International Journal of Computational Intelligence Systems (2022) 15:53
https://doi.org/10.1007/s44196-022-00109-1

RESEARCH ARTICLE

A Local Discrete Text Data Mining Method in High-Dimensional Data Space

Juan Li[1] · **Aiping Chen**[1]

Received: 20 February 2022 / Accepted: 14 July 2022

Abstract
Aiming at the problems of low accuracy, the long time required, and the large memory consumption of traditional data mining methods, a local discrete text data mining method in high-dimensional data space is proposed. First of all, through the data preparation and preprocessing step, we obtain the minimum data divergence and maximize the data dimension to meet the demand for data in high-dimensional space; second, we use the information gain method to mine the pre-processed discrete text data to establish an objective function to obtain the highest information gain; finally, the objective functions established in data preparation, preprocessing, and mining are combined to form a multi-objective optimization problem to realize local discrete text data mining. The simulation experiment results show that our method effectively reduces the time and improves the accuracy of data mining, where it also consumes less memory, indicating that the multi-objective optimization method can effectively solve multiple problems and effectively improve the data mining effect.

Keywords High-dimensional data space · Local discrete text · Data mining · Fuzzy rules · MTPIG algorithm · Multi-objective optimization

Abbreviation

MTPIG Multi-interval discretization based on term presence and information gain

1 Introduction

Against the background of the rapid development of network technology and the continuous increase in users, the amount of data both in reality and in network presents an exponential rising trend. In the face of massive data, how to find the required data has become a hot issue in current research in relevant fields [1]. How to obtain the law of data generation and development through mining and processing a large amount of data, and extract valuable information from it, has positive significance for the improvement of data application [2, 3]. Among various data types, there are not only normal data, but also some discrete data, which have certain limitations in collection, classification, retrieval, and mining due to their own particularities [4]. Therefore, to realize the effective mining of local discrete text data, it must be further processed.

At present, in the field of data mining, relevant scholars have put forward a large number of research methods, and the application effect of the methods has been verified, providing some valuable references for data processing. Among them, Christian proposed community resources for paired genome and metabolome data mining in high-dimensional data spaces. Genomics and metabolome are widely used to explore specific metabolite diversity. The PENTIomics Data Platform is a community initiative designed to systematically document the links between metabolome and (meta-) genomic data, helping to identify sources of natural product biosynthesis and metabolite structures [5]. Fernandez-basso et al. proposed a fuzzy mining method for energy efficiency in the framework of big data. The discovery and utilization of hidden information in the collected data has attracted attention in many fields, especially in the field of energy, because of their impact on the economy and the environment. Data mining techniques have thus become a suitable toolbox for analyzing the data collected in modern network management systems to gain a meaningful understanding of consumption patterns and device operation. Big data offers tremendous opportunities to implement new

✉ Juan Li
iamlj6@jit.edu.cn

[1] School of Computer Engineering, Jinling Institute of Technology, Nanjing 211169, China

Published online: 29 July 2022

solutions to manage these massive data sets. In addition, the value presented by these data, by their nature, complicates and conceals the understanding and interpretation of data and results. Therefore, using fuzzy methods to fully transform data can improve the interpretability of data. An automatic fuzzification method using a big data paradigm is introduced, which can detect interrelationships and patterns between different sensors and weather data recovered from office buildings in subsequent steps [6]. Shang et al., on the basis of artificial intelligence technology, such as the local outlier data mining method is put forward. Before data mining, first by feature extraction steps to get the data characteristics, and then based on the data characteristics of the test information entropy for preprocessing result. Finally, according to the result of data preprocessing, data mining using neural network, access to the data mining results. Research results show that this method addresses the problem that traditional methods are only effective for normal data mining, and improves the comprehensiveness of data mining methods [7].

It can be seen from the analysis that the above traditional methods can obtain the effect of data mining, but usually they can obtain better mining results when mining normal data. When facing local discrete data, there are problems of low accuracy and extra running time of data mining. Although the reference method improves the comprehensiveness of data mining to a certain extent, there is still a problem of large memory consumption in data mining.

Aiming to address the above problems, this paper proposes a local discrete text data mining method in high-dimensional data space, hoping to realize accurate and fast mining of local discrete text data without consuming too much memory. The main research content and innovation points of this method are described as follows: through data preparation and preprocessing, the objective function is established to minimize data divergence and maximize the data dimension, so that discrete text data can meet the demand for data in high-dimensional space. Based on the preprocessing results, the information gain method is used to mine the data, and the information gain is asymmetric and measures the difference between two probability distributions, P and Q. The information gain describes the difference between encoding using Q and then P. Usually P represents the distribution of samples or observed values, or it may be a theoretical distribution that is accurately calculated. Q represents a model that describes an approximation to P, and the objective function is established to obtain the highest information gain. By integrating all objective functions, the data mining problem is transformed into a multi-objective optimization problem to improve the data mining effect from multiple perspectives. By analyzing the experimental results, it can be seen that the proposed method can effectively improve the problem of low accuracy of data mining existing in traditional methods. Big data mining takes less time and consumes less memory, so it can meet the demand of low energy consumption while meeting the mining effect.

2 Local Discrete Text Data in High-Dimensional Data Sets

The rapid development of the Internet has brought a large number of data, including not only normal data, but also some discrete text data. These latter data are different from normal data and will bring some difficulties to data application. Therefore, they need special processing [8]. This paper will study this kind of data mining, specifically through feature clustering, preprocessing, and multi-objective optimization to eliminate interference data, so as to improve the efficiency and accuracy of data mining. Figure 1 shows the process of local discrete text data mining in high-dimensional data sets.

According to Fig. 1, the association degree mining method is used to extract the features of data samples from local discrete text data, and the association rules are used to solve the probability of data mining. Through the above steps, the description of the internal features of local discrete text data

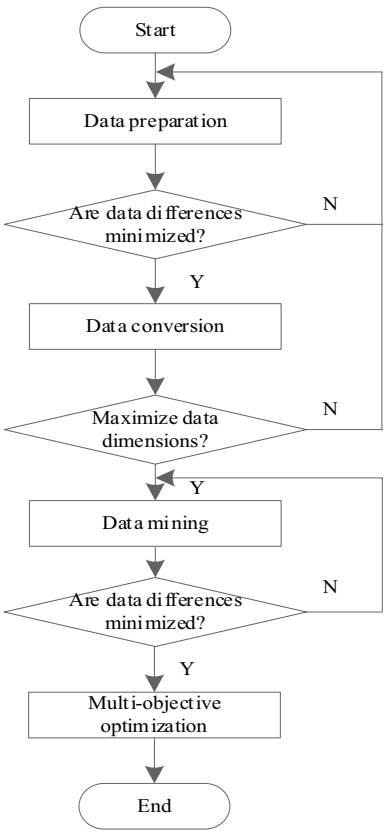

Fig. 1 Flow chart of local discrete text data mining

can be better realized, multi-objective optimization can be carried out, and the data mining process can be completed.

2.1 Discrete Text Data Feature Clustering

2.1.1 Discrete Text Data Preparation

Discrete text data preparation refers to the definition, representation, and processing of mining objects so that the data can be adapted to specific data mining requirements. In the actual process of data mining, to improve the mining efficiency and reduce the impact of invalid data and interference data on the mining results, it is necessary to prepare the data first, and the processing quality of this link directly affects the data mining effect [9].

Discrete text data refers to data with discrete features, which usually have different states, different attributes, and other characteristics. Usually, the processing effect of discrete data is not as good as that of continuous data. Therefore, the discrete text data in the high-dimensional data space is processed. Before mining, it is first clustered to obtain target data in a high-dimensional data space [10, 11]. This paper mainly uses fuzzy theory for data clustering, assuming that there is a data sample set $\partial=\{\partial_1, \partial_2, \ldots, \partial_n\}$, where n represents the number of samples. Combine the association rules to extract the data with similar features in the data set, use W to represent the data projection space normal vector, ∂_k to represent the projection of the sample data ∂_i in the data sample set, and obtain the data clustering center, which is represented by formula (1):

$$R^2 = \sqrt{x_i^2(\alpha,\beta) + y_i^2(\alpha,\beta)} \tag{1}$$

Among them, x_i^2 represents the density of each point in the data set; y_i^2 represents the distance distribution of each point in the data set; α represents the center point of the sequence; and β represents the cluster center.

The internal divergence of the sample data is further reduced according to fuzzy theory, which refers to the basic concept of fuzzy set or the theory of continuous membership function. From the point of view of practical applications, fuzzy theory is mainly applied to fuzzy systems, especially fuzzy control, and some fuzzy expert systems are applied to medical diagnosis and decision support. As fuzzy theory is still new from the point of view of theory and practice, more reliable practical applications will appear as the fuzzy field matures. The application of fuzzy theory is the most effective, and the most extensive field is fuzzy control. Fuzzy control in a variety of fields unexpected to solve the traditional control theory can not solve or difficult to solve the problem, and achieved some convincing results. The specific formula is as follows:

$$\partial_i(x) = K_i(x) \times \frac{R^2(x_i, x_j)}{S^2} \tag{2}$$

Among them, $K_i(x)$ represents the central tendency of the data; x_i represents the overall dispersion index; x_j represents the medium and high value dispersion index; and S^2 represents the data feature space.

Under the effect of formula (2), the data divergence problem is transformed into a minimization objective function, and the optimal clustering center of the data is obtained by solving it [12, 13]. The specific formula is as follows:

$$R' = S^2\left[F_i(x), F_j(x)\right]^T \tag{3}$$

Among them, $F_i(x)$ and $F_j(x)$ both represent the feature aggregation of similar target data; and T represent the number of iterations.

According to the calculation result of formula (3), the internal divergence of the sample data is minimized after the R' value is obtained.

2.1.2 Discrete Text Data Preprocessing

Based on the preparation of discrete text data, to further provide a reliable data foundation for data mining, we continue to preprocess the target data. Since discrete text data have a large number of unusable data without processing, it is necessary to transform these data in advance in the preprocessing process, then the transformed data can adapt to the high-dimensional data space [14, 15]. In this process, an objective function $\mu(f)$ is established, which is used to express the dimension of the data. To adapt to the high-dimensional space and maximize $\mu(f)$, the specific formula is as follows:

$$\mu(f)_{\max} = \begin{cases} G_r | a_i \le a_1 \quad i = 1 \\ G_r | a_{i-1} < a_i \le a_k \quad 1 < i \le k \end{cases} \tag{4}$$

Among them, G_r represents the support threshold; and a_i represents the dimension of the high-dimensional space data item.

The conversion of high-dimensional converted data is mainly realized by the MTPIG (Multi-interval Discretization based on Term Presence and Information Gain) algorithm. The core idea of the algorithm is to treat the high-dimensional space as a continuous space and divide the data in it [16, 17]. We first divide the attribute data in the continuous space to form an ordered subspace:

$$Q(x) = \left[q_1(x) + q_2(x) + \cdots + q_n(x)\right] \tag{5}$$

then calculate the probability of A attribute data appearing in this subspace after division:

$$A_{ij} = \begin{cases} 1 & \text{if } x_i = x_j \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

Among them, 1 represents that the A attribute data appears in the subspace, and 0 represents that the A attribute data does not appear in the subspace. Replace the internal divergence of the sample data with 1 or 0, thereby transforming the attributes of the sample data to adapt to discrete text data mining in a high-dimensional space.

2.2 Discrete Text Data Mining

Normally, the pre-processed data dimension can meet the requirements of high-dimensional space, but because part of the data is protected, it will also affect the discreteness of the data itself [18]. Therefore, this paper uses the information gain method to mine the pre-processed discrete text data to obtain complete and accurate data mining results [19, 20]. Let H_m denote the sample set containing m types of attribute data, and denote the information gain by ψ_z. The calculation formula is:

$$\psi_z = \frac{\int_{\Omega} H_m f(x_i, x_j) \mathrm{d}\Omega}{\psi_{m'}} \tag{7}$$

Among them, Ω represents the number of best features of the data; and $\psi_{m'}$ represents the information gain between a single feature and the entire data set.

Assuming that M_l represents the number of samples in H_m, the expected information required for data mining is expressed by formula (8):

$$\psi(d_{ij}) = \sum_{i,j=1}^{M} C_i \left| \delta^{(i)}(d_j) \right|^2 \tag{8}$$

Among them, C_i represents the attribute class; $\delta^{(i)}$ represents the attribute subset; and d_j represents the probability that the sample belongs to C_i. Considering the different attributes of the data, the high-dimensional space is divided into different subsets $Y = \{y_1, y_2, ..., y_m\}$, and the discrete text data of an attribute is selected arbitrarily in the subset, and it is mined accordingly. Firstly, the information encoding of the attribute data in the subspace is obtained, and then the information gain of the data attribute is calculated [21]. At this point, the highest information gain is obtained by establishing the form of the objective function:

$$\psi_{\max} = 1 - \frac{H_{\mathrm{ac}}}{\max(H_{\mathrm{ac}}) + \psi_z} \tag{9}$$

Among them, H_{ac} represents the attribute mark threshold.

For the information gain calculated above and the maximum information gain under the constraint of the objective function, we can further mine the discrete text data in the high-dimensional space, and establish a branch for each data attribute to form a decision tree to determine whether the data belong to the branch. If they belong, they will be classified into one class. If not, it will establish a subtree for further data mining, until the final mining results are obtained.

2.3 Multi-Objective Optimization Discrete Text Data Mining Method

Through the above steps, discrete text data can be mined. To improve the effectiveness of data mining, discrete text data preparation, discrete text data preprocessing and objective function established in discrete text data mining are combined to form a multi-objective optimization problem [22, 23]. Under the condition of meeting multiple optimization objectives, the optimal mining of discrete text data can be realized. Based on the objective function formula established above, considering the characteristics of discrete data and the particularity of its high-dimensional space, an equation system is established:

$$\begin{cases} R' = S^2 \left[F_i(x), F_j(x) \right]^T \\ \mu(f)_{\max} = G_r | a_{i-1} \\ \psi_{\max} = 1 - \dfrac{H_{\mathrm{ac}}}{\max(H_{\mathrm{ac}}) + \psi_z} \end{cases} \tag{10}$$

To ensure the dimension of high-dimensional spatial data items, formula (4) is simplified. The optimal value of multi-objective optimization [24–26] is obtained by solving formula (10). Before solving the optimal value, specific constraints need to be set:

$$\sigma_1 > ... \sigma_L > \sigma_{L+1} (L = 1, 2, ..., N) \tag{11}$$

Herein, σ represents the mining probability of data samples in the high-dimensional space; and L represents the frequency of occurrence of the objective function.

Under the constraint conditions shown in formula (11), the optimal value of multi-objective optimization is calculated [27, 28], and the formula is as follows:

$$N_i(k) = \left\| R' \times \psi_{\max} \right\|^2 \times \mu(f)_{\max} \tag{12}$$

According to the above analysis, the multi-objective optimization of discrete text data mining is realized by combining multiple objective functions.

2.4 Local Discrete Text Data Mining in High-Dimensional Data Sets

It can be seen theoretically from the above analysis that the multi-objective optimization method can obtain the results of local discrete text data mining in high-dimensional data

sets. The following summarizes our specific data mining process:

Step 1: Use fuzzy theory to cluster data, and combine association rules to extract data with similar characteristics, establish a data divergence minimization objective function, and obtain the optimal clustering center in the data extraction results [29, 30];

Step 2: Based on the results of data preparation, establish an objective function $\mu(f)$ and obtain its maximum value $\mu(f)_{\max}$, which is the highest dimension of the data, so that the data can adapt to the high-dimensional space [31–33]; If the maximum value can be obtained, proceed to the next step; If the maximum value is not obtained, return to step 1;

Step 3: Establish the objective function again, calculate the information gain of the data attribute, and obtain the highest information gain;

Step 4: Combine multiple objective functions to form a multi-objective optimization mode, establish constraint conditions, and if the conditions are met, the optimization of local discrete text data mining is realized by finding the optimal solution [34, 35]. If the constraint conditions are not met, return to step 3 until the conditions are met.

3 Experimental Analysis

To verify the practical value of the local discrete text data mining method in the improved dimension data space and to fully verify its theoretical and practical significance, a simulation experiment is conducted. In the experiment, the big data mining algorithm based on semantic relevance feature fusion (i.e., reference [5] method) and the data mining method integrating improved genetic algorithm and association rules (i.e., reference [6] method) are used as comparison methods, and the data mining accuracy, mining time, and memory consumption are taken as experimental evaluation metrics. The specific experimental design and experimental results are analyzed as follows.

3.1 Experimental Environment and Data Set

The data used in the experiment comes from the UCI database, which is an open source database. The data needed for the experiment can be selected arbitrarily from the UCI database. Therefore, for this paper we selected some samples from the database as the data required by the experiment, and divided them into five datasets, namely Monk, Letter, Vote, Banding, and Hypo, which were uniformly named dataset 1, dataset 2, dataset 3, dataset 4, and dataset 5. Among them, dataset 1-dataset 4 was the training sample set, and dataset 5 was the test sample set. To ensure the unity of experimental conditions and to avoid the influence of differentiation on the accuracy of the experimental results, before mining the characteristics of local discrete text data, we first cluster the data. The experiment sets a total of 5 groups of high-dimensional text data, with 200 samples in each group. It is required to cluster 5 data clusters and 20 high-dimensional text data in each cluster. The 100 text data given by the experiment are clustered by different methods, the cluster results and the number of text data contained in each cluster are obtained, and the results are compared with the set results. Setting the difference threshold of all data sets to 0.01, we conduct multiple experiments in the experiment to obtain the average value as the final experimental result. Under this condition, the detailed information of the experimental data set is given in Table 1.

The above experimental data were loaded in the hardware environment with an Intel Core 8I7-10700F processor, iGame GeForce RTX 3060 Ultra WOC graphics card, and 970 EVO Plus 500G NVMe M. 2 hard disk. The MATLAB software was used to process the experimental data.

3.2 Analysis of Experimental Results

Under the above experimental environment settings, feature clustering was carried out on the data before data feature mining. Eight groups of high-dimensional text data were set in the experiment, with each group containing 400 samples, and five data clusters were required to be clustered, each cluster containing 80 high-dimensional text data. In this paper, the multi-objective soft subspace clustering method and data flow soft subspace clustering method are used to cluster the 400 text data presented in the experiment. After clustering, the data cluster results and the number of text data contained in each cluster were obtained, and the results were compared with the set results. Based on the experimental data in Table 1, the methods in this paper, reference [5], and reference [6] were compared. The comparison results are analyzed in detail below.

3.2.1 Data Mining Accuracy Verification

To compare the data mining accuracy of the different methods, the accuracy was calculated as follows:

Table 1 Experimental data set information

Data set name	Data volume/GB	Number of data types/piece	Number of data attributes/piece
Data set 1	1087	15	2
Data set 2	396	11	4
Data set 3	1024	24	4
Data set 4	843	19	5
Data set 5	650	13	3

$$Z_{\mathrm{QL}} = \frac{1}{m}\left(k - k'\right) \times 100\% \tag{13}$$

In formula (13), m represents the number of data mining items and k represents the actual number of data mining items; k' represents the predicted number of data mining items. According to formula (13), different methods were used to compare the data mining accuracy, and the results are shown in Fig. 2:

According to the analysis of Fig. 2, under the condition of increasing data volume, the data mining accuracy of different methods shows a linear growth trend. When the data volume is 1600 GB, the data mining accuracy of this method is 61%, and the data mining accuracy of the reference [5] method and the reference [6] method are 34% and 41% respectively; when the data volume is 4000 GB, the data mining accuracy of this method is 95%, and the data mining accuracy of the reference [5] method and the reference [6] method are 55% and 75% respectively. It can be seen from the above result that the data mining accuracy of this method is significantly higher than that of traditional methods.

3.2.2 Data Mining Time Verification

Using different methods to mine the local discrete text data in data set 1-data set 5, we compared the data mining time of different methods. Figure 3 shows the comparison results:

According to Fig. 3, the method in this paper has the shortest mining time for dataset 5, which is only 1.7 min; the method in Reference [5] has the shortest mining time for dataset 3, which is 6.0 min; and the method in reference [6] has the shortest mining time for dataset 4, which is 4.9 min. In addition, the data mining time of the proposed method for the five data sets is lower than that of the two traditional methods. Therefore, it can be seen that the proposed method

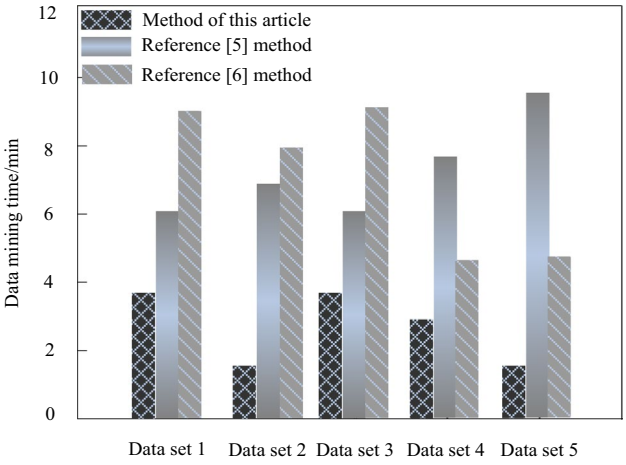

Fig. 3 Comparison results of data mining time

can achieve the research goal of fast mining of local discrete text data.

3.2.3 Memory Consumption Verification

To better verify the good performance of the method in this paper, different methods are compared from the perspective of memory consumption. The comparison is shown in Fig. 4.

It can be seen from Fig. 4 that at the beginning of the experiment, the memory consumption of the three methods is at a low level. With the increase in the number of iterations, the memory consumption of different methods increases gradually. Among them, the memory volume increasing range of this method is the smallest, that is, the memory consumption of this method in local discrete text data mining is the smallest, followed by the method in reference [6], whereas the memory consumption of the method

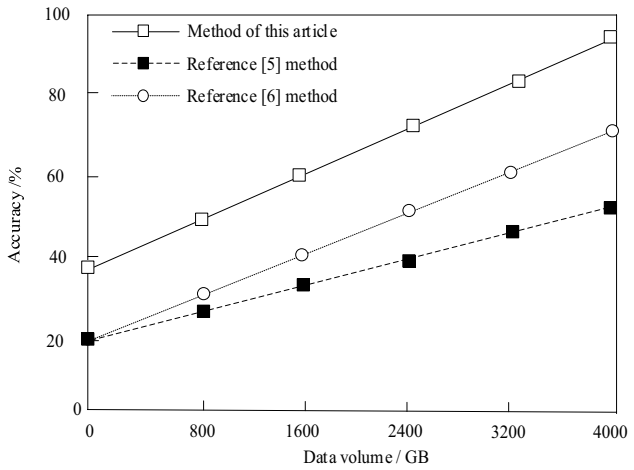

Fig. 2 Comparison results of data mining accuracy

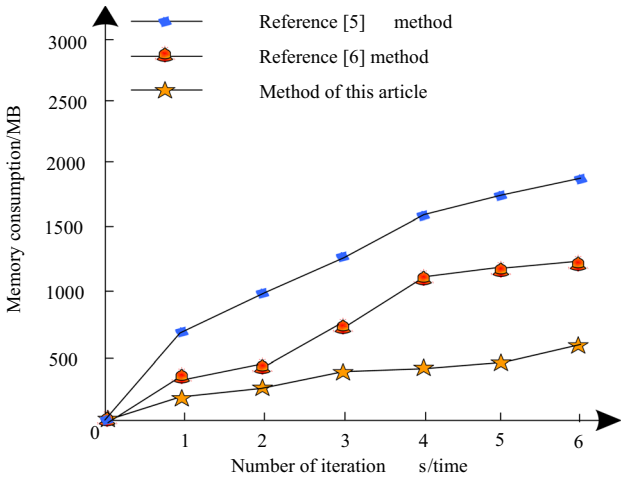

Fig. 4 Memory consumption comparison results

in reference [5] is the largest. Through comparison, it can be seen that the practical application value of this method is higher, which can save a great deal of system memory and improve the space accommodation.

In conclusion, compared with other methods, the local discrete text data mining method in high-dimensional data space has higher accuracy, shorter data mining time, minimum memory consumption, and the highest application value.

4 Discussion and Analysis

The data set was from the access records saved in the Web log of a website of a software certification training center on a certain day. The users of the website use the local discrete text data mining method in the high-dimensional data space proposed in this chapter to perform secondary processing on the search engine when conducting Web information retrieval, accessing approximate Web page clusters. The purpose of the research is to recognize the browsing path of users through Web click flow, so as to predict the sequence of users browsing Web pages, and to sort the word segmentation dictionary of related topics accordingly, so as to obtain the order of similar Web page clusters in the next search. This can make it conform to the interests of users. Data mining is a method of effectively classifying the features of data and excavating its internal correlation. It has been widely used in many scientific fields. The data mining algorithm is simple, fast, scalable, and maintainable. Users can change the keyword database they are interested in according to their own needs, which is convenient for users to find the information they really care about and has high application value.

5 Conclusion

In the era of big data, mining and applying effective information with the explosive growth of data is the key link to improving the quality of data use. However, when facing special data types, traditional methods often cannot show the consumption of mining performance. Therefore, this paper proposes a local data mining method in high-dimensional data space. This method improves the efficiency of data mining through discrete text data preprocessing, and improves the accuracy of local discrete text data mining through multi-objective optimization. The results show that the data mining accuracy of this method is relatively high, the data mining time is short, and the memory capacity consumed is the least. The application value of this method is the highest.

Although this method has achieved the preset goal, due to the real-time change of data in the big data environment, this feature will bring some difficulty to data mining. Next, this problem will be fully considered and the method in this paper will be further optimized to adapt to the big data environment.

Acknowledgements We thank the editorial board and all reviewers for their professional advises to improve this work.

Author contributions JL is a director of this work, who contributed to draft the manuscript. She investigated the research object. AC helped to draft the manuscript and investigated the research object. All authors read and approved the final manuscript.

Funding This research is funded by the Jiangsu Higher Education Reform Research Project (2021jsjg641), the Jiangsu Educational Science "14th five-year plan" Project (B/2021/01/13).

Availability of data and materials The data that support the findings of this study are available on request from the corresponding author Juan Li.

Declarations

Conflict of interest The authors have no competing interests to declare that are relevant to the content of this article.

Ethics approval and consent to participate Not applicable.

Consent for publication Not applicable.

References

1. Zhao, W., Luo, Z.: Web text data mining method based on Bayesian network with fuzzy algorithms. J. Intell. Fuzzy Syst. **38**(4), 1–9 (2020)
2. Zhou, J., Guo, Y., Sun, Y., Wu, K.: Data mining method based on rough set and fuzzy neural network. J. Intell. Fuzzy Syst. **38**(2–3), 1–9 (2020)
3. Gao, J., Liu, J., Guo, S., Zhang, Q., Wang, X.: A data mining method using deep learning for anomaly detection in cloud computing environment. Math. Probl. Eng. **2020**(1), 1–11 (2020)
4. Radhika, A., Masood, M.S.: Effective dimensionality reduction by using soft computing method in data mining techniques. Soft. Comput. **25**(2), 1–9 (2021)
5. Christian, M.H.: A community resource for paired genomic and metabolomic data mining. Nat. Chem. Biol. **17**(4), 363–340 (2021)

6. Fernandez-Basso, C., Ruiz, M.D., Martin-Bautista, M.J.: A fuzzy mining approach for energy efficiency in a Big Data framework. IEEE Trans. Fuzzy Syst. **28**(11), 2747–2758 (2020)
7. Shang, F.H., Cao, M.J., Wang, C.Z.: Local outlier data mining based on artificial intelligence technology. J. Jilin Univ. (Eng. Technol. Ed.) **51**(2), 692–696 (2021)
8. Salehi, H., Das, S., Biswas, S., Burgueno, R.: Data mining methodology employing artificial intelligence and a probabilistic approach for energy-efficient structural health monitoring with noisy and delayed signals. Expert Syst. Appl. **135**(11), 259–272 (2019)
9. Follett, L., Geletta, S., Laugerman, M.: Quantifying risk associated with clinical trial termination: a text mining approach. Inf. Process. Manage. **56**(3), 516–525 (2019)
10. Kim, L., Ju, J.: Can media forecast technological progress? A text-mining approach to the on-line newspaper and blog's representation of prospective industrial technologies. Inf. Process. Manage. **55**(4), 1506–1525 (2019)
11. Deotale, R., Rawat, S., Vijayarajan, V., Prasath, V.B.S.: POCASUM: policy categorizer and summarizer based on text mining and machine learning. Soft. Comput. **25**(14), 9365–9375 (2021)
12. Rivera-Quiroz, F.A., Petcharad, B., Miller, J.A.: Mining data from legacy taxonomic literature and application for sampling spiders of the Teutamus group (Araneae; Liocranidae) in Southeast Asia. Sci. Rep. **10**(1), 15787 (2020)
13. Alex, B., Grover, C., Tobin, R., Sudlow, C., Whiteley, W.: Text mining brain imaging reports. J. Biomed. Semant. **10**(1), 23 (2019)
14. He, B.R.: Simulation of time series data mining algorithm based on multi-objective decision. Comput. Simul. **36**(11), 243–246 (2019)
15. Borsato, N.W., Martell, S.L., Simpson, J.D.: Identifying stellar streams in Gaia DR2 with data mining techniques. Mon. Not. R. Astron. Soc. **492**(1), 1370–1384 (2020)
16. Tinoco, J., Granrut, M.D., Dias, D., Miranda, T., Simon, A.G.: Piezometric level prediction based on data mining techniques. Neural Comput. Appl. **32**(1), 4009–4024 (2020)
17. Campo-Vila, J.D., Takilalte, A., Bifet, A., Mora-López, L.: Binding data mining and expert knowledge for one-day-ahead prediction of hourly global solar radiation. Expert Syst. Appl. **167**(8), 114147 (2020)
18. Shafiabadi, M., Pedram, H., Reshadi, M., Reza, A.: An accurate model to predict the performance of graphical processors using data mining and regression theory. Comput. Electr. Eng. **90**(1), 106965 (2021)
19. Kazanidis, I., Valsamidis, S., Gounopoulos, E., Kontogiannis, S.: Proposed S-Algo+data mining algorithm for web platforms course content and usage evaluation. Soft. Comput. **24**(19), 14861–14883 (2020)
20. Nguyen, T.V., Zhou, L., Chong, A., Li, B., Pu, X.: Predicting customer demand for remanufactured products: a data-mining approach. Eur. J. Oper. Res. **281**(3), 543–558 (2020)
21. Sharma, G., Sazim, S., Pati, A.K.: Quantum coherence, coherent information and information gain in quantum measurement. EPL (Europhys. Lett.) **127**(5), 50004 (2019)
22. Mittal, S., Shukla, D.: Simulation guided design of spectroscopy experiments via maximizing kinetic information gain. Biophys. J. **116**(3), 183–184 (2019)
23. Kelly, J., Leahy, P.G.: Sizing battery energy storage systems: using multi-objective optimization to overcome the investment scale problem of annual worth. IEEE Trans. Sustain. Energy **11**(4), 2305–2314 (2020)
24. Srinivasan, B., Venkatesan, R.: Multi-objective optimization for energy and heat-aware VLSI floor planning using enhanced firefly optimization. Soft. Comput. **25**(5), 4159–4174 (2021)
25. Tam, N.T., Hung, T.H., Binh, H., Le, T.V.: A decomposition-based multi-objective optimization approach for balancing the energy consumption of wireless sensor networks. Appl. Soft Comput. **107**(2), 107365 (2021)
26. Grishchenko, A.V., Kruchek, V.A., Kurilkin, D.N., Khamidov, O.R.: Diagnostics of the technical condition of rolling bearings of asynchronous traction motors of locomotives based on data mining. Russ. Electr. Eng. **91**(10), 593–596 (2020)
27. Shichkina, Y., Irishina, Y., Stanevich, E., Salgueiro, A.D.J.P.: The main aspects of creating a system of data mining on the status of patients with Parkinson's disease. Procedia Comput. Sci. **186**(9), 161–168 (2021)
28. Taranto-Vera, G., Galindo-Villardón, P., Merchán-Sánchez-Jara, J., Salazar-Pozo, J., Moreno-Salazar, A., Salazar-Villalva, V.: Algorithms and software for data mining and machine learning: a critical comparative view from a systematic review of the literature. J. Supercomput. **77**(10), 11481–11513 (2021)
29. Sun, Z.J., Duncan, A., Kim, Y., Zeigler, K.: Seeking frequent episodes in baseline data of in-situ decommissioning (ISD) Sensor network test bed with temporal data mining tools. Prog. Nucl. Energy **125**(3212), 103372 (2020)
30. Bruch, S., Ernst, L., Schulz, M., Zieglowski, L., Tolba, R.H.: Best variable identification by means of data-mining and cooperative game theory. J. Biomed. Inform. **113**(7), 103625 (2020)
31. Mohamed, A., Molendijk, J., Hill, M.: Lipidr: a software tool for data mining and analysis of lipidomics datasets. J. Proteome Res. **19**(7), 2890–2897 (2020)
32. Yang, T., Zhang, L., Kim, T., Hong, Y., Peng, Q.: A large-scale comparison of artificial intelligence and data mining (AI&DM) techniques in simulating reservoir releases over the upper colorado region. J. Hydrol. **602**(6), 126723 (2021)
33. Guo, A., Jiang, A., Lin, J., Li, X.: Data mining algorithms for bridge health monitoring: Kohonen clustering and LSTM prediction approaches. J. Supercomput. **76**(2), 932–947 (2020)
34. Luo, Z., Hong, S.H., Ding, Y.M.: A data mining-driven incentive-based demand response scheme for a virtual power plant. Appl. Energy **239**(4), 549–559 (2019)
35. Liu, J., Dong, H., Wang, P.: Multi-fidelity global optimization using a data-mining strategy for computationally intensive black-box problems. Knowl.-Based Syst. **227**(3), 107212 (2021)