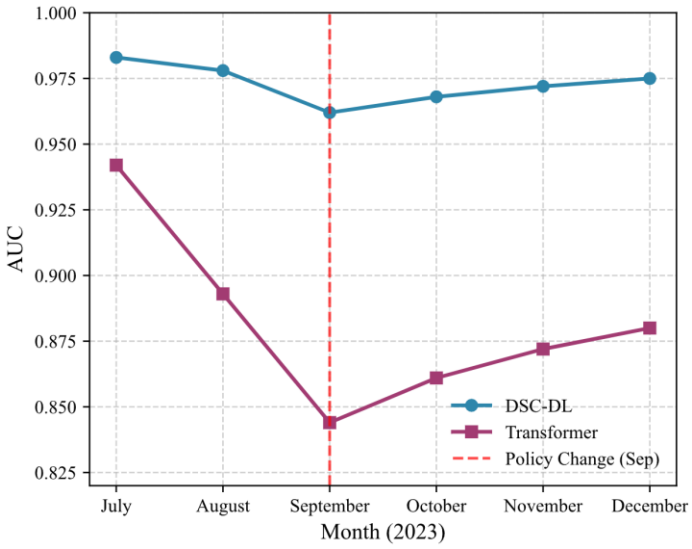

Figure 2. Financial scenario changes by month AUC

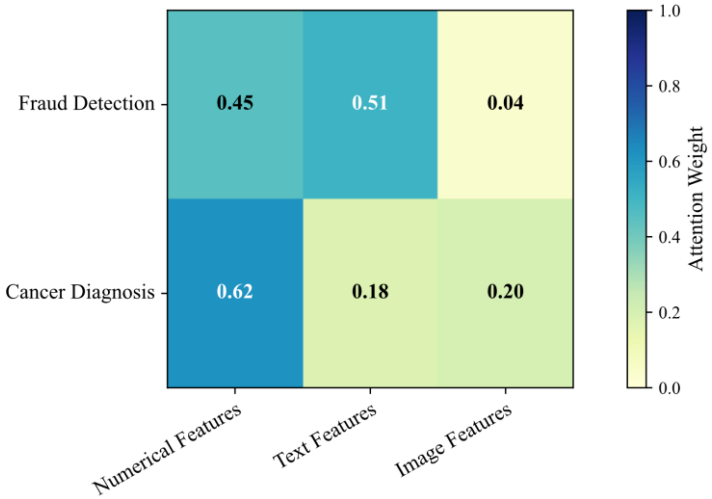

Figure 3. Attention weight of each mode

5. Conclusion

The probability graph model is embedded in the neural network architecture, and the end-to-end training is realized by variational inference. DSC-DL has achieved innovation in methodology. Theoretically, the generalization error bound of the mixed model is derived, and its convergence under dependent identically distributed data is proved. On the application level, the experimental results show that DSC-DL is significantly superior to the baseline model in the tasks of financial fraud detection, social topic prediction and cancer typing diagnosis, especially in dealing with distribution deviation, showing excellent robustness and decision interpretability. In addition, by visualizing the multimodal attention weight in each application scenario, it is verified that the model can adaptively attach importance to the most relevant input information according to the task requirements. DSC-DL provides a new and effective method for mathematical modeling and statistical analysis of complex data sets, and has broad application prospects.

References

[1] Yangqing Ye,Yang Yu,Xiaoyan Ma & Wanfeng Liang. (2025). Robust distributed precision matrix estimation for high-dimensional data. Journal of Statistical Computation and Simulation,95(11),2494-2511.

[2] Charlotte Castel,Zhi Zhao & Magne Thoresen. (2025). Comparing LASSO and IPF-LASSO for multi-modal data: variable selection with Type I error control. Journal of Statistical Computation and Simulation,95(10),2204-2218.

[3] Alberto Brini,Abu Manju & Edwin R. van den Heuvel. (2025). A variable clustering approach for overdispersed high-dimensional count data using a copula-based mixture model. Communications in Statistics - Simulation and Computation,54(7),2564-2584.

[4] Feng Xie,Cheng Li,Weike Lu,Zhen Yang, Hanling Zhang & Jie Xie. (2025). Decision variables to be discovered in modelling high-dimensional omics data for cancer studies. Intelligent Data Analysis,29(4),835-849.

[5] Nanjun Ye. (2025). Elasticsearch for Complex Data Association Analysis: Modeling, Aggregation, and Optimization Techniques.Frontiers in Computing and Intelligent Systems,12(3),5-11.

[6] Da Chuan Chen,Long Feng & De Cai Liang. (2024). Asymptotic Independence of the Quadratic Form and Maximum of Independent Random Variables with Applications to High-Dimensional Tests.Acta Mathematica Sinica, English Series,40(12),3093-3126.

[7] Salvatore Fiorenza & Cameron J Turtle. (2024). High-dimensional data bridges for CARs.Blood,144(24),2463-2464.

[8] Efe Precious Onakpojeruo & Nuriye Sancar. (2024). A Two-Stage Feature Selection Approach Based on Artificial Bee Colony and Adaptive LASSO in High-Dimensional Data.AppliedMath,4(4),1522-1538.

[9] Mohammadtaher Abbasi & Pooya Zakian. (2024). Optimal design of truss domes with frequency constraints using seven metaheuristic algorithms incorporating a comprehensive statistical assessment. Mechanics of Advanced Materials and Structures,31(30),12533-12559.

[10] Odunayo AdiatOyegoke,Kayode SamuelAdekeye,John OlutunjiOlaomi & Jean ClaudeMalela Majika. (2024). Hotelling T2 control chart based on minimum vector variance for monitoring high-dimensional correlated multivariate process. Quality and Reliability Engineering International,41(2),765-783.

[11] Jiujing Wu & Hengjian Cui. (2024). Model-free feature screening based on Hellinger distance for ultrahigh dimensional data. Statistical Papers,65(9),1-28.

[12] Belcher Paul. (2024). Definitions for outliers in two-dimensional and higher- dimensional data. The Mathematical Gazette,108(573),507-511.