Hindawi
Mathematical Problems in Engineering
Volume 2022, Article ID 3966850, 9 pages
https://doi.org/10.1155/2022/3966850

Research Article

Classification Algorithm for Library Electronic Documents Based on Continuous Attribute

Jin Zeng

Library of Army Medical University, Chongqing 400030, China

Correspondence should be addressed to Jin Zeng; 2016150290@jou.edu.cn

Received 9 January 2022; Revised 14 March 2022; Accepted 21 March 2022; Published 16 April 2022

Academic Editor: Dost Muhammad Khan

The automatic classification of document data will occupy an increasingly important position in digital libraries. Generally, the kernel method based on support vector machine is used to classify literature data on the standard test set, which has some shortcomings. In order to solve these problems, vocabulary expansion is used to preprocess the document vector to obtain a small but precise, orthogonal, and unambiguous new document vector; the document vector is sorted according to semantics to improve the access and calculation speed; the document is mapped to Lz with the help of wavelet kernel space for document classification. This paper analyzes the existing continuous attribute discretization methods in detail, discusses how to reduce the loss of information in the discretization process, and proposes a low-frequency discretization (LFD) algorithm based on the attribute low-frequency region. This method effectively reduces data loss by setting the segmentation point in the attribute interval with lower frequency, and through the research and analysis of the existing association rule mining algorithm, this paper combines low-frequency discretization, weighted multiple minimum support, and full confidence, and a weighted multiple minimum support association rule mining algorithm based on low-frequency discretization (WM-SamplingHT) is proposed. The algorithm first uses the low-frequency discretization algorithm to discretize the continuous attributes, then sets the respective weights and minimum support for the data items when mining frequent itemsets, removes the false patterns through the full confidence, and then obtains cleaner frequent itemsets. Using the real classification data of China Academic Journals Network, it is verified from the perspectives of abstract information and full-text documents. The results show that this method is superior to the nuclear method and has certain theoretical research and practical applications.

1. Introduction

In the age of scientific and technological information, with the rapid development of computer technology and the widespread application of large databases, information has shown explosive growth [1]. With the emergence of massive amounts of information, problems such as information overload, difficulty in distinguishing between true and false information, and the inability to guarantee information security have made it very difficult to process information. "Information is rich, knowledge is lacking" has become a research hotspot in data mining and related fields [2]. Data mining is to analyze a large amount of data, dig out potential and valuable information from it, and then use it to provide guidance for people's study and life. Data mining integrates knowledge in many fields such as data warehouse [3], machine learning, distributed computing [4], algorithm analysis, statistics, information retrieval [5], and artificial intelligence.

The ever-developing digital library (DLa), with its new digital media form, has greatly enriched the types and quantity of digital objects [6]. At the same time, in the broader digital protection (DP) and information life cycle management, the indexing and classification of these digital objects have become a part of access management and data. In recent years, researchers have used machine learning (ML) to assist certain regular circulation services of digital libraries, which has produced certain effects. These applications mainly include extracting image content from pictures of scientific and technological literature for

classification [7]; automatically assessing the resource quality of educational digital libraries [8] and characterizing their characteristics; assessing the quality of scientific and technological conferences [9]; web-based collection development; using support vector machines (SVM) automatically extracting document metadata; automatically extracting titles from general documents; constructing the information architecture of the digital library; eliminating duplicate documents; collaborative filtering and vocabulary classification for automatic expansion of domain-specific vocabulary; generation of visual thesaurus and documents semantic markup [10].

With the gradual expansion of the scale of digital libraries, automatic document classification (TAC) has become more and more important and will occupy an increasingly important position in digital libraries [11–13]. In this field, the current foreign research uses support vector machines (SVM) to conduct experimental analysis on the standard test set. Domestic research on the automatic classification of documents began in 1985. After more than 20 years of development, some important progress has been made successively, such as algorithm research, construction of knowledge bases, development of automatic classification systems for Chinese documents, and automatic Chinese documents based on "Chinese Library Classification" [14, 15]. Classification was conducted through an automatic classification method of Chinese documents based on N-Gram technology and research on automatic classification of scientific and technological documents based on SVM (support vector machine) and KNN algorithm. The shortcomings of the above research are mainly manifested in the following. (i) The traditional SVM classification method has the defects of the large scale of document vector, nonorthogonal and ambiguous kernel function, and time-consuming calculation of reproducibility. (ii) The standard test set is only the accumulation of several documents. Rather than the real database of the digital library, the experimental results are not convincing [16]. In order to solve these problems, firstly use semantic smoothing kernel and vocabulary expansion to preprocess the document vector, then use semantic sorting to increase the calculation speed, and finally use the wavelet kernel to classify documents in space. In order to verify the effect of the algorithm, we use the classified digital documents of the China Academic Journals Network for testing. Part of these classified documents is used to train algorithms and indexing rules, and the other part is used to test the learning effect of algorithms. The results show that compared with the traditional support vector machine method, the new method based on wavelet analysis has better results, and in the case of insufficient literature, it still has a strong learning function.

Association rule mining combines knowledge in many related fields such as computational linguistics [17], mathematical statistics [18], machine learning, and information retrieval, and then it effectively discovers the process of potentially useful information from massive data. The association rule mining process has gone through three stages from the initial database providing the original data to the user obtaining valuable knowledge from the database, namely, data preprocessing, association rule mining, and association rule evaluation, as shown in Figure 1.

The main research content of this paper has two points: a weighted multiple minimum support association rule mining algorithm based on low-frequency discretization. Since most data has numerical attributes, nowadays, data mining technology has developed, but many data mining methods still cannot handle numerical attributes. Therefore, in the field of data mining, the discretization of continuous attributes has gradually become a crucial issue. Important preparatory work and the loss of information in the discretization process have a direct impact on the quality of data mining. Although many existing continuous attribute discretization algorithms have their characteristics and advantages and disadvantages, most of them require users to define their assumptions. Therefore, these algorithms are more or less subject to various restrictions, and there are many data lost. Therefore, the first part of this article mainly studies the use of low-frequency discretization technology to discretize continuous attributes, thereby improving the accuracy of discretization and effectively reducing data loss. Because the importance of data in real life is not the same, to solve this problem and reduce the loss of information in the process of discretization of continuous attributes, this paper proposes multiple minimum support weighted association rule mining algorithm based on low-frequency discretization. The algorithm first uses the low-frequency discretization algorithm to discretize continuous attributes, then allows users to set their minimum support for data items in the database, and allows users to set corresponding weights for records with different importance, thereby effectively solving the problem of uneven distribution frequency of each data item when setting the unified support degree due to the different importance of each data item to the user, and through the full confidence, it removes the cross mode among them and then obtains clean frequent itemsets which are also easy to find more interesting rules.

This paper analyzes the existing continuous attribute discretization methods in detail, discusses how to reduce the loss of information in the discretization process, and proposes a novel low-frequency discretization algorithm (LFD) based on the attribute low-frequency region which provides promising results that improve the relevant algorithms' performance in terms of classification accuracy. The organization structure of this article is described in detail as follows. Section 1 is related to the research background. Section 2 briefly outlines how to conduct data mining and then introduces the association rule theory and Bayesian classification methods in detail. Section 3 is a new algorithm for the discretization of continuous attributes. Section 4 is the experimental results and analysis. Section 5 presents the summary and conclusion.

2. Research Background

2.1. Association Rule Theory. As an important data mining method, association rules represent rules that have a certain relationship between different items in the database. Since the association rules were proposed in 1993, people have

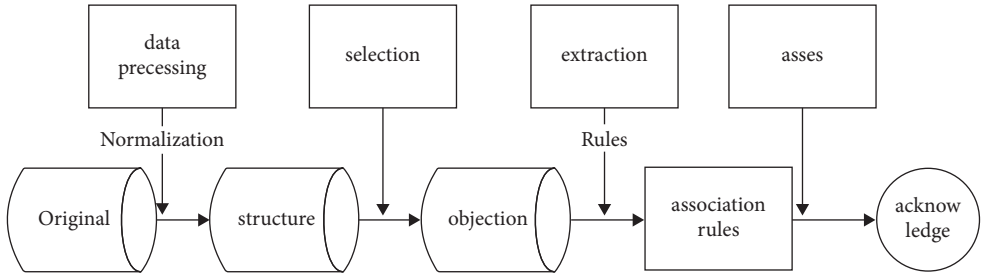

Figure 1: The processing of rules.

paid more and more attention to their mining and continue to expand the related fields [19]. When traditional algorithms mine frequent itemsets, all they get are hidden frequent itemsets. However, when the data set is dense, or the minimum support set by the user is too small, it will lead to the generation of a large number of frequent itemsets, which not only increases the difficulty of mining frequent itemsets, but also makes the mining results difficult to understand. At the same time, it also makes it more difficult to quickly find valuable information in the mining results. Therefore, people find ways to reduce the number of frequent itemsets generated in this situation, thereby reducing the redundant information of frequent patterns. From the current point of view, mining frequent closed itemsets and mining extremely frequent itemsets are both widely used [20]. The closed itemset in the data set means that the direct superset of the itemset does not have the same support count as it. Among them, similar to frequent itemsets, frequent closed itemsets refer to itemsets whose support is equal to or greater than the minimum support threshold in the closed itemsets [21]. The extremely frequent itemset refers to all itemsets in the frequent itemset that satisfy the direct superset and are not frequent.

In relatively dense data sets, long-frequent patterns often contain valuable information. Therefore, to dig out this potentially valuable information, it is usually possible to find the potentially extremely frequent itemsets in the data set. Among them, frequent itemsets without supersets are extremely frequent itemsets. Since the number of frequent closed itemsets is generally much greater than the number of maximum frequent itemsets, the maximum frequent itemsets can be used to effectively reduce the scale of the solution. This is helpful for users to quickly discover long-frequent patterns, and an effective understanding of long-frequent patterns has far-reaching significance.

2.2. Bayesian Methods. The Bayesian classification algorithm is a classification algorithm based on Bayes' theorem [22]. The algorithm uses posterior probability to represent the classification situation and predicts the class attribution of each target transaction through the obtained probability. An important concept in Bayesian classification is conditional probability. If both X and Y are random events, and the probability of Y always satisfies $P(Y) > 0$, then, under the premise that Y occurs, the probability of X occurring is

$$P(X|Y) = \frac{P(XY)}{P(Y)}. \tag{1}$$

The probability of event X and event Y occurring at the same time is called the joint probability of X and Y and is labeled $P(XY)$:

$$P(XY) = P(x|Y) * P(y). \tag{2}$$

A joint probability is still applicable in multiple events:

$$P(XY) = P(x|Y1) * P(y1) + P(x|Y2) * P(y2) + \cdots + P(x|Yn) * P(yn). \tag{3}$$

When it is known that event A and event B occur at the same time, the probability of event h is usually different from that of event A. However, there is a certain relationship between these two probabilities, and Bayes' theorem is used to describe this relationship. Bayes' theorem is a common method for calculating probability. It is generally believed that the probability of an event is directly related to whether the event occurs or not; that is, whether an event occurs is related to the probability of occurrence in the prior distribution. Suppose the probability of occurrence of event A is denoted as $P(A)$, we usually call it the prior probability; when we know that event A has occurred, the probability of occurrence of event B is denoted as P, which we call posterior probability.

$$P(b|a) = \frac{P(b)P(a|b)}{P(a)}. \tag{4}$$

The structure diagram of the Naive Bayes Classifier (NBC) is shown in the figure, where the root node represents the category and the leaf node represents the attribute. When an object is classified using the NBC model, first calculate the prior probability of each attribute, and then use the Bayesian formula to obtain the posterior probability of each attribute, and then you can determine which category the object belongs to, that is, the category with the largest posterior probability [23]. The structure of NBC is shown in Figure 2.

The naive Bayes classification model assumes that each attribute is independent of each other; that is, when there is a given category, it is assumed that the attributes are independent of each other so that the joint probability as shown in the formula can be obtained:

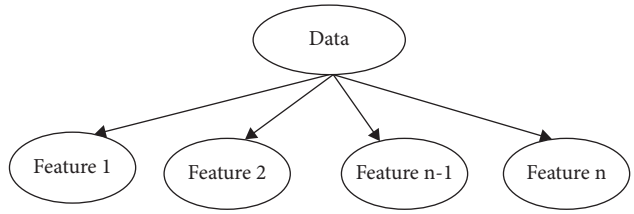

Figure 2: The structure of NBC.

$$
\begin{aligned}
P(C, A_1 = a_1, \ldots, A_n = a_n) &= P(C)\prod_{i=1}^{n} P(A_i = a_i|C),\\
P\left(X_1|Z_{1;t}\right) &= kP(Z_t|X_t)P\left(X_1|Z_{1,\Gamma-1}\right) = kp(Z_1|X_t)\int_{X\Gamma-1} P(X_t|X_{\Gamma-1})P\left(X_{\Gamma-1}|Z_{1;\Gamma-1}\right),\\
P\left(X_t|Z_{1,t}\right) &\approx kP(Z_t|X_t)\sum_i \omega_{r-1}^{(i)} P\left(X_t|X_{t-1}^{(i)}\right),\\
q(X_t) &= \sum_i \omega_{r-1}^{(i)} P\left(X_t|X_{\Gamma-1}^{(i)}\right).
\end{aligned}
\tag{5}
$$

Then, priority can be expressed as

$$\omega_t^{(i)} = P\left(Z_t|X_t^{(i)}\right). \tag{6}$$

2.3. Continuous and Discrete Attributes of Data. Discrete random variables and continuous random variables are determined by their value range or value form [24]. The value of the variable can only take the discrete natural number which is the discrete random variable; if the variable can take any real number in any certain interval, that is, the value of the variable can be continuous, such a random variable is called a continuous random variable. The field of data mining generally divides data into discrete data and continuous data. Among them, discrete data generally has an infinite number or a finite number of values, which can be described by language or represented by a small number of discrete values [25], for example, gender. Continuous data usually describes some measurable properties of the object and is generally represented by continuous intervals, for example, length. Generally speaking, in the database, there are not only continuous data but also discrete data.

The task of data discretization is to divide the specific value range of the entire value attribute into multiple independent subregions that do not cross each other by setting the split point. The obtained intervals are all discrete and use different values for these intervals or different symbols to represent [26]. Therefore, on wood, the discretization of continuous attributes is the process of dividing the continuous attribute space into multiple intervals with a certain dividing point because most of the data in the database under actual conditions are mixed data of continuous attributes and discrete attributes [27]. Moreover, under normal circumstances, the value range of continuous attributes is large, so there is no way to provide a unified discretization algorithm measurement index for all continuous attributes, but there are some principles that must be followed in the process of discretization of continuous attributes. Therefore, the discretization effect of the continuous attribute discretization algorithm can be simply and intuitively measured from the following aspects.

(i) Simplicity: If the spatial scale of the attributes is reduced by discretizing continuous attributes, that is, the number of division points and intervals is correspondingly reduced, then the discretization algorithm is relatively simple.

(ii) Consistency: Data sets with continuous attributes are usually consistent. If the discretization is successful, the generated discrete data set is also consistent with the original continuous attributes. Anyway, if the discretization is unsuccessful, a data set inconsistent with the original continuous attributes will be generated. At this time, information loss will occur. If the data set can be regarded as a simple information system, then the system should maintain consistency before and after discretization; that is, before and after discretization, the information system should have the same classification capabilities. The consistency level is usually used as a measure of consistency; that is, the discretization of continuous attributes should ensure that the data consistency level is not much different.

This section first introduces the process of data mining, then focuses on the related theories of association rules and classic association rule mining algorithms, and then

introduces the relevant theories of Bayesian classification and common Bayesian classification methods. In addition, this section also introduces the relevant knowledge of discretization of continuous attributes, including discretization tasks and evaluation criteria.

3. Weighted Multiple Minimum Support Association Rule Mining Algorithm Based on Low-Frequency Discretization

Association rule mining algorithm is an important branch of data mining, and the key to mining association rules lies in the mining of frequent itemsets [28]. In real life, people usually set different degrees of support for items with different occurrence probabilities according to their needs in order to obtain association rules that are useful to them. In practical applications, for example, stores may be more concerned about the sales of products with higher profits. Therefore, the products are weighted in a certain order so that people can obtain more valuable information [29]. However, in the related research of many scholars, few studies consider both the minimum support and attribute weight. Although many existing continuous attribute discretization algorithms have their characteristics, advantages, and disadvantages, most of them require users to define their assumptions or are based on various assumptions. Therefore, these algorithms are more or less affected by various assumptions, a kind of restriction. Moreover, the algorithm does not distinguish between high-frequency attributes and low-frequency attributes, resulting in a large amount of data loss. Therefore, this paper proposes a low-frequency discretization (LFD) algorithm. Based on the above methods, the WM-sampling algorithm can be obtained by applying the low-frequency discretization algorithm to the weighted multiple minimum support association rule mining algorithm. The algorithm uses a low-frequency discretization algorithm to discretize continuous attributes, sets the corresponding weight and minimum support for each data item, and uses full confidence to remove false frequent itemsets and strong association rules.

Through the research and analysis of the existing continuous attribute discretization methods, considering the relationship between frequency and data loss, this paper proposes discretization in the low-frequency data interval to reduce data loss. The "high-frequency" in the data value refers to a large number of values in the attribute data set. Correspondingly, the "low-frequency" in the value refers to a small number of values in the attribute data set. In this article, we show a discretization technique: low-frequency discretization (LFD); the algorithm only considers low-frequency values as possible boundary points, to minimize the impact of discretization. In other words, the purpose of the discretization algorithm is to improve the quality of discretization.

Rough set and its extended set attribute reduction algorithms both measure the changes in the relationship between conditional attributes and decision attributes before and after reduction by selecting appropriate feature evaluation functions and use this as an index to evaluate the ability of any candidate attribute set to approximate all attributes [30]. When the classification performance of the original data itself is poor, this evaluation index based on the ability to approximate all attributes will be difficult to guarantee or improve the classification performance of subsequent classifiers. Considering that, in practical applications, a large number of attribute reduction methods are ultimately evaluated by the subsequent classification performance, and the ultimate goal of all classification tasks is to find a sample that can distinguish different types of samples to the greatest extent (medium such as straight lines, planes, and hyperplanes). Therefore, this article makes full use of the distribution information of the data itself, directly from the perspective of improving classification performance, and defines the concepts of within-class distinguishability and inter-class distinguishability. The optimal reduction set is determined based on the reduction principle of minimizing the distinguishability within classes and maximizing the distinguishability between classes, to obtain the attribute subset that maximizes the separability of the data.

The inter-class distinguishability is to reflect the distinguishability between heterogeneous samples (samples with different labels) by calculating the distance between samples of different categories. The calculation idea is as follows. First, the domain of discourse is divided into A series of clusters; select two clusters from all clusters, select a sample from each cluster to combine, calculate the sum of the distances between all heterogeneous samples that conform to this two-tuple relationship, and divide by the distance the number of calculations to obtain the average value of the inter-class distances of all clusters. The greater the distance between classes, the greater the degree of distinguishability between classes.

$$\mathrm{Inter}D_b\left(d_i, d_j\right) = \frac{\sum_{i=1}^{|di|} S_b\left(d_{ik}, d_j\right)}{|di|}. \tag{7}$$

The within-class distinguishability is to reflect the distinguishability between samples of the same type (samples with the same label) by calculating the variance of samples of the same category. The calculation idea is as follows: first select a cluster from all clusters, calculate the corresponding sample of the cluster, and calculate the average within-class variance of all clusters in this way. The smaller the intra-class variance, the higher the degree of aggregation within each cluster, and the smaller the distinguishability within the class.

$$\mathrm{Var}S(B) = \frac{\sum_{b=1}^{|B|} \mathrm{var}(b)}{|B|}. \tag{8}$$

According to the pattern recognition theory, a good feature attribute set should have a smaller discriminability within a class and a larger discriminability between classes, to ensure the distinguishability of different categories. To this end, this paper defines an attribute importance measurement function based on distinguishability, which is used to evaluate the impact of each attribute on decision classification performance.

$$\mathrm{SIG}(a, B, D) = \frac{\mathrm{Inter}(B \cup a, D)}{\mathrm{Inter}(B, D)} - \frac{\mathrm{Var}S(B \cup a)}{\mathrm{Var}\,S(B)}. \quad (9)$$

Since the calculation of the distinguishability between classes and the distinguishability within classes depends on the label information of the sample, it is a supervision mechanism. Compared with the traditional unsupervised granulation method, it can avoid the introduction of inconsistent information in the reduction process and improve classification performance. Because the distinguishability function between classes satisfies monotonicity, but the distinguishability function within classes does not satisfy monotonicity, the attribute importance function obtained by the algebraic operation of the two is also nonmonotonic.

At present, the commonly used evaluation functions in rough set attribute reduction algorithms include dependency and information entropy, which all satisfy the principle of monotonicity, and Li pointed out that the attribute reduction algorithm based on this monotonic evaluation function has certain defects, such that when the classification performance of the original data set is poor, the corresponding monotonicity-based evaluation function metric value is also relatively low. Relevant studies have shown that the use of attribute reduction algorithms with nonmonotonic evaluation functions can achieve better classification performance. Therefore, the attribute reduction algorithm that satisfies nonmonotonicity proposed in this paper has a certain theoretical basis. To select the attribute set with the best classification performance, this paper draws on the algorithm design principle based on the nonmonotonic evaluation function in the paper, uses the distinguishability measurement function as the evaluation criterion of attribute importance, and uses the heuristic algorithm as the search strategy, a heuristic attribute reduction algorithm based on distinguishability (DISAR). The idea of the algorithm is as follows: first initialize the reduction set to an empty set, and for any attribute other than the reduction set, calculate the change in the distinguishability between and within the class after adding it to the reduction set, and maximize the importance of the attribute. The attributes are added to the reduction set, and the rules are executed in sequence until the algorithm terminates.

Since the attribute importance function based on distinguishability is nonmonotonic, the termination condition of the DISAR algorithm is as follows: SIG is less than 0. At this time, adding any attribute to the current reduction set cannot improve the reduction performance. Compared with the monotonic attribute reduction algorithm, the algorithm proposed in this paper does not need to select threshold parameters to control the degree of convergence of the algorithm. It only needs to terminate the algorithm when the algorithm is reduced to maximum attribute importance of less than or equal to 0, thereby avoiding the problem of the selection of threshold parameters. Considering that the termination conditions of the algorithm in practical applications are too strict, the selected reduced attribute set may have an overfitting problem. In order to solve this problem, this paper adopts a postpruning strategy. The idea is to first treat the reduction result calculated by the DISAR algorithm as a series of nested reduction attribute increase chains and then use classifiers such as SVM to increase the reduction chain. Perform one-by-one inspection, and finally use the reduction increasing chain corresponding to the highest accuracy as the final reduction set. Since this step does not affect the progress of the algorithm, there is no need to repeatedly debug the best classification, and it has a certain degree of objectivity, robustness, and operability. In terms of time complexity, since the distance measurement function in this article considers both Euclidean distance and Manhattan distance, the time complexity needs to be discussed separately. The calculation time of the algorithm proposed in this paper is mainly consumed in the calculation of sample distance. Suppose the number of samples in the universe U is n and the total number of conditional attributes is C. When the Euclidean distance is used, the sample needs to be recalculated every time an attribute is added. Distance and the algorithm need to traverse each attribute in the worst case, so the computational complexity of the inter-class discrimination is nlogn, and the computational complexity of the intra-class discrimination is c∗nlogn, so the overall time complexity is c∗nlogn; when the Manhattan distance is used, both the inter-class and intra-class discrimination only need to calculate the distance between samples under a single attribute once, so the overall time complexity is nlogn. It can be seen that, in terms of time efficiency, the algorithm proposed in this paper uses the Manhattan distance function to be far superior to the Euclidean distance function.

4. Design Analysis and Discussion

In order to test the performance of the DISAR algorithm, seven data sets are selected from the UCI machine learning database, as shown in Table 1 which describes the name of the datasets, the number of instances, number of features, and class.

All data sets are continuous data sets. In order to eliminate the influence of attribute dimensions, the maximum-minimum standardization method is used to normalize the data. The DISAR algorithm is compared with six representative attribute reduction algorithms in terms of the number of attributes and classification accuracy. The comparison algorithms are as follows:

(i) Attribute reduction algorithm based on neighborhood dependence (NRS)

(ii) Attribute reduction algorithm based on neighborhood variable precision rough set (NFARNRS)

(iii) Similarity-based attribute reduction algorithm (SIMR)

(iv) Fuzzy rough set attribute reduction algorithm based on Gaussian kernel approximation (FSGKA)

(v) Attribute reduction algorithm based on neighborhood combination measure (NCMAR)

(vi) Attribute reduction algorithm based on neighborhood combination entropy (ARNCE).

Among the 6 algorithms participating in the comparison, Algorithms 1, 2, 5, and 6 all need to set the

Table 1: Characteristics of the data sets used in the study.

No.	Name	Instances	Feature	Class
1	Iris	150	4	3
2	Diabetes	768	8	2
3	Wine	178	13	3
4	Heart	270	13	2
5	Glass	214	9	6
6	Sonar	208	60	2
7	dataR2	116	9	2

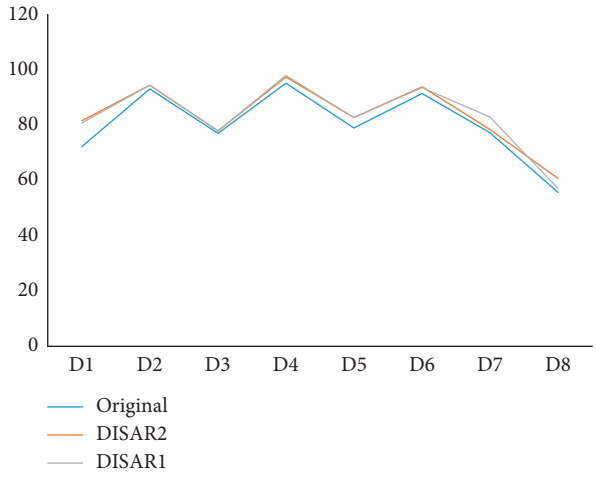

Figure 3: Support vector machine.

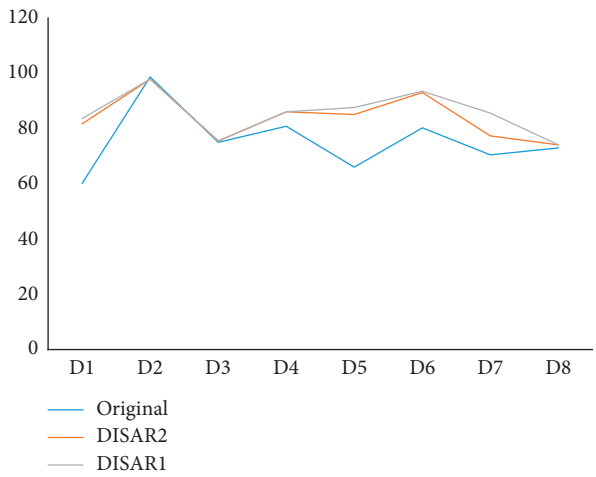

Figure 4: K-nearest neigbours (KNN).

neighborhood radius value. Here, the neighborhood radius is uniformly set to one-third of the standard deviation; the variable precision parameter in Algorithm 2 is taken. The value is set to 0.5–0.95, and the step size is 0.05; because Algorithm 3 can only process discrete data sets, the data set needs to be discretized before attribute reduction. The WEKA software was used in the original text but here the equal frequency discretization method is used. Through the three classifiers of SVM, CART, and KNN, the reduction set calculated by different algorithms is used to calculate the

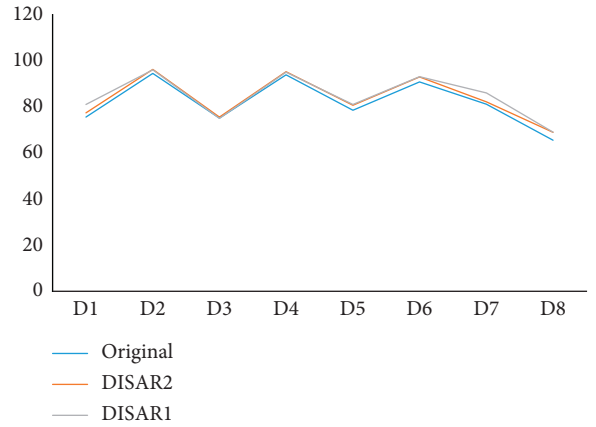

Figure 5: Classification and regression tree (CART).

accuracy of ten-fold cross-classification ten times, and the average value is taken as the final accuracy value. Since the reduction set when the classification accuracy of different classifiers is the highest is not necessarily the same, Algorithm 1 to Algorithm 6 select the reduction set with the highest classification accuracy as the final attribute reduction result. For sample data, too many features will not only take up a lot of storage space but also cause a serious burden on calculations. Therefore, in the actual reduction process, not only the classification accuracy of the reduction but also the scale of the resulting reduction must be considered.

In addition, in order to reveal the information changes of the attribute characteristics selected by the DISAR algorithm, the accuracy values of each data set on each attribute increase chain are recorded separately, and the corresponding line graph is drawn, which contains 9 data sets with larger feature dimensions. The abscissa corresponds to the number of attributes contained in the attribute increasing chain, the ordinate corresponds to the classification accuracy (%) under each classifier, SVM1/KNN1/CART1 corresponds to the accuracy of the DISAR1 algorithm on the SVM/KNN/CART classifier, and SVM2/KNN2/CART2 corresponds to the accuracy of the DISAR2 algorithm on the SVM/KNN/CART classifier, as shown in Figures 3–5.

As shown in Figures 3–5, for most data sets, in the process of increasing the number of attributes one by one, the classification accuracy of the attribute increase chain on different classifiers gradually increases. Oscillatory attenuation is mainly caused by the different data distributions and feature dimensions of each data set and the different preferences of different classifiers for the selection of reduction sets; at the same time, it can be observed that the attribute increase chain is classified on the three classifiers and the changing trend of accuracy is the same on the whole, which can also reflect the reliability of the classification ability of the classifier. W-T-L results are shown in Figure 6.

As shown in Table 2, what this article hopes is that the samples can be concentrated in the upper left corner of the bisecting line. At this time, the distance between classes is much larger than the distance within classes, and the classification performance is ideal. And from Figure 6, in these seven types of data sets, the distribution of each

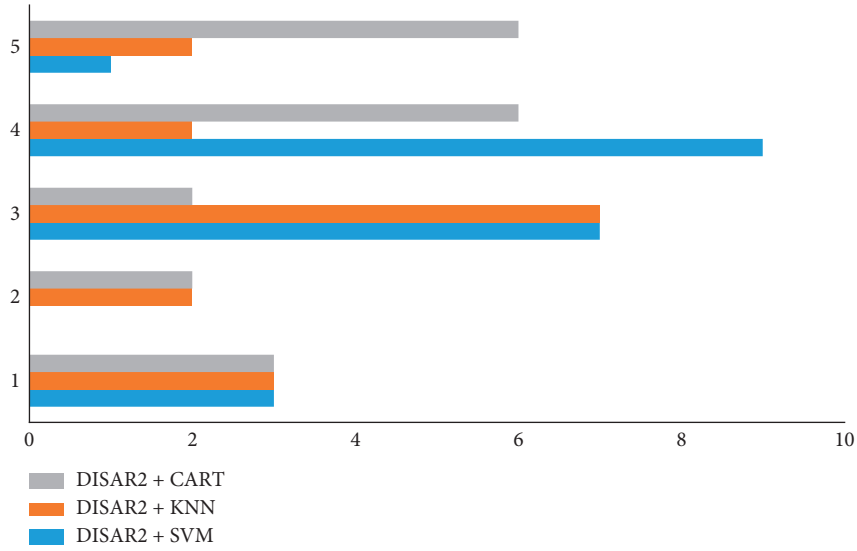

Figure 6: W-T-L results.

Table 2: W-T-L.

Data	DISAR2 + SVM	DISAR2 + KNN	DISAR2 + CART
1	3	3	3
2	2	2	2
3	7	7	2
4	9	2	6
5	1	2	6

reduced data has a certain degree of left shift relative to the distribution of the original data; especially in the iris data set, this change is more obvious so that the DISAR algorithm in this article can be selected. The advantage of easy-to-categorize attribute subsets can be demonstrated from a visual point of view.

5. Conclusion

This paper proposes a heuristic algorithm for attribute reduction based on a supervisory mechanism. It uses the distinguishability between classes as the evaluation criteria for attribute importance and quantitatively describes the changes in classification performance before and after the attribute is added during the reduction process; since there is no need to set any adjustable parameters in the algorithm, the resulting reduction is more objective and reliable. In the experimental simulation part, taking seven UCI data sets as the research object, through comparative analysis with 6 algorithms, it can be found that the average number of attributes of the DISAR2 and DISAR1 algorithms in this paper on the three classifiers is 4.70, which is much smaller than the others. The average number of attributes of the reduction set obtained by the six algorithms is 5.86, which significantly reduces the feature dimension of the data. At the same time, the average classification accuracy of the DISAR2 and DISAR1 algorithms is 85.36%, which is significantly higher than the average classification accuracy of the other six algorithms of 81.94%, which significantly improves the recognition accuracy of sample classification. It can be seen that the algorithm proposed in this paper has better classification performance. However, the algorithm in this paper is mainly researched on continuous data sets. How to introduce the algorithm in this paper to incomplete data sets and attribute reduction of mixed data sets is still a problem to be solved.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The author declares that he has no conflicts of interest.

References

[1] Z. a. Pawlak, "Rough sets," *International Journal of Computer & Information Sciences*, vol. 11, no. 5, pp. 341–356, 1982.

[2] X. Jia, Y. Rao, L. Shang, and T. Li, "Similarity-based attribute reduction in rough set theory: A clustering perspective," *International Journal of Machine Learning and Cybernetics*, vol. 11, no. 5, pp. 1047–1060, 2020.

[3] S.-H. Teng, M. Lu, A.-F. Yang, J. Zhang, Y. Nian, and M. He, "Efficient attribute reduction from the viewpoint of discernibility," *Information Sciences*, vol. 326, no. 1, pp. 297–314, 2016.

[4] Q. Hu, D. Yu, J. Liu, and C. Wu, "Neighborhood rough set based heterogeneous feature subset selection," *Information Sciences*, vol. 178, no. 18, pp. 3577–3594, 2008.

[5] Y. Zhou, R. Zhao, Q. Luo, and C. Wen, "Sensor deployment scheme based on social spider optimization algorithm for wireless sensor networks," *Neural Processing Letters*, vol. 48, no. 1, pp. 71–94, 2018.

[6] Z. Zhang and Y. Zhang, "Application of wireless sensor network in dynamic linkage video surveillance system based on kalman filtering algorithm," *The Journal of Supercomputing*, vol. 75, no. 9, pp. 6055–6069, 2019.

[7] X. Yu, H. Lu, X. Yang, Y. Chen, and W. Shi, "An adaptive method based on contextual anomaly detection in internet of things through wireless sensor networks," *International Journal of Distributed Sensor Networks*, vol. 16, no. 5, p. 155, 2020.

[8] Y. Yao and B. Yao, "Covering based rough set approximations," *Information Sciences*, vol. 200, no. 1, pp. 91–107, 2012.

[9] Z. Ying, J. Wang, and G. Hao, "An autonomous connectivity restoration algorithm based on finite state machine for wireless sensor-actor networks," *Sensors*, vol. 18, no. 2, p. 153, 2018.

[10] Y. Yang, S. Wang, W. Xu, and K. Wei, "Reliability evaluation of wireless multimedia sensor networks based on instantaneous availability," *International Journal of Distributed Sensor Networks*, vol. 14, no. 11, p. 155, 2018.

[11] B. H. Xiao, B. Q. Pei, and N. Wu, "Design of manned centrifuge control system based on motion controller," *Space Medicine & Medical Engineering*, vol. 031, no. 001, pp. 32–36, 2018.

[12] W. Wang and M. Zhang, "Self-adaptive gathering for energy-efficient data stream in heterogeneous wireless sensor networks based on deep learning," *IEEE Wireless Communications*, vol. 27, no. 5, pp. 74–79, 2020.

[13] T. Wang, H. Xiong, H. Ding, and L. Zheng, "Tdoa-based joint synchronization and localization algorithm for asynchronous wireless sensor networks," *IEEE Transactions on Communications*, vol. 68, no. 5, pp. 3107–3124, 2020.

[14] J. Wang, R. Zhu, S. Liu, and Z. Cai, "Node location privacy protection based on differentially private grids in industrial wireless sensor networks," *Sensors*, vol. 18, no. 2, p. 410, 2018.

[15] R. Tian, L. Wang, J. Zhang, H. Yang, and L. Zhang, "Wireless energy-efficient system based on the directed diffusion routing method for the high density seismic array survey," *IOP Conference Series: Earth and Environmental Science*, vol. 660, no. 1, p. 7pp, Article ID 012140, 2021.

[16] N. N. Tang, X. F. Zheng, J. Z. Wu, H. Chen, and H. X. Li, "Design of off-line led control system based on 4g-lte," *Chinese Journal of Liquid Crystals and Displays*, vol. 33, no. 01, pp. 55–60, 2018.

[17] S. Sivakumar and P. Vivekanandan, "Efficient fault-tolerant routing in IoT wireless sensor networks based on path graph flow modeling with Marchenko-Pastur distribution (EFT-PMD)," *Wireless Networks*, vol. 26, no. 6, pp. 4543–4555, 2020.

[18] R. Sharma, V. Vashisht, and U. Singh, "WOATCA: A secure and energy aware scheme based on whale optimisation in clustered wireless sensor networks," *IET Communications*, vol. 14, no. 8, pp. 1199–1208, 2020.

[19] S. Yang, Z. Gong, K. Ye, Y. Wei, Z. Huang, and Z. Huang, "EdgeRNN: A compact speech recognition network with spatio-temporal features for edge computing," *IEEE Access*, vol. 8, pp. 81468–81478, 2020.

[20] Y. Qi, C. Ma, H. Yu, and X. Bian, "A key pre-distribution scheme based on μ-pbibd for enhancing resilience in wireless sensor networks," *Sensors*, vol. 18, no. 5, p. 1539, 2018.

[21] M. S. Hossain and G. Muhammad, "An audio-visual emotion recognition system using deep learning fusion for a cognitive wireless framework," *IEEE Wireless Communications*, vol. 26, no. 3, pp. 62–68, June 2019.

[22] Q. Liu, "Coverage reliability evaluation of wireless sensor network considering common cause failures based on d–s evidence theory," *IEEE Transactions on Reliability*, vol. 23, no. 99, pp. 1–15, 2020.

[23] L. Liu, G. Han, Z. Xu, J. Jiang, and M. Martinez-Garcia, "Boundary tracking of continuous objects based on binary tree structured svm for industrial wireless sensor networks," *IEEE Transactions on Mobile Computing*, vol. 41, no. 99, p. 1, 2020.

[24] Q. Hu, L. Zhang, D. Chen, W. Pedrycz, and D. Yu, "Gaussian kernel based fuzzy rough sets: Model, uncertainty measures and applications," *International Journal of Approximate Reasoning*, vol. 51, no. 4, pp. 453–471, 2010.

[25] K. Khoshraftar and B. Heidari, "A hybrid method based on clustering to improve the reliability of the wireless sensor networks," *Wireless Personal Communications*, vol. 113, no. 2, pp. 1029–1049, 2020.

[26] J. Zhang and D. Tao, "Empowering things with intelligence: A survey of the progress, challenges, and opportunities in artificial intelligence of things," *IEEE Internet of Things Journal*, vol. 8, no. 10, pp. 7789–7817, 2021.

[27] B. J. Fan, Y. R. Li, and Y. Z. Liu, "The study on the controller of aviation switched reluctance starter/generator integrated system," *Computer Simulation*, vol. 36, no. 01, pp. 102–107, 2019.

[28] A. Facchinetti, L. Gasparetto, and S. Bruni, "Real-time catenary models for the hardware-in-the-loop simulation of the pantograph-catenary interaction," *Vehicle System Dynamics*, vol. 51, no. 4, pp. 499–516, Jan. 2013.

[29] Y. Cui, L. Zhang, Y. Hou, and G. Tian, "Design of intelligent home pension service platform based on machine learning and wireless sensor network," *Journal of Intelligent and Fuzzy Systems*, vol. 40, no. 2, pp. 2529–2540, 2021.

[30] S. M. Chowdhury and A. Hossain, "Impact of error control code on characteristic distance in wireless underground sensor networks," *IET Communications*, vol. 12, no. 13, pp. 1540–1549, 2018.