Efficient Feature Selection Algorithm for High-Dimensional Non-equilibrium Big Data Set

Shuang-cheng Jia(✉) and Feng-ping Yang

Alibaba Network Technology Co., Ltd., Beijing 100102, China
xindine30@163.com

Abstract. When the traditional algorithm is used to calculate the feature classification of high-dimensional non-equilibrium and large data set, it is easy to appear the problem of low accuracy and recall rate of feature selection. Therefore, a feature selection algorithm based on granular fusion is designed. By using the regularization feature of the data, the original big data aggregate is transformed into a small-scale data subset. On the basis of this, the feature selection function of the data particle is obtained. Finally, the weight fusion calculation of each feature subset is carried out. The feature classification of high-dimensional non-equilibrium big data set is realized. The experimental results show that the feature selection algorithm based on granular fusion can realize the feature selection and recall of high dimensional unbalanced data sets. The accuracy of the method is higher than that of the traditional method, which shows that the method is feasible and effective.

Keywords: High dimensional data · Non-equilibrium feature · Granulation fusion · Feature selection

1 Introduction

Feature selection, as one of the preprocessing steps of data analysis and mining, is widely used in the fields of machine learning and pattern recognition [1]. With the rapid development of network and data acquisition technology, data sets with ultra-high dimension and unbalanced data are emerging constantly. Ultra-high-dimensional unbalanced data usually has a large number of redundant, independent features, which makes feature selection more difficult. The massive size of the data greatly affects the computing efficiency of feature selection, and sometimes ordinary microcomputers cannot even load all of the data. Therefore, the exploration is more efficient. The feasible feature selection algorithm for massive high-dimensional non-equilibrium data has important theoretical and practical significance. The granular fusion computing theory is an imprecise solution method to study big data's analysis. On the premise of guaranteeing the value of the data, the data scale is reduced, and the input of the problem is converted into multiple information grains from the original big data set. Can significantly reduce the size of the amount of data [2], the application of granular fusion computing theory to large-scale data processing has attracted the attention of many scholars. Liang et al. used clustering technology to granulate big data on cloud platform to reduce the loss of data information and improve the efficiency of time and

S. Liu and L. Xia (Eds.): ADHIP 2020, LNICST 347, pp. 399–408, 2021.
https://doi.org/10.1007/978-3-030-67871-5_36

resource utilization [3]. Yuan et al. aimed at large-scale time series data, the fuzzy information granulation method is used for granulation, and support vector machine is used for regression analysis and prediction on the particle, so as to improve the speed of time series data analysis [4].

Inspired by the above research, this paper proposes and designs a feature selection algorithm based on granular fusion. Making use of the advantages of granulation fusion theory, the selection process of high-dimensional non-equilibrium data sets is transformed into the feature selection process of small data scattered points. In order to ensure the effectiveness of the data feature selection algorithm designed in this paper, the experimental results show that the feature selection algorithm based on particle fusion has the advantages of traditional algorithm, and its accuracy and recall rate are higher than the traditional algorithm. It shows that the proposed algorithm is effective and practical.

2 Design of Feature Selection Algorithm for High-Dimensional Data Based on Granular Fusion

Facing the severe challenge to the traditional feature selection algorithm caused by massive high-dimensional non-equilibrium data, based on the granulation fusion perspective, this paper proposes a feature selection algorithm based on data set based on the granulation fusion theory. The algorithm is divided into three steps: granulation, feature selection on granulated data grains, fusion of granulated features to select [5], as follows:

2.1 High Dimensional Data Granulation Processing

In feature selection of high-dimensional non-equilibrium data set based on granulation fusion theory, statistical random sampling theory is used to split the original large-scale data set in the process of granulation [6]. In determining the size of the data set, the variance of the whole data set must be calculated first, and then, according to the attribute characteristics of the massive data, the hierarchical sampling theory is used to granulate the data set [7]. After granulation, the high-dimensional non-equilibrium large data sets are represented as shown in Fig. 1.

In this paper, the granulation of massive data is realized based on the theory of granulation fusion. The particle size can be calculated directly according to the size of the original data set N and given parameters r, which will greatly improve the efficiency of granulation of massive data, and meanwhile, the data volume contained in each particle is greatly reduced, and the feasibility of data particle selection in a subsequent single-machine environment is ensured [8].

When calculating, first enter the dataset X of the sample size N, and then output the eigenvalues of P data grains:

$$p = \frac{(X - X_P)}{N^r} \tag{1}$$

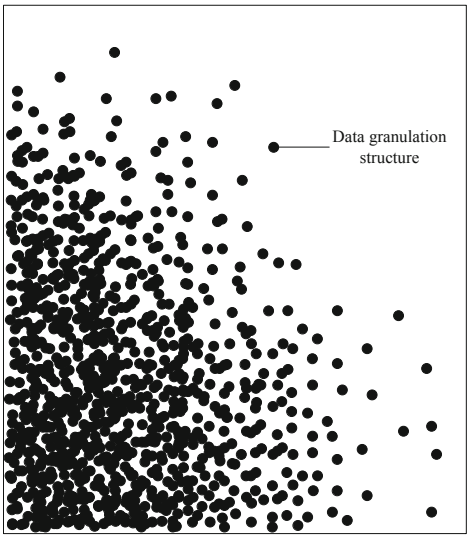


Fig. 1. Representation of granulated large datasets

Of which, p represents the eigenvalues of data grains; According to the theory of granulation fusion, the range of parameter r is as follows: $0.5 \leq r \leq 0.9$.

The above calculation realizes the granulation process of high-dimensional non-equilibrium large data sets, in which the eigenvalue p of the data grains involved can be calculated as a parameter of feature selection. The analysis shows that the p value is determined by X_P and N^r, and the exact p value is obtained, then the backup output is carried out, which provides the basis for the calculation of the characteristics of the data grains in the next step.

2.2 Data Particle Feature Selection

According to the original feature space, the candidate feature subset of the data particle is generated and used as the input of the feature selection algorithm [9]. For the search of feature items, the starting point of the search determines the search direction, so to start the search from the empty set H_i, the subsequent search process is the process of adding the selected features to the candidate feature subset in turn, then the function expression of the forward search is as follows:

$$H = \frac{x|pw_i|}{q_1} + \frac{x|pw_i|}{q_2} + \cdots + \frac{x|pw_i|}{q_n} \tag{2}$$

Of which, H represents forward search term of empty set H_i; $\frac{x|pw_i|}{q_1}$ represents degree of granulation of unbalanced coefficient x when the probability of scatter point is 1. In the same way, $\frac{x|pw_i|}{q_n}$ represents the degree of granulation of the disequilibrium coefficient x when the probability of scatter point is n.

In order to avoid falling into the local optimal linearity, the constraint calculation of the forward search term H is performed using the random search strategy: Assuming that the original feature set contains N_q feature vectors, the candidate feature subset may have $2N_q$ feature vectors. When $H \geq 2N_q$, directly using the exhaustive method to search spatial data features, all feature subsets will be accessed immediately according to the search direction, and the target subset will be marked, and the evaluation criteria will be given. If $H < 2N_q$, then it is necessary to access the optimal results of each feature subset by searching completely, and calculate the complementary balance factor of the data using N_q. On this basis, a feature vector is added from the candidate feature set, and a non vector feature is randomly deleted to improve the efficiency of feature selection in the algorithm. At the same time, the uncertainty due to high computational complexity is avoided [10–13].

Through the above calculation, the initial feature N_q of the data particle is obtained, and N_q is only used as a coefficient reference to prepare for the fusion calculation and selection of the feature subset.

2.3 Feature Subset Fusion Calculation

Feature subset evaluation is one of the important steps in the feature selection process of big data set. Every candidate feature subset needs to be evaluated by evaluation criteria. By introducing the relevant content of granular fusion theory, the fusion calculation process of feature selection of large data sets is transformed into the evaluation process of feature subset. The fusion evaluation form of the subset is shown in Fig. 2.

Figure 2 shows that the fusion of data particle subset mainly includes unidirectional fusion, pattern fusion, euclidean distance fusion, equal distance fusion and independent fusion. Introduce the above model into this calculation and write it down as u_i, and $u_i = \{u_1, u_2, u_3, \cdots, u_i\}$, the independent standard and the fusion standard are combined to evaluate the characteristics by the intrinsic attributes of the data.

At the same time, according to the specific learning algorithm, the distance criterion is used to measure the similarity between sample data in order to represent the contribution or effectiveness of features to classification and recognition. Feature selection using distance measures is generally based on the following assumptions:

The samples belonging to different classes of feature subsets are taken as the distance criterion f_1, and the absolute value coefficients of f_1 is calculated and measured in the form of $f_1 \to f_n$. If that measure set is found to be a null set, that $f_1 \to f_n = 0$, then we stop the transmission of high-dimensional data and continue to wait for the mining function in the data buffer until $f_1 \to f_n \neq 0$, then get the characteristic selection coefficient of the high-dimensional non-equilibrium data grains:

$$\sigma = \frac{1}{f_1 \to f_n \neq 0} \sum_{j=1}^{i} \left| u_{ij} \right| \tag{3}$$

Of which, σ represents characteristic selection coefficients of high-dimensional non-equilibrium data grains; u_{ij} represents Information entropy characteristics of data.

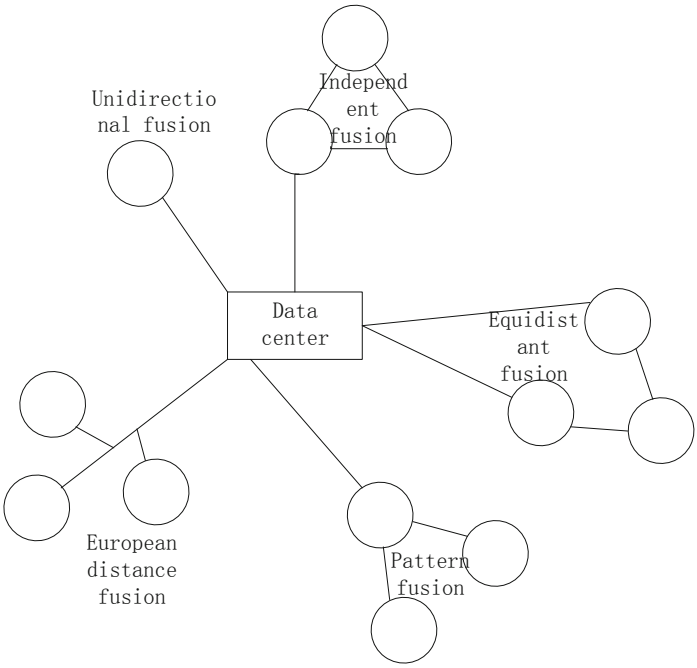


Fig. 2. Fusion form of data granular subset

If the selected big data feature can be constrained by σ, it shows that the feature of the data has the metric and can be recalled effectively. The correlation criterion of σ is used to select the feature subset of the data. The feature of big data can be constrained by σ. This indicates that the feature of the data has the metrology and can be recalled effectively. If the large data characteristic to be selected does not meet the constraint criterion of σ, the measurement performance of the data feature is not high, the effective recall rate is low, the feature selection is abandoned, the selection range is redefined until the large data characteristic to be selected can be restricted by σ, and the characteristic selection of the high-dimensional non-balanced large data set is finished according to the correlation coefficient between the features.

3 Experiment

In order to ensure the validity of the feature selection algorithm based on granulation fusion designed in this paper, the experimental analysis is carried out. The experimental environment is shown in Table 1:

The object of the experiment is a high-dimensional unbalanced and large data set, and the testability detection is carried out. The testability detection results accord with the standard, which shows that the experimental data have practical significance. At the same time, in order to ensure the preciseness of the experiment, the traditional data feature selection algorithm is used for comparison, and the accuracy and recall rate of

Table 1. Configuration information of development environment

Name	Configuration situation
Operating system	Microsoft Windows XP
Processor	Intel(R)Celeron(R) 2.6 GHz
Memory	6.0 GB
Hard disk	4.0 GB
Database management software	Microsoft SQL Server 2010 R2
JDK	1.6
Mathematical software	MATLAB

the two algorithms are counted. In this experiment, the accuracy and recall rate of the feature selection of the high-dimensional unbalanced big data set of the two algorithms are investigated, and the experimental process is supervised by Fisher Score. Therefore, the illustrative result of the experiment can be guaranteed.

Among them, the calculation formula of the accuracy rate and recall rate of big data set feature selection is as follows:

$$\gamma = \frac{A}{(A+B)^2} \tag{4}$$

$$\lambda = \frac{C}{(C+B)^n} \tag{5}$$

Of which, γ represents the accuracy rate of the characteristic selection of the large data set, λ represents the recall rate selected on behalf of the large data set feature, γ and λ are considered in percentage; A represents number of constraints representing data; B represents clustering coefficients with negative data constraints; C represents data set parameter mean; n represents constant, represents the number of experiments.

The experimental process is as follows: firstly, two kinds of feature selection algorithms are used to select the most valuable feature subset, then the original data is projected into the low-dimensional feature subspace to cluster, and the clustering algorithm adopts the simple mean algorithm. Finally, the feature selection results of large data sets are modified and backed up.

The traditional algorithm and this algorithm compare the feature selection recall rate of high-dimensional non-equilibrium large data set as shown in Table 2.

From the Table 1, we can see that the performance of the proposed algorithm is obviously better than that of the traditional feature selection algorithm. Although the value of data is lower than the characteristic value of constraint information, the clustering performance of data is greatly improved after global or local monitoring. The results show that the data feature selection algorithm based on granular fusion can effectively use the theory of granular fusion and unsupervised information for feature selection, improve the data recall rate, and verify the effectiveness of the algorithm.

Table 2. Comparison of maximum recall rates of dataset feature (%)

Data set	Sample	Features	Category	Traditional algorithm	Algorithm in this paper
Heart	270	16	2	66.7	74.6
Sphere	362	36	2	72.1	92.3
Sonar	246	49	3	59.6	86.4
Digits	189	16	4	78.3	91.6
Wine	4563	42	8	92.1	99.6
Image	961	26	9	59.3	86.4
Zoo	143	15	10	89.3	98.9

At the same time, in the data set Wine, the data recall rate of the two feature selection algorithms is ideal, which can make use of more features to achieve the highest clustering performance. However, the proposed algorithm still has some advantages, and its maximum data recall rate is 99.6%. It greatly increases the efficiency of feature selection for large data sets.

The accuracy of feature selection for high-dimensional unbalanced large data sets is compared by the two algorithms, as shown in Fig. 3.

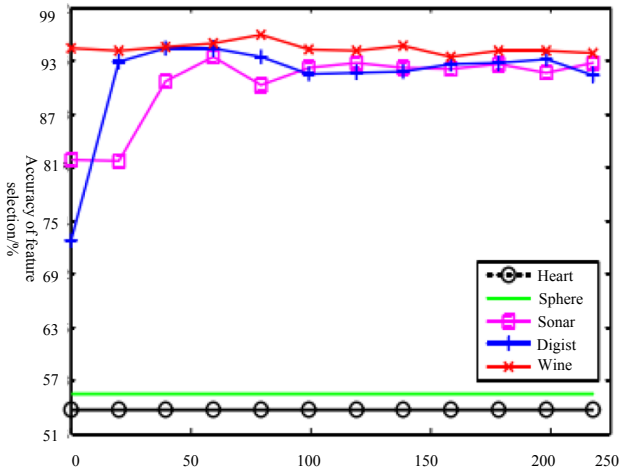
In Fig. 3(a) shows the accuracy of feature selection of large data sets under the traditional method; (b) shows the accuracy rate of feature selection of large data sets in this paper. The analysis shows that, no matter how many pairs of constrained data are used, the accuracy of the proposed method is higher than that of the traditional method. At the same time, for the large data set in the form of global variance, the accuracy of feature selection between the two algorithms is not much different. But we can still see the advantages of this algorithm.

In addition, when using pairwise constraint data for feature selection, the accuracy of this algorithm for Sphere data selection is low, which is due to the linear characteristics of Sphere data. However, the accuracy of feature analysis of linear data is not high, which leads to the low accuracy of feature selection. In addition, all of the algorithms in this paper have overall advantages.

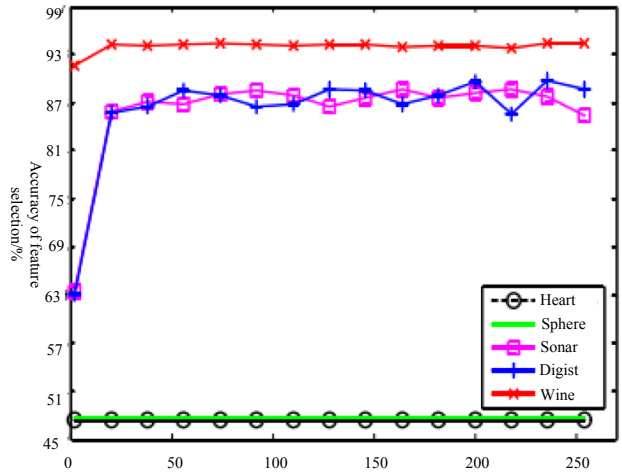
To sum up, the accuracy and recall rate of the proposed algorithm are higher than those of the traditional method, so it can be said that the high-dimensional non-equilibrium big data feature selection algorithm designed in this paper is effective and feasible.

In order to further verify the effectiveness of the algorithm in this paper, the efficient selection time of the high-dimensional unbalanced large data set features of the algorithm in this paper and the traditional algorithm is compared and analyzed. The comparison result is shown in Fig. 4.

According to Fig. 4, as the number of experiments increases, the efficient selection time of the high-dimensional unbalanced large data set features of the algorithm in this paper and the traditional algorithm is gradually increasing, but the efficient selection time of the high-dimensional unbalanced large data set features of the algorithm in this paper is steadily increasing. The time is within 20 s, while the traditional algorithm's high-dimensional unbalanced large data set feature efficient selection time is unstable,



（a）Traditional algorithm



（b）Algorithm in this paper

Fig. 3. Comparison of the accuracy of the feature selection of the big data set

and the time is within 60 s, indicating that the high-dimensional unbalanced large data set feature efficient selection time of this algorithm is short.

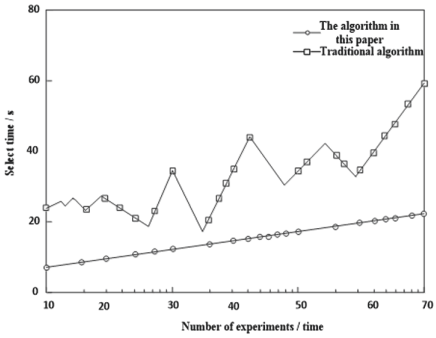


Fig. 4. Comparison and analysis of efficient selection time

4 Conclusion

A feature selection algorithm based on granularity fusion is designed for high-dimensional unbalanced large data sets. The original big data is aggregated into a small-scale data subset by using the regularization features of data. On this basis, the feature selection function of data particles is obtained. Finally, the weight of each feature subset is calculated. F realizes the natural classification of high-dimensional unbalanced large data sets. The effectiveness of the algorithm is verified by experiments. However, there are still a series of deficiencies in the research process of this paper. I hope that the next research can make theoretical analysis and practical test again to improve the applicability of the algorithm.

References

1. Zhe, J., Jianjun, H.: Prediction of lysine glutarylation sites by maximum relevance minimum redundancy feature selection. Anal. Biochem. **550**, 1–7 (2018)
2. Xiaofeng, N.: Research on locally discrete text data mining in high-dimensional data sets. Mod. Electron. Technol. **40**(19), 138–141 (2017)
3. Dan, Z., Chunming, W.: Feature selection based on improved quantum evolutionary algorithm. Comput. Eng. Appl. **54**(1), 146–152 (2018)
4. Hongxiang, D., Qiuyu, Z., Moyi, Z.: FCBF feature selection algorithm based on normalized mutual information. J. Huazhong Univ. Sci. Technol. (Nat. Sci. Ed.) **45**(1), 52–56 (2017)
5. Xin, Z., Haitao, W., Xuehong, C.: Feature selection algorithm based on random forest in Hadoop environment. Comput. Technol. Dev. **28**(255(07)), 94–98+104 (2018)
6. Zhao, X., Zhang, L.: High-dimensional unbalanced data set classification algorithm based on SVM. J. Nanjing Univ. (Nat. Sci.) **54**(2) (2018)
7. Liping, Y., Yunfei, L.: Zhu World Bank: anomaly detection algorithm based on high-dimensional data stream. Comput. Eng. **44**(1), 51–55 (2018)
8. Liu, S., Liu, D., Srivastava, G., et al.: Overview and methods of correlation filter algorithms in object tracking. Complex Intell. Syst. (2020). https://doi.org/10.1007/s40747-020-00161-4
9. Liu, S., Bai, W., Liu, G., et al.: Parallel fractal compression method for big video data. Complexity **2018**, 2016976 (2018). http://doi.org/10.1155/2018/2016976

10. Hongjun, Z.: Research on visualization algorithm of high-dimensional data in multi-dimensional data sets. Microelectron. Comput. **34**(5), 110–113 (2017)
11. Shuai, L., Weiling, B., Nianyin, Z., et al.: A fast fractal based compression for MRI images. IEEE Access **7**, 62412–62420 (2019)
12. Gaber, M.M., Philip, S.Y.U.: Data stream mining in fog computing environment with feature selection using ensemble of swarm search algorithms. New Gener. Comput. **25**(1), 95–115 (2018)
13. Li, J., Liu, H.: Challenges of feature selection for big data analytics. IEEE Intell. Syst. **32**(2), 9–15 (2017)