ISSN: 2664-9640

Feature selection algorithm of massive high-dimensional data based on artificial intelligence technology

Junhong Tong

Guangzhou College of Commerce, Guangzhou 511363, China

1383239@qq.com

Abstract

With the development of computer technology and artificial intelligence, data explosion is one of the hottest issues in contemporary times. In the ultra-high dimensional data, the sample size of the data increases significantly. At this time, only a few covariates are associated with the response variables. The model presents the characteristics of sparsity and the interpretability of the model parameters is poor. Statisticians are faced with the task of identifying the most important features and constructing the optimal interpretation model to connect these important features with the response variables. Extracting useful features from ultra-high dimensional data is the basis of ultra-high dimensional data modeling. Because the model is sparse at this time, it is important to delete the most obvious non impact characteristics before any accurate analysis of ultra-high-dimensional data. Because the dimension is too high, many traditional modeling methods and high-dimensional data variable selection methods are not suitable for ultra-high-dimensional data analysis. In recent years, mathematicians have developed some algorithms for this goal. A more feasible strategy is to establish a two-stage feature selection process. In the first stage, a fast and efficient variable screening process is used to reduce the feature dimension to an appropriate scale below the sample size, and all important features can be retained. On this basis, some effective methods are used to select important variables for the reduced high-dimensional data. Aiming at massive high-dimensional data, this paper proposes a feature selection algorithm based on artificial intelligence technology. The algorithm can effectively complete the extraction of feature attributes, and the execution efficiency of the algorithm is very high. Experimental results show that the proposed algorithm has a high speedup ratio.

Keywords

Feature selection; High-dimensional data; Sparsity principle; Data mining.

1. Introduction

With the emergence of massive video, picture, text, voice and social relationship data, as well as the rise of Internet of things and cloud computing, data is growing and accumulating at an unprecedented rate [1]. On the one hand, the huge amount of data provides us with enough data to understand things and provide sufficient basis for judgment and decision-making of data analysis; On the other hand, it also poses a challenge for us to mine useful information in data [2]. In the face of the dilemma of rich data and lack of information, data mining technology can play a great role [3]. Data mining makes full use of the theories and methods of machine learning, pattern recognition, mathematical statistics, artificial intelligence, fuzzy logic, neural network, evolutionary algorithm and so on. Its research contents include synthesizing and improving various methods and technologies and effectively integrating them, as well as studying new technologies for data mining [4]. Data mining allows us to find hidden, effective and potentially valuable laws and patterns in data [5]. However, with the deepening of research

ISSN: 2664-9640

in the field of data mining, the research object becomes more and more complex, and the feature dimension increases sharply, resulting in a large number of high-dimensional data, especially in the aspects of text data, transaction data, gene data and network access data [6]. There are a large number of redundant, uncorrelated and noisy features in high-dimensional data. These features not only increase the time complexity and space complexity of machine learning algorithm, but also consume a lot of resources, and also greatly reduce the solution accuracy of the algorithm, which has a negative impact on the final analysis and decision [7].

In recent years, with the rapid development of technologies such as big data, artificial intelligence and cloud computing, the amount and dimension of data have exploded. How to mine valuable information from these data plays a crucial role in classification learning [8]. However, the classification learning task of such high-dimensional samples becomes difficult due to the high-dimensionality of the feature space and the uneven distribution of categories. Among them, the high dimensionality of the feature space can easily lead to the "curse of dimensionality" [9]. Therefore, dimensionality reduction has become a key problem that must be solved before classification learning. As one of the data dimensionality reduction techniques, feature selection plays a significant role [10]. In addition, the class imbalance problem of high-dimensional sample data can easily cause the classification model to fail to recognize small class samples. How to avoid the disaster of dimensionality and the problem of inability to identify small samples in the classification learning of high-dimensional sample data has become one of the difficult problems to be overcome by feature selection technology [11].

Firstly, this paper expounds the significance of feature selection in high-dimensional sample data classification. Secondly, in order to further remove data redundancy and screen out important data features, the sparsity principle is introduced to improve the feature selection algorithm, which is affected by high-dimensional data and reduce the processing time of the algorithm.

2. Feature selection method of high-dimensional data

2.1. Feature selection

Feature selection, also known as feature subset selection, or attribute selection, refers to selecting a feature subset from all features to make the constructed model better [12]. In the practical application of machine learning, the number of features is often large, among which there may be irrelevant features, and there may be interdependence between features [13]. Feature selection can eliminate irrelevant or redundant features, thereby reducing the number of features, improving model accuracy, and reducing running time. On the other hand, picking out truly relevant features simplifies the model and makes it easier for researchers to understand how the data is generated. The feature sequence-based ensemble method is shown in Figure 1.

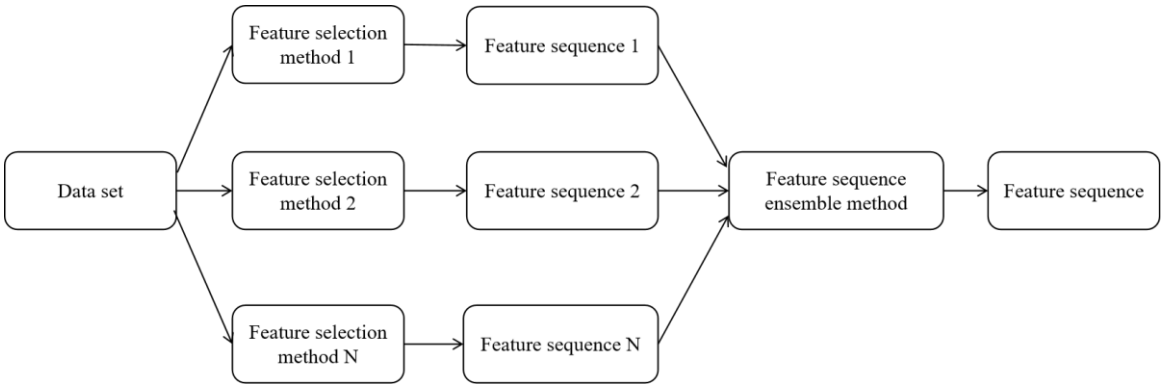

Figure 2 Integration method based on feature sequence

ISSN: 2664-9640

At present, many methods have been proposed for feature selection on high-dimensional data. There are methods of filtering irrelevant features, the basic idea of which is to calculate the degree of association between each feature and class through certain rules, and arrange them in descending order, and select the first N features whose values are greater than the selected threshold or the first AA features to form the final feature subset, including the method of eliminating redundant features through clustering, and the combined feature selection algorithm. There are also genetic algorithms, simulated annealing algorithms, neural networks and so on. Many of the above algorithms are very sensitive, even for some slight changes of training samples, which is caused by inaccurate estimation of statistical parameters such as sample mean and standard deviation used in feature evaluation criteria. The instability of feature selection results will cause researchers to lose their research enthusiasm and confidence in research work. This chapter mainly introduces the stable feature selection method.

2.2. Stable feature selection method

At present, researchers have proposed a variety of solutions to the stability problem of feature selection for high-dimensional data, mainly including: integrated feature selection method, prior feature correlation method, $Group$ feature selection method and sample injection method. The integrated feature selection method is the most commonly used method to solve the problem of feature selection stability. The idea of this idea is similar to that of classifier integration: different feature selection algorithms get different feature subsets. The feature subsets obtained by different feature selection algorithms can achieve complementary effects. Each feature selection algorithm has its own characteristics that limit its search ability in the feature space, so the ensemble method can obtain a solution that approximates the global optimal solution. The integrated feature selection method mainly includes two steps: constructing different feature selectors and integrating the results of these feature selectors to obtain relatively consistent results. The existing stable feature selection method is shown in Figure 2.

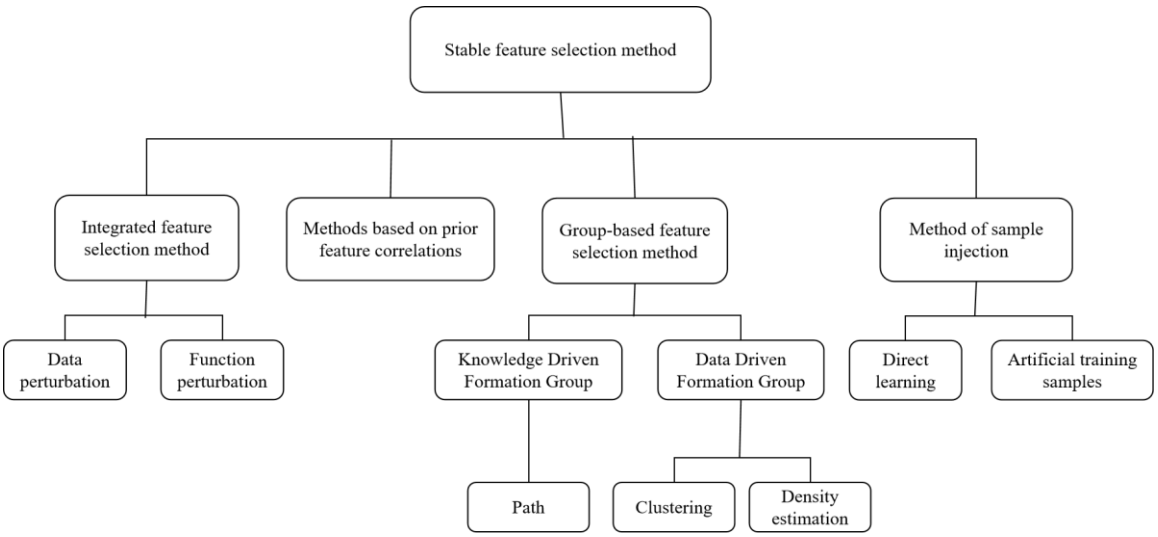

Figure 2 Existing stable feature selection methods

Theoretical and experimental results show that the integration of different learners is one of the key factors for the success of integrated learning. In order to construct different learners, the following two strategies are commonly used: data disturbance and function disturbance. This method of data perturbation is to select features in different subsets or separate feature subspaces. Function disturbance is a method that uses different learners to select features, and then integrates the feature selection results of each learning machine. There are two differences between this method and data perturbation: first, it uses different feature selection methods

ISSN: 2664-9640

instead of the same method; Second, this method operates on the original data set without sampling. Under this strategy, a new integrated feature selection method is proposed. Because the function disturbance is limited by the number of available feature selection methods, it is not as flexible as the data disturbance method. Generally, the number of selected learners will not be more than 4. It should also be noted that there must be differences between the selected learners, because if two feature selection algorithms get similar results, it will be meaningless to integrate their results.

3. Result Analysis and Discussion

3.1. Feature selection method combined with sparse online learning

At present, the feature selection algorithm of big data is mainly based on data clustering. The data information is divided into different subsets by using statistical principle, and the features of the data are extracted. The data are clustered according to the distance between the data features and the clustering center, and on this basis, the corresponding improved optimization methods are introduced. In addition, the existing methods mostly use offline feature selection, that is, it is assumed that the sample set and its corresponding data features have been obtained in advance. However, due to the streaming dynamic characteristics of big data, it is difficult to guarantee that the assumption is always true, and in many cases, it is necessary to obtain the features of big data timely and effectively. Obviously, the offline processing method is difficult to meet the needs. In order to effectively improve the clustering effect of big data, sparsity principle is introduced to improve the feature selection algorithm which is influenced by high-dimensional data and reduce the processing time of the algorithm. The feature selection algorithm of big data, which combines the competitive entropy weighting and thinning method, is integrated into the online learning framework to realize the timely and accurate selection of data features.

High-dimensional data clustering is the key and difficult point in the field of data mining and machine learning, and subspace clustering is an effective way to complete high-dimensional data clustering. Now subspaces have been used in many fields. . From the perspective of machine learning, the main subspace clustering methods are divided into French methods, algebraic methods, statistical methods and methods based on sparse classes, among which methods based on sparse classes have received extensive attention in recent years. This type of method usually consists of two stages. The first stage uses sparse representation or low-rank approximation to obtain an association matrix, and the second stage processes the association matrix through sparse classes to obtain the final data partition. At present, the sparse subspace clustering algorithm is regarded as the most excellent subspace clustering algorithm by many scholars. In the first stage, the algorithm uses sparse representation technology to obtain a good correlation matrix, thereby improving the data segmentation in the subsequent sparse class stage. precision. However, this method has the problem of high computational complexity.

3.2. Parameter Settings

In order to optimize the data feature selection algorithm to cope with large-scale high-latitude data, the principle of sparsity is introduced. , using the L_1 criterion to sparse the data feature set, so as to complete the dimensionality reduction processing of big data. According to the L_1 criterion, if the vector x is projected on the L_1 sphere, then for the vector x, all its components will move to its maximum component. Therefore, according to the L_1 criterion, the fractional value components contained in the vector x can be removed, and only the The vector has a large influence factor, and the corresponding L_1 criterion has an inequality as follows:

ISSN: 2664-9640

$$\left\|x-x^{m}\right\| \leq \xi_{q}\|x\|_{1}(m+1)^{1/(q-1)} \tag{1}$$

Where x^{m} is the m maximum vectors containing only x, and $m=1,2,\cdots,d,\xi_{q}$ is the constant associated with q. At this time, the classifier restricted by L_1 criterion is described as:

$$\Delta_{R}=\left|w\in R^{d}:\|w\|_{1}\leq R\right| \tag{2}$$

In order to enhance the sparsity of the classifier, it is introduced that the decision threshold θ,θ is a positive real number. In $x=[x_1,x_2,\cdots,x_d]\in R^{d}$, the classifier is expressed as:

$$T(x,\theta)=[T(x_1,\theta),T(x_2,\theta),\cdots,T(x_d,\theta)] \tag{3}$$

After K iterations, the characteristic value greater than the threshold θ is retained; The eigenvalues less than the threshold and greater than half the threshold are linearly amplified to retain the influence factor of this eigenvalue; Set eigenvalues less than half the threshold to zero. The algorithm framework is shown in Figure 3.

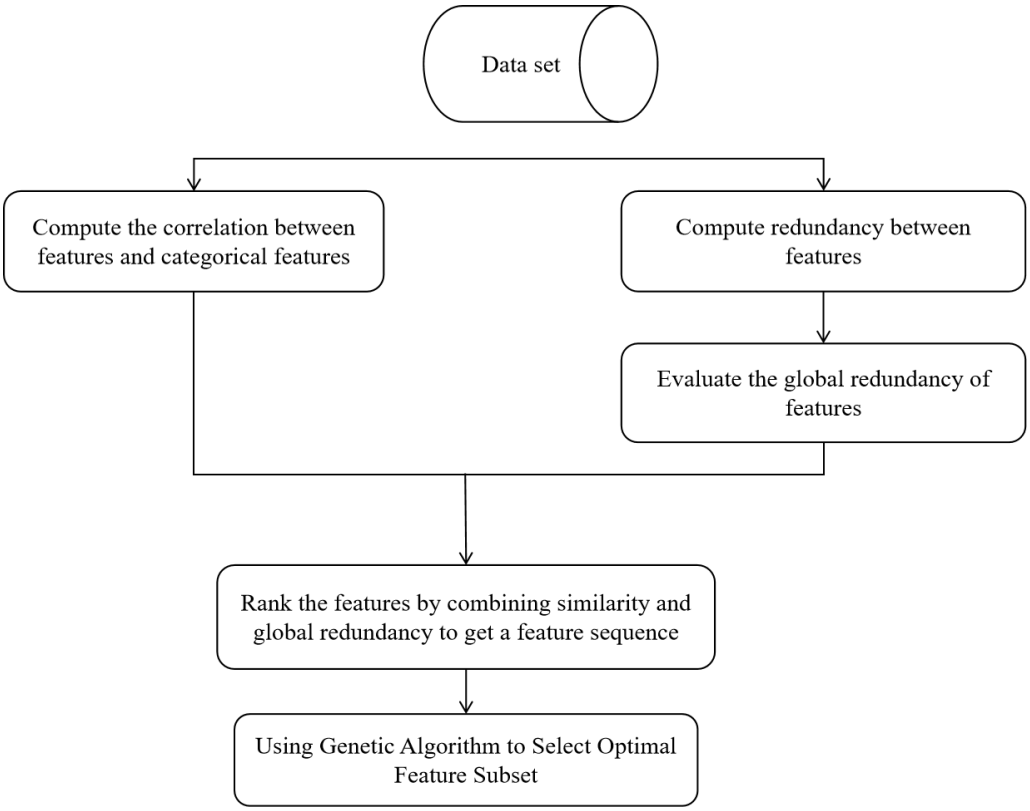

Figure 3 Algorithm framework

Sparse principle can effectively remove the data features with less influence in big data. The sparse fraction factor of data object x_1 is calculated by L_1 norm, and its expression is as follows

$$\min\|S_i\| s\cdot t\cdot x_i = X^{'}s_i \tag{4}$$

Here, $X^{'}$ represents the X matrix of the data object. According to the factor, the objective function corresponding to the sparse score is:

$$S(r)=\frac{\sum_{i=1}^{n}(x_{ir}-(Xs_i))^{2}}{Var(X(r))} \tag{5}$$

ISSN: 2664-9640

Where x_{ir} is the data feature of dimension r; Xs_i is the reconstructed data feature of r dimension; Is the cumulative variance of $Var(X(r))$ X matrix. Using the objective function $S(r)$, the processing of features, such as marking and sorting operations, can be completed. The importance of big data features is inversely proportional to the objective function, so the minimum value of the objective function is used as the evaluation standard of feature importance. With the increase of the number of samples processed by the algorithm, the failure rates of the two methods for feature selection of big data show a downward trend, and the trend is the same. Among them, the failure rate curve of the method in this paper decreases faster, and the failure rate of feature selection is lower when the number of samples is the same, and the performance of feature selection is obviously better than that of the comparison method.

4. Conclusion

Aiming at the feature selection of massive high-dimensional big data in the cloud computing environment, the sparse principle is introduced to mark local data features, determine the importance of the features and sort them, and filter out redundant data, so as to effectively deal with the high-dimensional features of big data, and the overall algorithm is embedded. In the online learning framework, the processing ability of big data is further improved. In practical applications, the high-dimensional data we encounter are often high-dimensional small sample data, such as gene expression data, spectral data, etc. High-dimensional small samples are an important reason for the instability of feature selection methods. Some classical feature selection methods focus on high classification performance while ignoring its stability. Recently, the stability of feature selection methods has received increasing attention, especially in areas such as biomarkers. This chapter systematically introduces the definition of feature selection method stability, several existing stable feature selection methods for high-dimensional data, and stability measurement methods. Through the comparison of simulation experimental data, it is verified that the algorithm proposed in this paper significantly improves the accuracy of feature selection of massive high-dimensional big data in cloud computing environment, and has stable performance for different data sets.

References

[1] Bu F , Chen Z , Zhang Q , et al. Incomplete high-dimensional data imputation algorithm using feature selection and clustering analysis on cloud. Journal of supercomputing, vol. 72, no. 8, pp. 2977-2990, 2016.

[2] Xia S , Wang G , Yu H , et al. Vibration-Based Outlier Detection on High Dimensional Data. International Journal on Artificial Intelligence Tools, vol. 25, no. 03, pp. 1650013, 2016.

[3] Wu S , Hu W , Zhang L , et al. An Intelligent Key Feature Selection Method of Power Grid Based on Artificial Intelligence Technology. Zhongguo Dianji Gongcheng Xuebao/Proceedings of the Chinese Society of Electrical Engineering, vol. 39, no. 1, pp. 14-21, 2019.

[4] Liang M , Zhao J . Feature selection for chemical process fault diagnosis by artificial immune systems. Chinese journal of chemical engineering, vol. 26, no. 08, pp. 7-12, 2018.

[5] Conrad T O F , Genzel M , Cvetkovic N , et al. Sparse Proteomics Analysis – a compressed sensing-based approach for feature selection and classification of high-dimensional proteomics mass spectrometry data. Bmc Bioinformatics, vol. 18, no. 1, pp. 160, 2017.

[6] Hu X , Li J , Yang Y , et al. Reliability verification-based convolutional neural networks for object tracking. IET Image Processing, vol. 13, no. 1, pp. 175-185, 2018.

[7] Zhang Z , Chen S . Real-time seam penetration identification in arc welding based on fusion of sound, voltage and spectrum signals. Journal of Intelligent Manufacturing,vol. 28, no. 1, pp. 207-218, 2017.

ISSN: 2664-9640

[8] Jaddi N S , Abdullah S , Nazri M Z A . A Recurrence Population-based Great Deluge Algorithm with Independent Quality Estimation for Feature Selection from Academician Data. Applied Artificial Intelligence, vol. 35, no. 13, pp. 1081-1105, 2021.

[9] Zhang Z . Big data analysis with artificial intelligence technology based on machine learning algorithm. Journal of Intelligent and Fuzzy Systems, vol. 39, no. 5, pp. 1-8, 2020.

[10] Saucedo-Dorantes J J , Jaen-Cuellar A Y , Delgado-Prieto M , et al. Condition monitoring strategy based on an optimized selection of high-dimensional set of hybrid features to diagnose and detect multiple and combined faults in an induction motor. Measurement, vol. 178, no. 4, pp. 109404, 2021.

[11] Abdulrahman A , Abdulmalik A S , Mansour A , et al. Ultra Wideband Indoor Positioning Technologies: Analysis and Recent Advances. Sensors, vol. 16, no. 5, pp. 1-36, 2016.

[12] Yusof N M , Muda A K , Pratama S F . Swarm Intelligence-Based Feature Selection for Amphetamine-Type Stimulants (ATS) Drug 3D Molecular Structure Classification. Applied Artificial Intelligence, vol. 35, no. 12, pp. 914-932, 2021.

[13] Rostami M , Berahmand K , Nasiri E , et al. Review of swarm intelligence-based feature selection methods. Engineering Applications of Artificial Intelligence, vol. 100, no. 1, pp. 104210, 2021.