Hybrid feature selection model based on machine learning and knowledge graph

Xiaoqing Peng\textsuperscript{1,2,a}, Yong Shuai\textsuperscript{3b},* Yaxi Gan\textsuperscript{4c}, Yaokai Chen\textsuperscript{3d}

\textsuperscript{1}Chongqing Ceprei Industrial Technology Research Institute, Chongqing, China
\textsuperscript{2}Chongqing Key Laboratory of Reliability Technologies for Smart Electronics, Chongqing, China
\textsuperscript{3}Chongqing Public Health Medical Center, Chongqing, China
\textsuperscript{4}Chongqing Vanguard Radio Energy Transmission Research Institute Co., Ltd, Chongqing, China
\textsuperscript{a}email:372870591@qq.com \textsuperscript{b}email:alexshuai@sina.com
\textsuperscript{c}email:dearx1019@163.com \textsuperscript{d}email:yaokaichen@hotmail.com
\textsuperscript{*}Corresponding Author: \textsuperscript{b}email:alexshuai@cqgwzx.com

Abstract. Aiming at the problem that the current feature selection algorithm can not adapt to both supervised learning data and unsupervised learning data, and had poor feature interpretability, this paper proposed a hybrid feature selection model based on machine learning and knowledge graph. By the idea of hybridization, this model used supervised learning algorithms, unsupervised learning algorithms and knowledge graph technology to model from the perspective of data features and text features. Firstly, the data-based feature weights were obtained through the machine learning model, and then the text-based weights were obtained by using the knowledge graph technology, and the weight sets are combined to obtain a feature matrix with good explanatory properties that meets both the data and text features. Finally, the case analysis proves that the method proposed in this paper has good effects and interpretability.

1. Introduction

With the advent of the big data era, mass data has increased the number of data features. However, the high dimensionality affected the efficiency of the machine learning model, reduced the accuracy of the model, and made the calculation results of the model difficult to interpret. Therefore, data scientists used feature selection to reduce feature dimensions to solve these problems. Feature selection is a dimensionality reduction method. Compared with feature extraction, feature selection can not only remove irrelevant or redundant features, but also retain some of the original features that are convenient for machine learning, thereby improving the efficiency and accuracy of machine learning models, and the interpretability of the result, which is of great significance for machine learning modeling\cite{1,2}.

The current feature selection methods can be classified in many ways. According to the results of the feature selector, it can be divided into a subset evaluation model and a feature ranking model. According to the inference model, it can be divided into Filters methods, Wrappers methods and Embedded methods. According to the idea of supervised learning, it can be divided into supervised
learning model and unsupervised learning model.[1] However, these methods can only be used in a specific environment. Moreover, due to the impact of data differentiation, the interpretability of features for machine learning models is reduced.[3-6] Therefore, we proposed a hybrid feature selection model. The model regarded feature selection as a process-based modeling method. It adopted the feature selection method of supervised learning and unsupervised learning, and combined the knowledge graph technology for feature selection, which can improve the efficiency and the interpretability of the model.

2. Method
The hybrid feature selection model proposed in this paper took a large number of features as input and a small number of interpretable, more robust features as output for modeling. Firstly, a machine learning algorithm was used to output the weight matrix of the feature by the characteristics of the data. Then, the knowledge graph technology was used to obtain the feature text weight matrix, and the weight values of each feature in the weight matrix were added together, and a case was used to verify the effectiveness of the model finally.

When the sample was a data set with results, we firstly used the supervised learning and unsupervised learning algorithms to obtain the feature weight matrix. Secondly, we combined this matrix with the feature weight matrix obtained by the knowledge graph to get the final feature weight matrix. Finally, we used the data set to verify the validity of the feature matrix by classification algorithm, and the verification indicator was F1-Measure.

When the sample was a data set with no results, we firstly used the unsupervised learning algorithm to obtain the feature weight matrix. Secondly, we combined this matrix with the feature weight matrix obtained by the knowledge graph to get the final feature weight matrix. Finally, we used the data set to verify the effectiveness of the feature matrix by clustering algorithm, and the verification index was Silhouette Coefficient Index.

From Figure 1 we can find that for a data set with results, this model was a hybrid model. In the hybrid model, we used Recursive Feature Elimination (RFE) to obtain the feature matrix for supervised learning modeling, and used Principal Component Analysis (PCA) to obtain the feature matrix for unsupervised learning modeling. For a data set with no results, the model becomes a series model. In the series model, the unsupervised learning modeling used was principal component analysis (PCA).

2.1. RFE modeling process.
RFE is a feature selection algorithm. It is implemented by fitting a given algorithm model (such as linear regression), sorting features by importance, discarding the least important features, and refitting the model. The modeling process using python's sklearn.feature_selection.RFE is as follows:

![Figure 1. Overall modeling flowchart](image-url)
① Load Data set;
② Initialize the linear regression model;
③ Use the linear regression model to initialize the RFE model;
④ Run the RFE model, and output rfe.ranking_ as the importance weight matrix of each feature.

### 2.2. PCA modeling process.
PCA is a method of selecting features through feature projection. This method can obtain the weight value of each feature and the credibility of the weight value through the function of PCA. The modeling process of using python's sklearn.decomposition.PCA is as follows:
① Read the data set and standardize it;
② Initialize the PCA model and determine the n_components parameters;
③ Run the PCA function and output pca.components_ and pca.explained_variance_ratio_, where pca.components_ can be understood as the weight coefficient under each credibility, and pca.explained_variance_ratio_ is the credibility matrix.

### 2.3. Knowledge graph modeling process.
Knowledge graph was a large-scale semantic network that can assist language cognition, feature selection, search and recommendation, and conduct knowledge questions and answers. This article adopted the knowledge graph construction method of literature [8]. On the basis of the completed case data knowledge graph, we firstly used graph algorithms and deep traversal of graphs to mine information. Secondly, we found the frequency of feature appearance in the neo4j graph database, and standardized the frequency of feature appearance. Finally, The processed result was used as the weight value matrix of the feature.

### 2.4. Overall modeling process.
The overall modeling process is as follows
① Data preparation, including data collection and integration, data preprocessing, and data standardization processing.
② For the supervised learning model, calculate the feature weight matrix P; for the unsupervised model, calculate the feature weight matrix Q. Matrices P and Q are both 1×N matrices, N is a positive integer and N>1.
③ Use the StandardScaler function to standardize the matrices P and Q, and add the two matrices to get the matrix S, S is a 1×N matrix.
④ Use the knowledge graph technology to obtain the text weight matrix R, and use the StandardScaler function for standardization. R is a 1×N matrix.
⑤ Standardize the matrix S and add it to the matrix R to get the matrix T as the final feature weight matrix.
⑥ Sort the eigenvalues in the matrix T to get the sorted matrix T*.

### 3. Case Analysis
This article used travel feature selection as a case for classification. When travelers plan to travel to different cities, the features they pay attention to are also different. In current travel websites, most of the choices for travel features are listed in full, or personalized through manual services. This method wastes manpower, is highly subjective, and cannot fully meet the needs of passengers. Therefore, we use the model proposed in this paper to assist in screening travel characteristics, and compare with other feature selection models to verify the effectiveness and interpretability of the model.

Chongqing is a hot tourist city in China. The model proposed in this paper is used to model the feature selection of Chongqing tourism. Firstly, we prepare the data. Use "Chongqing" and "travel" as keywords to search on the Internet, crawl the shared content of travel websites and travel forums, where travel-related words are the characteristics, and travel to Chongqing or not to Chongqing is
regarded as the result. Perform text preprocessing on the retrieved content, and use the jieba data package in python to extract the keywords of the text. Through analysis, there are 18 main features of travel attention, including online celebrity check-in, number of attractions, cost performance, food, attraction level, travel environment, single travel, parent-child travel, honeymoon travel, season, city traffic, accommodation, friendliness, Transportation outside the city, natural scenery, tourism characteristics, travel days, per capita consumption. Use these characteristics to collect relevant data to establish a data set, and use the model proposed in this paper for analysis.

During the modeling process, the confidence level of pca.explained_variance_ratio_ is set to 0.95, and the feature weights of RFE and the feature weights obtained from the knowledge graph are standardized. The final feature weight matrix \( T \) is
\[
\begin{bmatrix}
0.442, 0.312, 0.387, 0.425, 0.223, \\
0.332, 0.278, 0.121, 0.325, -0.112, 0.223, 0.309, 0.296, -0.279, 0.212, \\
0.335, -0.112, 0.315
\end{bmatrix}.
\]
The 10 most important features are [net celebrity check-in, number of attractions, value for money, food, travel Environment, honeymoon trip, accommodation, friendliness, tourism characteristics, per capita consumption].

In order to verify the effectiveness of the algorithm, we compare the model in this paper with other feature selection algorithms, including PCA, Latent Dirichlet Allocation (LDA), Locally Linear Embedding (LLE), and RFE. The verified data set is the data set corresponding to the 10 most important features. When the data set includes results, supervised learning algorithms (including logistic regression (LR), support vector machine (SVM), random forest (RF)) are used for classification verification. When the data set does not include results, unsupervised learning algorithms (including Kmeans, Density-Based Spatial Clustering of Applications with Noise (DBSCAN), Balanced Iterative Reducing and Clustering Using Hierarchies (BIRCH)) are used for clustering verification. The evaluation results of supervised learning are shown in Table 1, and the evaluation results of unsupervised learning are shown in Table 2.

| Feature Selection Method                  | F1-measure        |
|------------------------------------------|-------------------|
|                                          | Classified by SVM | Classified by LR | Classified by RF |
| PCA                                      | 0.321             | 0.429            | 0.323            |
| Latent Dirichlet Allocation (LDA)        | 0.214             | 0.609            | 0.334            |
| Locally Linear Embedding (LLE)           | 0.276             | 0.563            | 0.412            |
| RFE                                      | 0.377             | 0.554            | 0.322            |
| Method in this paper                     | 0.421             | 0.653            | 0.452            |

| Feature Selection Method | Silhouette Coefficient Index (n_cluster=2) |
|--------------------------|-------------------------------------------|
|                          | Clustered by Kmeans | DBSCAN | BIRCH |
| PCA                      | 0.432             | 0.332  | 0.312 |
| DBSCAN                   | 0.225             | 0.422  | 0.261 |
| BIRCH                    | 0.432             | 0.313  | 0.337 |
| Method in this paper     | 0.627             | 0.451  | 0.559 |

From the analysis results in Table 1 and Table 2, it can be seen that the method proposed in this paper can be applied to data sets with results and data sets without results, and the result of feature selection has the best effect on classification and clustering models, which shows that the model has good robustness and applicability. At the same time, judging from the characteristics of the paper selection, the most concerned feature of Chongqing travel is that this result is also very realistic, indicating that the model has good interpretability.
4. Conclusion
With the improvement of computer storage performance, the application of massive high-dimensional data will become more and more common. In order to reduce the complexity of modeling, save the cost of extracting unnecessary features, and remove the noise in the data set, we propose a hybrid feature selection model based on machine learning and knowledge graph technology. The model has stronger robustness and interpretability, and has a good effect in feature selection of travel data. In the future, we will also improve the algorithms of supervised learning and unsupervised learning, so that these algorithms can be more adapted to sparse and noisy data, thereby expanding the scope of application of the model.

Funding:
This study was supported by the Chongqing Science and Technology Bureau Project (cstc2019jscx-fxyd0298, cstc2020jscx-cylh0004).

Contributors:
XQP and YS contributed equally to this work. XQP and YXG collected and provided the data. YS extracted data and cleaned data. XQP and YS designed the study, constructed the model, and analyzed the data. YKC, YXG and XQP provided the correlative Knowledge. YS drafted of the manuscript. All authors read and approved the final manuscript.

Data availability:
No data are available. Data will not be shared, but reasonable requests may be directed to the corresponding author (alexshuai@sina.com).

Acknowledgments:
We would like to express our gratitude to all participants involved in this study and the staff of Chongqing Public Health Medical Center.

References
[1] Jundong Li, Kewei Cheng, Suhang Wang, Fred Morstatter, et al. (2017) Feature Selection: A Data Perspective. ACM Computing Surveys, 50: 1-45.
[2] Zhiqin Li, Jianqiang Du, Bin Nie, Wangping Xiong, et al. (2019) Summary of Feature Selection Methods. Computer Engineering and Applications, 55: 10-19.
[3] Zsolt János Viharos, Krisztián Balázs Kis, Ádám Fodor, Ádám Máthé Büki. (2021) Adaptive, Hybrid Feature Selection. Pattern Recognition, 107: 116, 1-13.
[4] Shaobo Lin, Jian Fang; Xiangyu Chang. (2020) Learning With Selected Features. IEEE transactions on cybernetics, 6: 1-15.
[5] Zhanshan Li, Zhaogeng Liu. (2019) Feature selection algorithm based on XGBoost. Journal on Communications, 40: 101-108.
[6] Yu Zhang , Xiaojie Liu, Beibei Li. (2021) A feature selection method for remote access trojan's traffic. Journal of Sichuan University (Natural Science Edition), 58: 1-9.
[7] Mingming Yang, Junchuan Yang. (2021) Feature Selection Based on Distance Measurement. Journal of New Media, 3: 19-27.
[8] Zhonghao Jia, Tianlong Gu, Chenzhong Bin, Liang Chang, et al. (2019) Tourism knowledge-graph feature learning for attraction recommendations, CAAI Transactions on Intelligent Systems, 14: 430-437.