Hindawi
Wireless Communications and Mobile Computing
Volume 2022, Article ID 5659979, 12 pages
https://doi.org/10.1155/2022/5659979

Research Article

Integrated Classification Algorithm for Unbalanced Data Streams Based on Joint Nonnegative Matrix Factorization

Jin Li and Ruibo Zhao

Tencent Technology Company Limited, Beijing 100086, China

Correspondence should be addressed to Ruibo Zhao; zhaor3@cuc.edu.cn

Received 14 February 2022; Revised 15 March 2022; Accepted 13 April 2022; Published 13 June 2022

Academic Editor: Lisheng Fan

The purpose of this paper is to study the unbalanced data flow integration classification algorithm based on joint nonnegative matrix factorization, in order to solve the problem that the basic clustering results obtained from the original data set have some information loss, thereby reducing the effective information in the integration stage. In this paper, the accuracy of the unbalanced data and the detection time consumption are selected as the research object. Six data sets with imbalanced proportions of minority and majority samples are selected for experiments. Mathematical statistical analysis is first used to observe text classification, disease diagnosis, and network intrusion detection and the classification accuracy of majority class and minority class; the commonly used algorithm for unbalanced data is statistical analysis method. Comparing the univariate method for comprehensive classification of unbalanced data flow based on nonnegative matrix factorization with the unbalanced data algorithm, the observation has accurate rate and detects time-consuming changes. Among them, the comprehensive classification algorithm of unbalanced data flow is based on the classification of data, classifying the data, judging whether two data points belong to the same category, and determining their degree of balance. The research data shows that the unbalanced data flow integrated classification algorithm based on joint nonnegative matrix decomposition can reasonably evaluate the classification performance of the classifier for a few classes, and the detection speed is faster and saves more time. The experimental research shows that the algorithm combines the relationship matrix and information matrix from the original data set into a consensus function, uses NMF technology to obtain the membership matrix, effectively uses potential information, improves the accuracy rate of 69.73%, and shortens 71.65% of the time consumed.

1. Introduction

The number is huge, and the dynamically changing incoming data is called the data stream. The classification of data streams is widely used for e-commerce and real-time monitoring of sensor networks and networks. However, the distribution by class in these applications is often uneven. This kind of data flow characterized by unbalanced distribution is called unbalanced. The data related to the unbalanced distribution of these categories gives traditional data extraction and classification algorithms and even poses serious problems for the existing data flow classification. Unbalanced mixed data processing is an important application in real life, especially in medical treatment, transportation, fault handling, and so on. Therefore, using various classification algorithms to process unbalanced mixed data has become an important research content in data mining.

With the rapid development of information technology, these data contain a lot of information in a series of application fields (such as wireless sensor networks, real-time traffic systems, network traffic monitoring, and credit card fraud detection), which prompt us to mine urgently. At present, there are few classification algorithms that can process unbalanced mixed data at the same time. This paper explores the comprehensive classification algorithm of unbalanced data flow based on joint nonnegative matrix factorization, in order to provide a significant contribution to data mining research.

Lu and Miao's decomposition of data into a small number of basic components is usually an effective strategy for data exploration, analysis, and interpretation [1]. Various working methods, such as principal component analysis (PCA) and nonnegative matrix factorization (NMF), are developed along this line of thought. These methods impose different constraints (e.g., orthogonality of PCA) to obtain compact or physically meaningful basis. Ying-Ying et al. discuss the molecular typing and prognosis prediction of gastric cancer based on nonnegative matrix factorization (NMF) [2]. The gene expression spectrum (GEO) was detected in patients with gastric cancer. The expression profile of INc RNA was analyzed using INC RNA mining method. The NMF model was established using consistent clustering +software package.

In order to improve the performance of traditional data stream integration algorithms in big data mining applications, a parallel data level integration algorithm was designed and implemented with the help of cloud computing-related technologies and nonnegative matrix factorization methods [3]. The purpose of this paper is to study the unbalanced data flow integration classification algorithm based on joint nonnegative matrix factorization, in order to solve the problem that the basic clustering results obtained from the original data set have some information loss, thereby reducing the effective information in the integration stage.

2. Programs Method

2.1. Basic Content of Nonnegative Matrix Factorization. Nonnegative matrix factorization makes all components after factorization nonnegative (requiring a purely additive description) and at the same time achieves nonlinear dimensionality reduction. This nonnegativity restriction leads to a certain degree of sparsity in the corresponding descriptions, and sparsity representations have been shown to be an efficient form of data descriptions between fully distributed descriptions and those of a single active component. Nonnegative matrix factorization (NMF) method has been widely used in multidimensional data similarity data clustering, text clustering, and social network clustering, but its serial calculation is the most difficult. The time is for big data processing operations [4, 5]. Previously, in the field of parallel data cluster processing for multidimensional data parallelization, there were cluster computer and shared memory computing methods, as well as grid computing, peer-to-peer computing, and widely distributed computing model spectra, all with excellent results. However, in the era of cloud computing, predistributed distributed computing models used for large amounts of PB often appear to be insufficient, so proper attention should be paid to cloud-based data classification groups. Optimization of traditional data aggregation methods is based on nonnegative matrix factorization. And NMF has gradually become one of the most popular multidimensional data processing tools in research fields such as signal processing, biomedical engineering, pattern recognition, computer vision, and image engineering.

(1) Unbalanced data stream integrated classification algorithm and nonnegative matrix factorization

The so-called unbalanced data refers to the fact that there are more samples of some classes than other classes in a data set. The class with more samples is generally called the majority class, and the class with fewer samples is called the minority class. The unbalanced data flow integrated classification algorithm is based on the classification of the data to classify the data, whether the two data points belong to the same category, and determine how balanced they are [6–8]. When the balance between them is greater than a certain value, they belong to the same cluster; otherwise, the two data points belong to different clusters. However, there are still some difficulties. Due to the serious skew in quantity, the performance of classification algorithms for classifying unbalanced data sets is not satisfactory. Because minority class samples are usually more difficult to identify than common samples, most data mining classification algorithms have great difficulty in dealing with minority class samples.

There are large-scale data in the presence of practical problems, which makes the matrix that stores these large data very large, and the stored information is unevenly distributed, which means that existing methods cannot process the data contained in the matrix efficiently and quickly [9]. In order to better deal with such data, the effective method category is the decomposition of the matrix, which can greatly reduce the size of the description problem and can compress and summarize the data. To this end, there are many methods of factorization matrix, such as the decomposition of exogenous values, the analysis of independent components, and the analysis of principal components [10]. Matrix factorization is a method of reducing a matrix to its constituent parts. This approach simplifies more complex matrix operations that can be performed on the decomposed matrix rather than the original matrix itself. The decomposition results obtained from cluster analysis are based on the decomposition of nonnegative matrices, which can ensure that its elements are not negative and represent their actual physical meaning. Therefore, in recent years, they have become the object of special attention.

The clustering method based on nonnegative matrix factorization NMF is as follows: considering that the data set can be represented as a vector set $X = \{X_1, X_2, \cdots, X_n\}$, and each vector represents the m-dimensional data point $X_j = \{x_{j1}, x_{j2}, \cdots, x_{\mathrm{jm}}\}$, NMF method is to divide X into two nonnegative low-rank matrices W and H, which can be achieved by optimizing the following formula as much as possible:

$$\min_{x>0} \left\| G - \hat{X}\hat{X}^T \right\|_F^2, \tag{1}$$

where $X^{\wedge}$ can be obtained by the following multiplication update rule:

$$\hat{X}_{ik} \longleftarrow \hat{X}_{ik}\left[\frac{1}{2} + \frac{\left(G\hat{X}\right)_{ik}}{\left(2\hat{X}\hat{X}^{T}\hat{X}\right)_{ik}}\right]. \tag{2}$$

(2) After decomposition, NMF can retain more information reflected by the original sample. The result obtained after decomposition is nonnegative and has good physical meaning, and the implementation process is simple and fast. As the name implies, NMF decomposes a nonnegative matrix into two nonnegative matrices, and the result of multiplying these two matrices is equal to the original matrix before decomposition [11]. The objective function is shown in

$$\min \|X - WH\|_F^2 \tag{3}$$

Among them, the nonnegative data set $X \in R^{m*n}n$ is the original matrix, $X_{m*n} = (x_1, x_2, \cdots, x_n)$, and x_i represents an m -dimensional column vector; that is, the information of a sample [12]. The basis matrix $W \in R^{m*r}$, $W_{m*r} = (w_1, w_2, \cdots, w_r)$, and w_i represents an m-dimensional column vector, representing a basis vector. The coefficient matrix $H \in R^{r*n}$; $H_{r*n} = (h_1, h_2, \cdots, h_n)$, where h_i is the column vector of r dimension, which can be regarded as the coordinates of the projection of the x_i vector in the new space defined by the W-based matrix, satisfying $x_i = W * h_i$, where x_i is the projection coefficient. r satisfies the condition $(m + n) * r < m * n$, that is to decompose a high-dimensional nonnegative matrix into the product of two low-rank nonnegative matrices. The iteration rules are shown in equations (4) and (5); Θ represents the Hada code program.

$$W \longleftarrow W\Theta\frac{\left(XH^{T}\right)_{ij}}{\left(WHH^{T}\right)_{ij}}, \tag{4}$$

$$H \longleftarrow H\Theta\frac{\left(W^{T}X\right)_{ij}}{\left(W^{T}WX\right)_{ij}}. \tag{5}$$

In the NMF iteration process, the base matrix has no constraints, and there is a lot of redundancy between the data [13]. Therefore, in recent years, many improved algorithms for NMF have been proposed.

(3) Joint nonnegative matrix initialization method

As with other models based on iterative optimization, since the local minimum is not unique, the result of nonnegative matrix decomposition is usually more sensitive to the initial value of the factor matrix, which means that the initialization of W and H will affect the convergence speed and final result of the algorithm [14–16]. For the sake of simplicity, random initialization is often used to assign initial values to W and H in many studies, which often makes the algorithm's convergence rate slower. To this end, some researchers have proposed some other methods to initialize NMF. Common initialization methods include the following categories:

(1) Multiple initialization

The core idea of this type of method is to perform multiple random initializations on the factor matrix, run the NMF algorithm once for each initialization, and then select the best estimate as the final decomposition result [17]. Due to the need to perform NMF decomposition multiple times, the computational overhead of such algorithms is often relatively large.

(2) Initialization based on matrix factorization

Nonnegative matrix factorization is actually a low-rank factorization technique with constraints, so we can use the results of other low-rank factorization algorithms as NMF initialization. Typical examples include SVD-based initialization and CUR decomposition-based initialization.

(3) Cluster-based initialization

Based on the characteristics of nonnegative matrix factorization, we can regard nonnegative matrix factorization as a clustering process, so the results of other clustering algorithms, such as k-means and fuzzy clustering, are used as NMF initialization. Compared with the initialization method based on matrix decomposition, using this kind of method as the preprocessing process is often too complicated and may cause the algorithm to terminate at a poor local solution. In practical applications, the selection of NMF initialization methods cannot be generalized, and the initialization method that is effective for one data set may not be applicable to another data set. Therefore, it is often necessary to select a suitable initialization method based on practical problems and certain prior knowledge.

(4) Common nonnegative matrix factorization constraints

(a) Sparseness constraint

The sparsity constraint helps to improve the uniqueness of the calculation results of nonnegative matrix factorization, and at the same time, it helps to strengthen the characteristics based on the partial representation. If W is regarded as a base matrix and H is regarded as a coefficient matrix, then applying a sparsity constraint to each column of W will make each base vector only affect a small part of the original observations: column sparsity constraints; then, each observation will only be represented by a linear combination of a few base vectors; and if sparsity constraints are imposed on each row of H, then each base vector will only be used to approximate some of the training data, or it can be understood that each basis vector is derived from part of the training data, which has a strong correlation with clustering.

(b) Orthogonality constraint

The addition of orthogonality constraints in nonnegative matrix factorization is to minimize the redundancy between basis vectors [18]. If the orthogonality constraint is applied to each column of W, that is, $W^TW = I$, then it will make the basis vectors have the greatest discrimination; and the orthogonality constraint is applied to each row of H, that is, $VV^T = I$, which will improve the accuracy of clustering. It is worth noting that applying orthogonality constraints to W and H, respectively, is actually equivalent to clustering the rows and columns of the input data matrix, respectively. If one factor matrix in NMF is regarded as a clustering center, the other is equivalent to an indicator vector. Therefore, orthogonality constraints have also been applied in clustering research.

(c) Discriminant constraints

From the perspective of pattern recognition, the traditional NMF algorithm can be regarded as an unsupervised learning process. By combining discriminative information and decomposition process, the basic NMF algorithm can be extended to supervised learning, and the model generation and classification tasks can be integrated into a framework. This method has been successfully applied to classification applications such as face recognition and expression recognition.

2.2. Components of an Integrated Classification Algorithm for Unbalanced Data Streams

2.2.1. Characteristics of Unbalanced Data. The data imbalance problem mainly exists in supervised machine learning tasks. When encountering unbalanced data, traditional classification algorithms with overall classification accuracy as the learning objective will pay too much attention to the majority class, thus degrading the classification performance of minority class samples. Unbalanced data is mainly composed of two types of interclass imbalance and intraclass imbalance [19, 20]. The imbalance between classes leads to uneven data distribution between classes, as shown in Figures 1(a) and 1(b). In some practical applications, the data shows that the data between the classes is extremely unbalanced, and the unbalance rate can reach 1000: 1 or greater in some cases. The imbalance in a category refers to the imbalance in the sample size of a category and its subcategories, or the data of a category has multiple different terms that are not so important, as shown in Figures 1(c) and 1(d). A large number of studies have shown that the imbalance of data between categories is not the only factor affecting classification learning, and the imbalance of data within categories is a key factor affecting the effect of classification [21]. Therefore, the classification problem of unbalanced data is mainly due to the complexity of the data distribution, as shown in Figures 1(b) and 1(c) (data overlap) and Figure 1(d) (small fragmentation problem); all of these problems will directly affect the classifier's learning result. Unbalanced data scenarios appear in all aspects of Internet applications, such as click prediction of search engines (clicked web pages often occupy a small proportion), product recommendation in the field of e-commerce (the proportion of recommended products being purchased is very low), credit card fraud detection, network attack identification, and cancer detection.

2.2.2. Integrated Classification Technology for Unbalanced Data. The integrated classification algorithm was aimed at improving the accuracy of the overall learning and cannot be directly used to deal with the classification learning of unbalanced data. Based on the currently available results, the imbalanced data can be classified through integrated algorithms at two levels: algorithm or data [22, 23]. Algorithm processing includes introducing a cost factor in the formation process of the comprehensive classification algorithm. According to whether the cost of heterogeneous samples is different from the cost of incorrect classification, the different cost factors are attributed to the cost factor to form an integrated cost-sensitive type classification algorithm. Since the AdaBoost algorithm is a series of trainings for different basic classifiers, they are obtained by changing the weight values of the training samples, so an integrated cost-sensitive classification algorithm is usually introduced to form a cost factor in the update. The legal values of the training champion. Data processing refers to the technique of rebalancing the data sampling during the establishment of the basic classifier, so that the integrated algorithm can build the classifier on the balanced training data that does not affect the learning performance. The combination of different data balancing strategies for resampling and integrated classification algorithms has led to integrated classification algorithms based on data processing boosting, integrated classification based on data processing bagging, and integrated classification based on mixed data processing.

2.2.3. Unbalanced Data Classification Algorithm Evaluation System. The classification of the decomposed data requires the classifier to achieve a high degree of classification accuracy for a limited number of classified samples without harming most of the classified samples. The evaluation criteria commonly used in learning machines are indicators of global classification accuracy and are not suitable for classification algorithms that evaluate classified data [24, 25]. Evaluation criteria that can provide more information are usually adopted in existing studies, such as single evaluation indicators based on confusion matrix, precision curve before recall, ROC curve, and cost curve.

2.2.4. Unbalanced Data Processing. Maintain the function of the original sample distribution; on the other hand, in order to make better use of most of the information in the sample (useful for the computer imbalance phase), first provide a random sample of the cyclic feature subset and then a small percentage relative to the number of these samples. The number of samples in each weight category is calculated comprehensively with the majority shared sample type. The base and composite weights for each sample category plus the number of sample types make up the percentage of sample formation. At last, the processed training sample

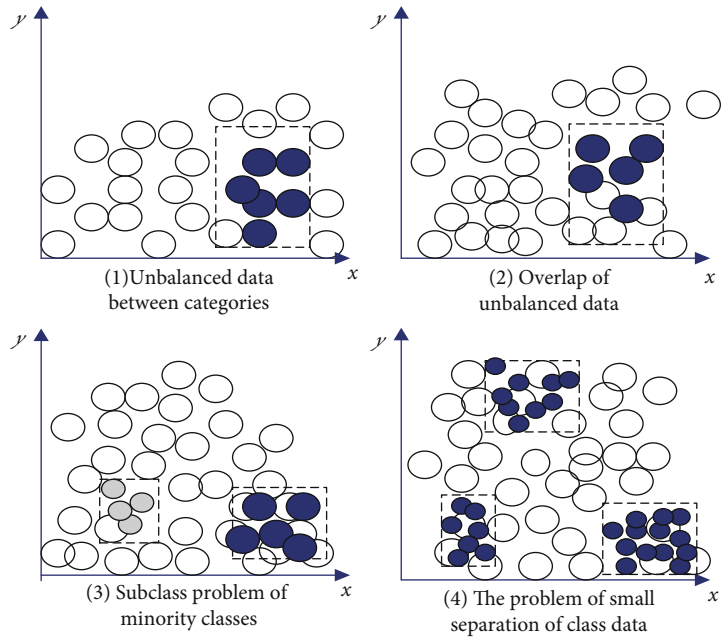

Figure 1: Characteristics of unbalanced data.

subset and feature subset are finally obtained. Among them, for the unbalanced data sampling method, the basic idea is to eliminate or reduce the imbalance of the data by changing the distribution of the training data. The specific process is shown in Figure 2. The way to deal with data imbalance is as follows: in the case of very few positive and negative samples, data synthesis should be used. In the case where there are enough negative samples and very few positive samples and the proportion is very disparate, the classification method should be considered. In the case where there are enough positive and negative samples and the proportions are not particularly disparate, sampling or weighting methods should be considered.

3. The Experiments

3.1. Experimental Data Set. This experiment selects 10 data from two sources: artificial data set and UCIE data set, of which 2d4c is a randomly generated artificial data set based on Gaussian distribution, and the rest are all from UCI's real data set, among which balance, heart, liver. They are the abbreviations of data set balance-scal, heart-statlog, liver disorders, and contraceptivc-method-choice. The relevant statistical information of all test data sets is listed in Table 1.

3.2. Experimental Design. Set the number of runs of the imbalanced data flow integrated classification algorithm $M = 10$, and combine the different result sets into an information matrix for experiments. Set the relationship matrix weight parameter δ to 0.0001, which is empirically obtained.

Compare the algorithm in this paper with the traditional algorithm, observe the use of the integrated nonnegative matrix decomposition-based unbalanced data stream integrated classification algorithm and common algorithm, and observe the classification of text classification, disease diagnosis, network intrusion detection, and the majority and minority categories rate. At the same time, compare the classification accuracy of the two algorithms in various scenes and the speed of the time consumption.

3.3. Evaluation Criteria. This experiment will use $F1$ and RI (Rand index) to evaluate the experimental results.

The definition of $F1$ is as follows:

$$\mathrm{F1} = \frac{2 * \mathrm{PR}}{\mathrm{P} + \mathrm{R}}, \tag{6}$$

where P is the precision rate, which represents the proportion of extracted correct objects in the extracted objects, and R is the recall rate, which represents the proportion of extracted correct objects in the samples.

The definition of RI is as follows:

$$\mathrm{RI}\left(\prod, \pi\right) = \frac{n_{11} + n_{00}}{n_{11} + n_{10} + n_{01} + n_{00}} \tag{7}$$

where II is the real data set, π is the clustering result label, n_{11} represents the number of data objects in a cluster in both the II and π sets, n_{01} represents the number of different clusters in the π set that are in the same cluster but in II, and the meanings of n_{00} and n_{10} are the same.

According to the above definition, the larger the values of $F1$ and RI, the better the clustering effect.

4. Discussion

4.1. Effectiveness of Using Integrated Nonmatrix Decomposition-Based Unbalanced Data Flow Integrated Classification Algorithm

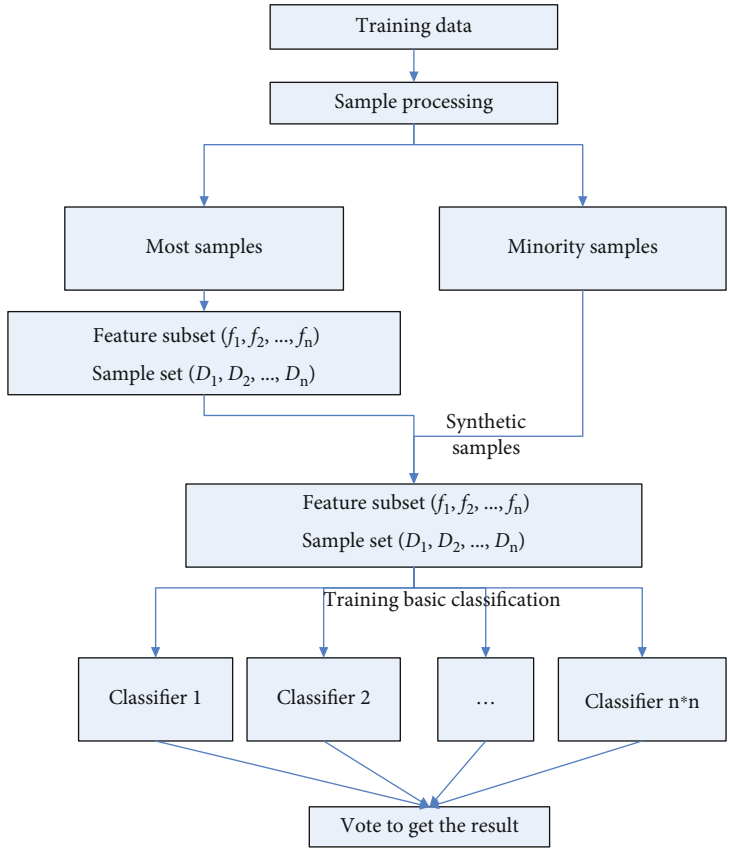

Figure 2: Method flow.

Table 1: Relevant information description of experimental test data set.

Data set	Number of samples	Attributes	Number of categories	Source
2d4c	200	2	4	Artificial
Wine	190	13	3	UCI
Iris	150	4	3	UCI
Glass	239	9	6	UCI
Segment	2319	19	7	UCI
Balance	572	2	3	UCI
Diabetes	721	9	2	UCI
Heart	281	13	3	UCI
Liver	273	6	2	UCI
Emc	1252	9	3	UCI

(1) In this experiment, the unbalanced data stream integrated classification algorithm based on joint nonnegative matrix decomposition is used to observe the classification accuracy of the majority and minority categories of text classification, disease diagnosis, and network intrusion detection. The data shows that after five tests, the text classification, disease diagnosis classification, and network intrusion detection classification have obtained obvious correct rates. The text classification accuracy rate is 78.45% on average, the disease diagnosis classification is 65.72% on average, the average detection classification of network intrusion is 83.23%. Based on the comparison of classification results based on joint nonnegative matrix factorization, only when the recall and precision rates are large, the nonnegative matrix factorization will be correspondingly large. Therefore, nonnegative matrix factorization can reasonably evaluate the classification performance of the classifier for minority classes. The data collection table is shown in Table 2 and Figure 3

Table 2: Effects of using nonnegative matrix factorization (unit: %).

Test	Text categorization	Disease diagnosis	Network intrusion detection
Test 1	78.34	63.34	80.34
Test 2	79.84	65.24	82.38
Test 3	77.65	64.92	84.23
Test 4	78.69	65.86	82.47
Test 5	78.82	64.38	83.23

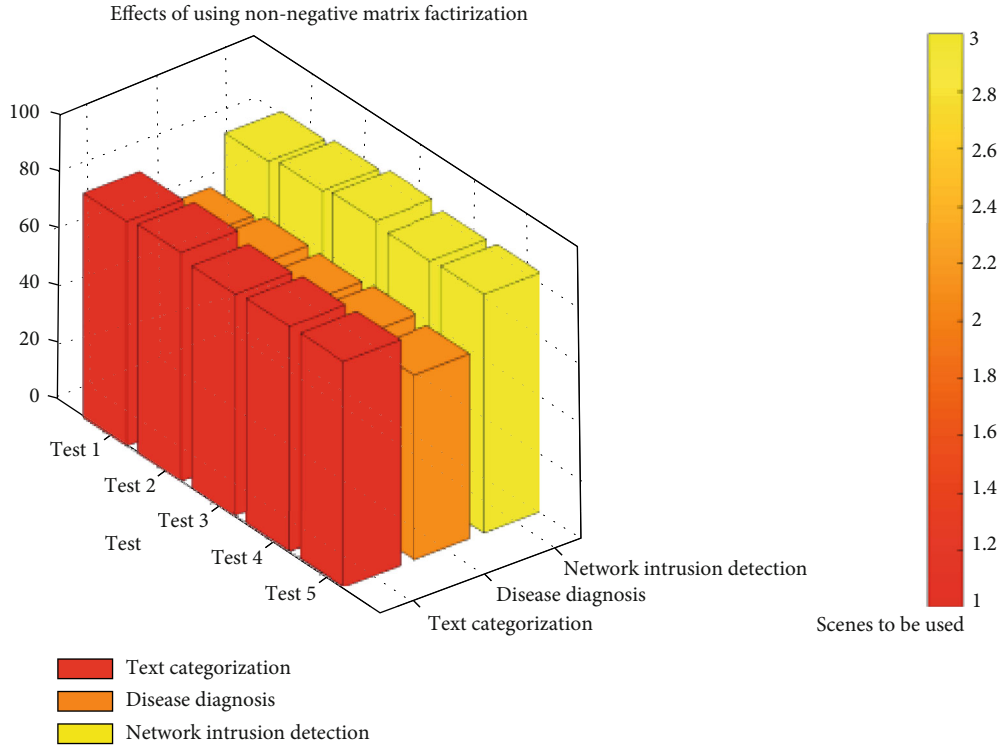

Figure 3: Effects of using nonnegative matrix factorization (unit: %).

(2) After using the general unbalanced data flow integration classification algorithm in this experiment, observe the classification accuracy of the majority and minority categories of text classification, disease diagnosis, and network intrusion detection. The data shows that after 5 tests, the text classification, disease diagnosis classification, and network intrusion detection classification can only get a lower correct rate. The average accuracy rate of the text classification is 22.45%, and the disease diagnosis classification is 24.64%. The average network intrusion detection classification is 19.54%. Using ordinary unbalanced data flow integrated classification algorithms, the recall and precision are small, and it is difficult to improve the accuracy of minority and majority classification. Therefore, the general unbalanced data flow integrated classification algorithm is not suitable for the classification of unbalanced data. The data collection table is shown in Table 3 and Figure 4

4.2. Convenience of Using Integrated Nonmatrix Decomposition-Based Unbalanced Data Flow Integration Classification Algorithm

(1) In this experiment, the unbalanced data flow integrated classification algorithm based on joint nonnegative matrix decomposition and the general unbalanced data flow integrated classification algorithm are used to classify the majority and minority categories in text classification, disease diagnosis, and network intrusion detection. The data shows that after five tests, the text classification based on the unbalanced data flow integrated classification algorithm under the joint nonnegative matrix decomposition, the disease diagnosis classification, and the network intrusion detection classification has obtained obvious correct rates; and based on the general classification, the accuracy of the

Table 3: Classification accuracy of common algorithms (unit: %).

Test	Text categorization	Disease diagnosis	Network intrusion detection
Test 1	20.12	25.45	19.45
Test 2	22.33	26.05	18.64
Test 3	23.74	25.84	19.56
Test 4	23.58	25.38	20.18
Test 5	21.83	23.12	20.23

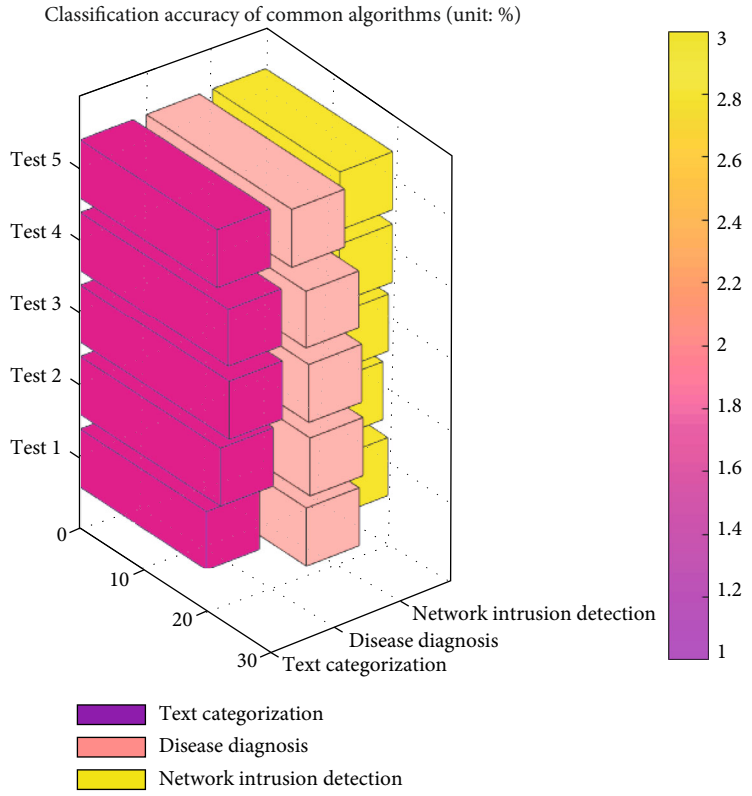

Figure 4: Classification accuracy of common algorithms (unit: %).

Table 4: The accuracy of the two algorithms compared (unit: %).

Test	Text categorization		Disease diagnosis		Network intrusion detection	
	Nonnegative matrix factorization	Ordinary decomposition	Nonnegative matrix factorization	Ordinary decomposition	Nonnegative matrix factorization	Ordinary decomposition
Test 1	78.34	20.12	63.34	25.45	80.34	19.45
Test 2	79.84	22.33	65.24	26.05	82.38	18.64
Test 3	77.65	23.74	64.92	25.84	84.23	19.56
Test 4	78.69	23.58	65.86	25.38	82.47	20.18
Test 5	78.82	21.83	64.38	23.12	83.23	20.23

unbalanced data obtained under the algorithm is very low, mainly because the recall and precision are small. Therefore, it can be seen that the unbalanced data flow integrated classification algorithm based on joint nonnegative matrix decomposition is more suitable for the data classification of

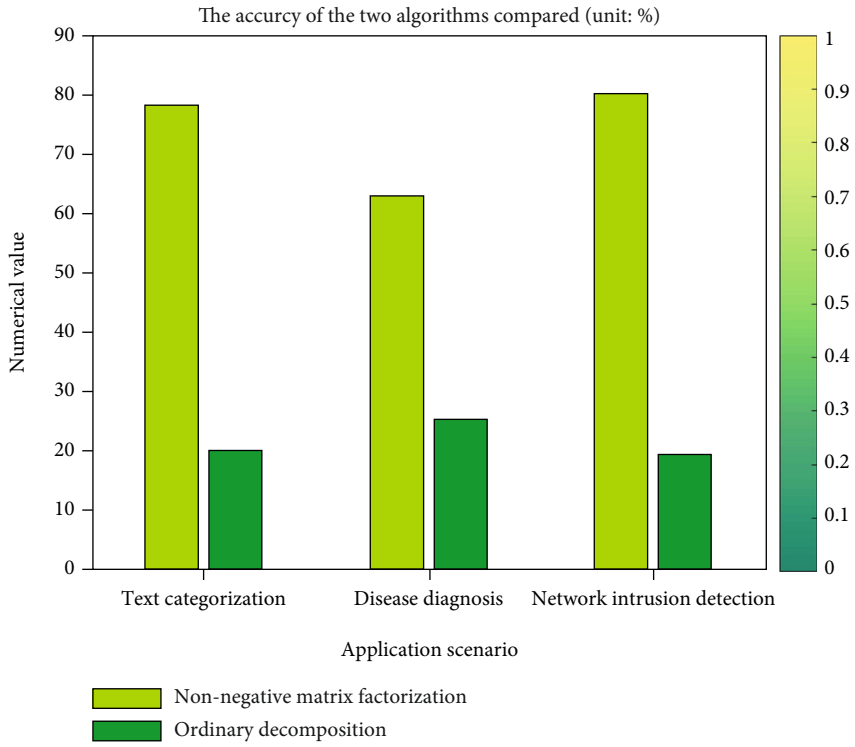

Figure 5: The accuracy of the two algorithms compared (unit: %).

Table 5: Comparison of the detection time consumption of the two algorithms (unit: h).

Test	Text categorization		Disease diagnosis		Network intrusion detection	
	Nonnegative matrix factorization	Ordinary decomposition	Nonnegative matrix factorization	Ordinary decomposition	Nonnegative matrix factorization	Ordinary decomposition
Test 1	1.50	5.3	1.84	5.6	1.32	4.84
Test 2	1.34	5.1	1.85	5.83	1.33	5.02
Test 3	1.24	5.22	1.78	5.89	1.23	4.99
Test 4	1.54	5.63	1.58	5.84	1.07	4.85
Test 5	1.50	5.23	1.63	6.09	1.13	5.12

unbalanced data. The data collection table is shown in Table 4 and Figure 5

(2) This experiment compares the time between the majority class and the minority class in text classification, disease diagnosis, and network intrusion detection using the unbalanced data flow integrated classification algorithm based on joint nonnegative matrix decomposition and the general unbalanced data flow integrated classification algorithm. Consume quickly. The data shows that after five tests, the text classification in the integrated classification algorithm based on unbalanced data flow under the joint nonnegative matrix decomposition, disease diagnosis classification, and network intrusion detection classification detection time consumption is shorter; and based on the general classification, under the algorithm, the detection time consumption of unbalanced data is shorter. The high accuracy of the integrated classification algorithm of unbalanced data flow based on joint nonnegative matrix decomposition reduces unnecessary errors, improves the consumption of detection time and speed, and obtains faster accurate classification results. The data collection table is shown in Table 5 and Figure 6

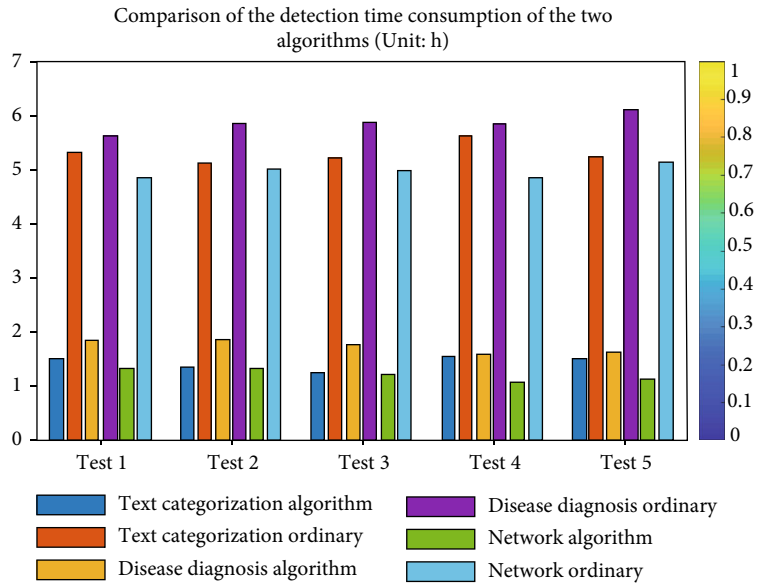

Figure 6: Comparison of the detection time consumption of the two algorithms (unit: h).

5. Conclusions

(1) In recent years, with the continuous deepening of research on class imbalance, the problem of class imbalance in data flow has attracted a large number of researchers. This paper proposes an ensemble classification framework to address the class imbalance problem in data streams. Its adaptive algorithm uses sampling techniques to deal with imbalance problems. The study led to the definition of a method for dealing with unbalanced data streams, which not only added examples of positive classes but also added classification errors in negative classes and also proposed a new method for defining class boundaries and negative to improve the integration effect of the classifier. An integrated classifier model for handling unbalanced data streams has been proposed, which combines weighted based integrated classifiers and sampling techniques. The data imbalance problem mainly exists in supervised machine learning tasks. When encountering unbalanced data, traditional classification algorithms with overall classification accuracy as the learning objective will pay too much attention to the majority class, thus degrading the classification performance of minority class samples. The vast majority of common machine learning algorithms do not work well with imbalanced data sets

(2) Classification is one of the main means of acquiring knowledge in the field of machine learning and data extraction. The most common classification algorithms, such as decision trees, tree networks, support vectors, and neural networks, have been used on a large scale. Existing classification algorithms usually assume that the data used for training is balanced; that is, the number of samples present in each type is approximately equal. In the case of imbalanced class data, the traditional classification algorithm (taking the accuracy of population classification as the learning target) pays too much attention to most classes, thereby reducing the ability to classify a few samples. But in fact, the cost of classification errors for a limited number of category samples is higher than most categories. For example, when predicting software defects, the size of defective samples is much smaller than the size of nondefective samples, but the purpose of classification is to identify a limited number of samples of defect categories. Other areas include medical diagnosis, oil spill control, cyber conspiracy control, and credit card fraud. The classification of unbalanced data is related to the performance of the learning algorithm when the class data is unbalanced or underexpressed. Based on the results of existing research, cost-sensitive techniques or sampling techniques can be used to reclassify data to solve the classification problem of classified data

(3) The classification learning of unbalanced data has a wide range of applications in many fields, such as software defect prediction and network intrusion detection. Due to the advantages of integration technology in dealing with unbalanced data learning, it is a research hotspot in the field of machine learning in recent years. The purpose of this paper is to study the unbalanced data flow integration classification algorithm based on joint nonnegative matrix factorization, in order to solve the problem that the basic clustering results obtained from the original data set have some information loss, thereby reducing the effective information in the integration stage. In this paper, the accuracy of the unbalanced data and the detection time consumption are selected as the

research object. Six data sets with imbalanced proportions of minority and majority samples are selected for experiments. Mathematical statistical analysis is first used to observe text classification, disease diagnosis, and network intrusion detection and the classification accuracy of majority class and minority class; the commonly used algorithm for unbalanced data is statistical analysis method. Comparing the univariate method for comprehensive classification of unbalanced data flow based on nonnegative matrix factorization with the unbalanced data algorithm, the observation has accurate rate and detects time-consuming changes. Experimental data shows that the unbalanced data flow integrated classification algorithm based on joint nonnegative matrix decomposition can reasonably evaluate the classification performance of the classifier for a few classes, and the detection speed is faster and saves more time. The experimental research shows that the algorithm combines the relationship matrix and information matrix from the original data set into a consensus function, uses NMF technology to obtain the membership matrix, effectively uses potential information, improves the accuracy rate of 69.73%, and shortens 71.65% of the time consumed. With the development of artificial intelligence deep learning, the advantages of deep network structure are becoming more and more obvious. In order to further study deep nonnegative matrix factorization, I think that with the research and development of deep nonnegative matrix factorization, more algorithms for optimizing deep model can be proposed, which will further improve the clustering performance of the model

Data Availability

No data were used to support this study.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this article.

References

[1] N. Lu and H. Miao, "Structure constrained nonnegative matrix factorization for pattern clustering and classification," *Neurocomputing*, vol. 171, pp. 400–411, 2016.

[2] C. Ying-Ying, Z. Xiao-Qiang, and C. Hao-Yan, "Case study of the molecular classification and prognostic prediction of gastric cancer based on nonnegative matrix factorization," *Journal of Shanghai Jiaotong University(Medical Science)*, vol. 37, no. 9, pp. 1187–1194, 2017.

[3] L. I. Xu, T. U. Ming, and W. Xiaofei, "Single-Channel speech separation based on non-negative matrix factorization and factorial conditional random field," *Acta Electronica Sinica*, vol. 27, no. 5, pp. 1063–1070, 2018.

[4] F. Segovia, J. M. Górriz, J. Ramírez, F. J. Martinez-Murcia, and M. García-Pérez, "Using deep neural networks along with dimensionality reduction techniques to assist the diagnosis of neurodegenerative disorders," *Logic Journal of the IGPL*, vol. 6, p. 6, 2018.

[5] F. Zhuang, P. Luo, and D. Changying, "Triplex transfer learning: exploiting both shared and distinct concepts for text classification," *IEEE Transactions on Cybernetics*, vol. 44, no. 7, pp. 1191–1203, 2014.

[6] T. Afzal, K. Iqbal, and G. White, "A method for locomotion mode identification using muscle synergies," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 25, no. 6, pp. 1–1, 2016.

[7] W. Sun, M. Jiang, and W. Li, "Band selection using sparse self-representation for hyperspectral imagery," *Geomatics & Information Science of Wuhan University*, vol. 42, no. 4, pp. 441–448, 2017.

[8] H. Rajaguru and S. K. Prabhakar, "Variational Bayesian matrix factorization and certain post classifiers for classification of epilepsy from EEG signals," *Research Journal of Pharmacy & Technology*, vol. 9, no. 6, p. 750, 2016.

[9] C.-H. Yeh, C.-Y. Lin, K. Muchtar, and P.-H. Liu, "Rain streak removal based on non-negative matrix factorization," *Multimedia Tools & Applications*, vol. 77, no. 15, pp. 20001–20020, 2018.

[10] S.-S. Wang, A. Chern, Y. Tsao et al., "Wavelet speech enhancement based on nonnegative matrix factorization," *IEEE Signal Processing Letters*, vol. 23, no. 8, pp. 1101–1105, 2016.

[11] L. Sun, G. Zhao, and D. Xinpeng, "CUR based initialization strategy for non-negative matrix factorization in application to hyperspectral unmixing," *Journal of Applied Mathematics & Physics*, vol. 4, no. 4, pp. 614–617, 2016.

[12] L. Tong, J. Zhou, Y. Qian, X. Bai, and Y. Gao, "Nonnegative matrix factorization based hyperspectral unmixing with partially known endmembers," *IEEE Transactions on Geoscience & Remote Sensing*, vol. 54, no. 11, pp. 6531–6544, 2016.

[13] Z. Zhang and Y. Liu, "A list-wise matrix factorization based POI recommendation by fusing multi-tag, social and geographical influences," *Journal of Internet Technology*, vol. 19, no. 1, pp. 127–136, 2018.

[14] W. Pak and Y. J. Choi, "High performance and high scalable packet classification algorithm for network security systems," *IEEE Transactions on Dependable & Secure Computing*, vol. 14, no. 1, pp. 37–49, 2017.

[15] T. Kim, B. Do Chung, and J.-S. Lee, "Incorporating receiver operating characteristics into naive Bayes for unbalanced data classification," *Computing*, vol. 99, no. 3, pp. 1–16, 2016.

[16] M. Nakata and T. Hamagami, "Revisit of rule-deletion strategy for XCSAM classifier system on classification," *Transactions of the Institute of Systems Control & Information Engineers*, vol. 30, no. 7, pp. 273–285, 2017.

[17] A. Care, F. A. Ramponi, and M. C. Campi, "A new classification algorithm with guaranteed sensitivity and specificity for medical applications," *IEEE Control Systems Letters*, vol. 2, no. 3, pp. 393–398, 2018.

[18] L. F. Gao, S. J. Zhao, and Y. U. Dong-Mei, "Unbalanced support vector machine coupling negative-samples cutting with asymmetric misclassification cost," *Acta Electronica Sinica*, vol. 45, no. 12, pp. 2978–2986, 2017.

[19] B. Zou, X. Chen, and L. Yang, "k-Times Markov sampling for SVMC," *IEEE Transactions on Neural Networks & Learning Systems*, vol. 29, no. 4, pp. 1328–1341, 2018.

[20] S. K. Mishra, S. C. Swain, and L. N. Tripathy, "Fault detection & classification in UPFC integrated transmission line using DWT," *International Journal of Power Electronics & Drive Systems*, vol. 8, no. 4, pp. 1793–1803, 2017.

[21] S. K. Mishra, S. C. Swain, and L. N. Tripathy, "A time-frequency transform based fault detection and classification of STATCOM integrated single circuit transmission," *International Journal of Power Electronics & Drive Systems*, vol. 8, no. 4, p. 1804, 2017.

[22] B. Hu, X. W. Li, S. T. Sun, and M. Ratcliffe, "Attention recognition in EEG-based affective learning research using CFS +KNN algorithm," *IEEE/ACM Transactions on Computational Biology & Bioinformatics*, vol. 15, no. 1, pp. 38–45, 2018.

[23] K. Ding and P. Y. Jiang, "Social sensors (S2ensors): a kind of hardware-software-integrated mediators for social manufacturing systems under mass individualization," *Chinese Journal of Mechanical Engineering*, vol. 30, no. 5, pp. 1150–1161, 2017.

[24] M. Marbac, C. Biernacki, and V. Vandewalle, "Latent class model with conditional dependency per modes to cluster categorical data," *Advances in Data Analysis and Classification*, vol. 10, no. 2, pp. 183–207, 2016.

[25] A. Cano, D. T. Nguyen, S. Ventura, and K. J. Cios, "Ur-CAIM: improved CAIM discretization for unbalanced and balanced data," *Soft Computing*, vol. 20, no. 1, pp. 173–188, 2016.