International Journal of Computer Science and Information Technology
ISSN: 3005-9682 (Print), ISSN: 3005-7140 (Online) | Volume 3, Number 3, Year 2024
DOI: https://doi.org/10.62051/ijcsit.v3n3.41
Journal homepage: https://wepub.org/index.php/IJCSIT/index

Feature Engineering and Data Visualization Analysis in Artificial Intelligence in Big Data Era

Zongze Li

School of Electrical & Electronic Engineering, Nanyang Technological University, Singapore

ABSTRACT

In the environment of massive data, the selection and construction of feature engineering plays a crucial role in the performance and accuracy of sgon models. It is true that the classic hand-driven feature building method can incorporate insights from the professional field, but this method is potentially accompanied by the hidden trouble of information omission, and does not necessarily touch the boundary of the optimal solution. In order to solve these problems, this paper proposes two strategies of feature extraction: ensemble learning and deep learning. Ensemble learning enhances generalization by combining the opinions of multiple models, while deep learning allows models to automatically learn features, reducing the need for human intervention. Both of these methods can overcome the limitations of manual feature design to varying degrees. In addition, the paper also introduces the application of parallel coordinate graph in feature selection. By using the parallel axis system to implement projection transformation of high-dimensional data, scholars can intuitively analyze the data structure, so as to promote the process of feature selection and optimization. This method not only gives insight into the subtle relationship between the data, but also cleverly stimulates the potential of human pattern recognition and further improves the comprehensive performance of the model.

KEYWORDS

Big data; Feature construction; Feature extraction; Parallel coordinate graph; Data visualization

1. INTRODUCTION

With the rapid evolution of information technology, human society is encountering unprecedented data flooding. The rise of big data is not only reflected in the explosive growth of data volume, but also in the innovative changes in its processing, access and analysis methods. This transformation has given rise to disruptive innovations in research practice, business management and even public policy construction. In this era of big data, how to extract practical knowledge and insight from the vast ocean of data constitutes the core challenge of current research. As a cornerstone of machine learning and artificial intelligence, feature engineering involves the preprocessing, transformation and selection of initial data to extract representations that can fully reveal the core characteristics of a problem. High-quality feature selection can not only greatly improve the performance of the model, but also reduce the time consumption and the demand for computing resources in the training process [1]. Therefore, feature engineering is essential for building efficient and robust AI systems.

This paper explores the latest advancements in feature engineering & data visualization in AI under big data, focusing on manual feature design, ensemble learning, deep learning extraction, & parallel coord. graphs. It proposes a comprehensive framework integrating these techniques for efficient data analysis. The study validates the framework's effectiveness & practicality through empirical research.

2. LITERATURE REVIEW

At the heart of the machine learning ecosystem is the process of feature-building, which can extract valuable insights from raw data sources to build ever more sophisticated predictive models. In the early stage of feature engineering, it is highly dependent on the expertise and intuition of domain experts, who manually perform data purification, transformation and screening tasks [2]. With the evolution of technology, automated feature generation strategies have emerged, effectively reducing the need for human intervention. Data visualization, as a means of transforming numerical information into visual display, strongly promotes the cognitive improvement of users on the statistical characteristics and internal structure of data. Traditional chart forms such as bar charts, graphs, and scatter plots have proved effective in revealing the complex properties of low-dimensional data. However, in the face of the rapid expansion of data dimensions, these classical techniques appear inadequate. In recent years, advanced visualization techniques such as parallel coordinate system and thermal matrix have gradually emerged and become the preferred means for analyzing complex and high-dimensional data. Ensemble learning, as an intelligent architecture, can break the boundaries of improving model performance by fusing the prediction output of multiple base learners. In the field of feature engineering, ensemble learning uses multi-model ensemble to evaluate the value of features, which helps to prevent the occurrence of overfitting. Strategies such as random forests and gradient boosting trees have shown excellent results in many practical scenarios [3]. On the other hand, deep learning, especially convolutional neural network (CNN) and recurrent neural network (RNN), has been widely used in image recognition and natural language processing due to its powerful feature learning mechanism. These models have the ability to self-learn feature representations, greatly reducing the dependence on external artificial feature engineering. In a variety of situations, their representational abilities are comparable to or even exceed human intelligence.

Parallel Coordinate Plots (PCPs), a groundbreaking visualization technique for high-dimensional data, expose intricate dataset structures with distinctive simplicity. By mapping each data point's attributes via linear segments across parallel axes, PCPs consolidate all dimensional information into a singular, insightful display. This characteristic empowers users to keenly unravel correlations among attributes and discern concealed patterns and temporal trends with remarkable efficiency. Having permeated financial analytics, medical research, bioinformatics, and numerous other domains, PCPs exhibit an undeniable versatility and practical significance.

3. FEATURE ENGINEERING

3.1. Manual Feature

Specialized domain experts harness their profound knowledge to define features, capitalizing on prior domain insights, albeit with a risk of high subjectivity that can lead to diverse feature sets, data loss, and inefficient handling of extensive datasets. Ensemble learning mitigates these limitations by aggregating multiple models' predictions, enhancing overall predictive prowess, particularly in feature selection. It not only boosts a model's generalization capability and reduces overfitting susceptibility but also facilitates feature importance assessment, enabling the selection of the most distinctive features, thereby reducing their quantity and augmenting model interpretability. This enhances resilience against outliers and noise [4]. Notably, Random Forests, through constructing numerous decision trees and ensemble voting, and Gradient Boosting Trees, which enhance performance by sequentially appending weak models, both offer insightful feature importance analysis for streamlined feature selection.

3.2. Deep Learning Feature Extraction

Essentially, ensemble learning harnesses synergies among multiple models to enhance overall predictive prowess, bypassing certain constraints of manual feature engineering, particularly during feature selection. This approach bolsters generalization capability, mitigates reliance on training data, and optimizes performance on unseen data [5]. In high-dimensional settings, ensembles effectively assess feature relevance, facilitating the selection of discriminative attributes, reducing dimensionality, and enhancing model interpretability while fortifying resilience against noise. Random Forest, an ensemble technique, constructs a parallel ensemble of decision trees, employing majority voting for final predictions, thus curbing overfitting and boosting predictive accuracy. Its inherent feature importance measurement serves as a robust basis for feature selection. On the other hand, Gradient Boosting iteratively strengthens weak models, further elevating the collective performance of the ensemble. Both methodologies excel in capturing non-linear associations and intricate feature interactions, enabling profound analysis of feature relationships and uncovering latent patterns and dynamic trends. Their applicability spans diverse disciplines such as financial modeling, medical research, and bioinformatics, underscoring their universality and profound impact.

4. DATA VISUALIZATION ANALYSIS

4.1. Principles and Techniques of Data Visualization

Visual representation of data, a potent tool, simplifies intricate datasets into comprehensible visuals, enabling intuitive comprehension of underlying structures, trends, and anomalies. This technique encompasses diverse chart types and tactics, each with distinct utilities and strengths. For instance, scatterplots excel in elucidating correlations, such as linear associations, outliers, and clustering tendencies. Heatmaps, employing gradient colors to accentuate numerical intensity, frequently illuminate correlation matrices or distance matrices, facilitating the discovery of patterns and connections within data. Bar comparisons uniquely quantify disparities among categories, like sales performance or market share contrasts. Time-series analysis finds its forte in line graphs, adept at depicting evolving trends, be it stock movements or climate change dynamics. Parallel coordinates plots, tailored for multidimensional data, portray categorical distinctions through linear trajectories across axes, promptly revealing the impact of individual features on class separation, especially in high-dimensional domains like finance or bioinformatics. By scrutinizing these line patterns, salient characteristics can be swiftly identified in an uncluttered manner.

4.2. Parallel Coordinate Diagram Introduction

arallel coordinate plots, a potent tool for visualizing multi-dimensional data, employ a series of parallel axes to map individual attributes, depicting data instances as curves connecting these axes. Each axis singularly represents a distinct attribute semantics. For instance, an age axis might span from 20 to 60 years, while an income axis could range from 30,000to120,000. The exact position of curve points closely aligns with the corresponding attribute's value. By scrutinizing the line distribution, similarities and disparities among data points become evident. Converging lines on certain axes suggest shared attribute values among those points, whereas dispersed lines denote significant dissimilarities.

Parallel coordinate plots exhibit exceptional utility in feature selection. The arrangement of lines along each axis enables instantaneous differentiation of datasets, revealing which attributes are pivotal for class or group distinction. Outliers often manifest as deviant lines, facilitating swift identification and scrutiny. By scrutinizing dynamic patterns in data trajectories, one can uncover covariance relationships among features. For instance, concurrent increases along lines representing income and expenditure might suggest a positive correlation. This visual approach not only facilitates

dataset discrimination and outlier detection but also elucidates underlying feature associations, thus expediting feature selection and fostering in-depth Data analysis.

5. EXPERIMENTAL RESULTS AND DISCUSSION

5.1. Experimental Design

To enhance the generalizability and interdisciplinary impact of our study, we used a range of publicly available data sets. The essential properties of these data sets and evaluation criteria are described in detail below. Sample size The database used in this study contains enough 10000 samples to ensure a rigorous empirical test of the performance and stability of the algorithms used. In terms of document length, each sample has an average of about 200 words, revealing the depth and diversity of text information in the dataset, which is suitable for text prospecting and natural language processing exploration. In the feature dimension, the dataset of 500 features provides rich analysis materials and modeling basis, which comprehensively covers the multivariate attributes of patients, the fine diagnosis of diseases, and the diversity of treatment options.

5.2. Experimental Setup

In order to ensure the reliability and repeatability of the experimental results, we carried out detailed pre-processing steps for the original data. These steps include the removal of stops, punctuation, and the drying or morphing of the text to reduce noise and improve the accuracy of the model. In the data preprocessing phase, we take the following measures: First, we remove the stop words, which are frequent in the text but contribute less to the meaning of the topic, such as "of", "and", "in", etc. Removing these words helps reduce redundant information in the data. Second, we removed punctuation, because punctuation usually does not carry semantic information, and removing them reduces unnecessary features and simplifies the data set. Finally, we have a text that is desiccated or morphed, which is the process of reducing words to their root form, and the principle of morphing is the conversion of words to their basic form. Both methods help to reduce the size of the feature space and improve the efficiency of the model.

According to the results of data visualization analysis in Chapter 4, we select the most representative features for model training. These features are based on parallel coordinate plots and other visualization tools, and they show high potential for distinguishing between different categories or predicting targets. In the feature extraction stage, we focus on identifying the contribution of each feature to the prediction performance of the model, which can screen out the features with substantial information to promote the improvement of prediction accuracy. In view of the accumulation of previous research, we selected several representative machine learning strategies and deep learning architectures as benchmark models, including logistic regression, support vector machine (SVM), convolutional neural network (CNN), and recurrent neural network (RNN). To quantify the generalizability of the model, we adopted the classic 10-fold cross validation technique. The original data set is randomly divided into ten equal parts without putting back, one fold is used as the validation set, and the remaining 10% is used for training, which ensures that all examples can be traversed and verified, and further improves the reliability and scalability of the evaluation. This experimental arrangement not only allows us to investigate the behavior of the model on different data segmentation, but also ensures the stable reproducibility of the experimental results, which lays a solid empirical foundation for subsequent in-depth exploration.

5.3. Result Analysis

Table 1. Performance of different models on data sets

Model	Accuracy rate	Recall	F1 score
Logistic regression	0.85	0.84	0.85
SVM	0.87	0.86	0.87
CNN	0.90	0.89	0.89

At the same time, the ability to identify positive examples and balance accuracy and recall rate are better than the other two models. As shown in Table 1, when evaluating the performance of Logistic regression, support vector machine and convolutional neural network, CNN came out on top in all indicators, with an accuracy as high as 0.90, recall rate and F1 score as high as 0.89. SVM ranked second with precision of 0.87, recall of 0.86, and corresponding F1-score of 0.87. Logistic regression was in third place with a precision of 0.85 and a recall of 0.84, corresponding to a F1 index of 0.85. Therefore, for such specific classification challenges, CNN shows excellent performance in accurately predicting instance attribution.

6. CONCLUSIONS AND FUTURE WORK

6.1. Conclusion

In the era of overwhelming data abundance, the significance of feature engineering and data visualization for in-depth analysis is underscored. Our examination delves into the evolution and future prospects of feature engineering, with a focal emphasis on the pivotal role of feature selection and construction in enhancing machine learning model efficacy. Two innovative approaches to boost engineering are presented: ensemble learning and deep learning-based feature extraction. Ensemble methods enhance model generalization by aggregating diverse model types, whereas deep learning facilitates automatic feature learning, minimizing human intervention. Furthermore, we propose the utilization of parallel coordinate plots as a potent visualization aid for guiding feature selection. By transposing data points into lines and scrutinizing trends within the parallel coordinates, identifying the most discriminative features for classification or clustering becomes significantly streamlined. This technique not only aids in feature selection but also integrates human intuition, thereby escalating overall model performance.

6.2. Future Work

Future investigations will delve into sophisticated ensemble learning techniques, automatizing feature engineering for the creation of advanced automated tools that minimize human intervention in feature identification and construction. Research will also focus on leveraging reinforcement learning to optimize feature selection, enabling models to adapt feature weights based on feedback. Exploring variants like Bagging and Boosting, as well as efficacious fusion strategies for diverse model types, is crucial. The potential of exploiting various layers within deep learning architectures to derive enhanced feature representations, along with integrating traditional and deep learning features, will be exhaustively studied. Augmenting visualization methodologies, we'll investigate alternative high-level visualizations such as dimensionality reduction techniques (e.g., t-SNE, PCA) and interactive tools, catering to intricate data analysis demands. Moreover, our efforts will extend to real-time stream data processing, where feature engineering and visualization technologies will be applied to develop intuitive interactive tools for profound data exploration and dynamic analysis in ever-evolving big data landscapes.

REFERENCES

[1] Michaels G S, Carr D B, Askenazi M, et al. Cluster analysis and data visualization of large-scale gene expression data[C]//Pacific symposium on biocomputing. 1997, 98: 42-53.

[2] Azzam T, Evergreen S, Germuth A A, et al. Data visualization and evaluation [J]. New Directions for Evaluation, 2013, 2013(139): 7-32.

[3] Wang W, Lu C. Visualization analysis of big data research based on Citespace [J]. Soft Computing, 2020, 24(11): 8173-8186.

[4] Dzemyda G, Kurasova O, Zilinskas J. Multidimensional data visualization [J]. Methods and applications series: Springer optimization and its applications, 2013, 75(122): 10-5555.

[5] Huang B, Jiang B, Li H. An integration of GIS, virtual reality and the Internet for visualization, analysis and exploration of spatial data [J]. International Journal of Geographical Information Science, 2001, 15(5): 439-456.