Научная новизна. Разработан DE-алгоритм с элитарной стратегией для применения в кластерном анализе по методу k-средних. Так как DE-алгоритм представляет собой метод для поиска оптимального решения путем имитации естественного эволюционного процесса, его отличительной особенностью является его скрытый параллелизм и способность эффективно использовать глобальную информацию, таким образом, новый и улучшенный алгоритм более устойчив и может избежать попадания в ловушку локального оптимума и значительно усилить эффект кластеризации. Исследования этого аспекта ранее не проводились.

Практическая значимость. Применение элитарной стратегии DE-алгоритма может повысить эффективность и точность кластерного анализа по методу K-средних. Результат экспериментального моделирования показал, что новый метод, представленный в этой статье, значительно улучшил производительность оптимизации, что доказывает его целесообразность и эффективность.

Ключевые слова: *кластерный анализ, метод K-средних, дифференциальная эволюция, элитарная стратегия, оптимизация производительности, целесообразность, эффективность*

Рекомендовано до публікації докт. техн. наук В.В.Гнатушенком. Дата надходження рукопису 17.04.15.

Liu Ning

Shangluo University, Shangluo, China

ENSEMBLE CLASSIFICATION ALGORITHM BASED IMPROVED SMOTE FOR IMBALANCED DATA

Лю Нін

Шанлонський університет, м. Шанло, КНР

ПОКРАЩЕНА SMOTE-СТРАТЕГІЯ КЛАСИФІКАЦІЇ НЕЗБАЛАНСОВАНИХ ДАНИХ НА ОСНОВІ АНСАМБЛЕВОГО АЛГОРИТМУ

Purpose. In practical application, the accuracy of the minority class is very important and the research on imbalanced data has become one of the most popular topics. In order to improve the classification performance for imbalanced data, the classification algorithm based on data sampling and integration technology for imbalanced data was proposed.

Methodology. Firstly, the traditional SMOTE algorithm was improved to K-SMOTE (an over-sampling method based on SMOTE and K-means). In K-SMOTE, the dataset was to perform clustering operation, and the interpolation operation was performed on the connection of the cluster center and the original data point. Secondly, ECA-IBD (an ensemble classification algorithm based improved SMOTE for imbalanced data) was proposed. In ECA-IBD, over-sampling was conducted by K-SMOTE, and random under-sampling was carried out to reduce the problem scale to form a new dataset. A number of weak classifiers were generated and integration techniques were used to form the final strong classifier.

Findings. Experiment was carried out on the UCI imbalanced dataset. The results showed that the proposed algorithm was effective by using the F-value and G-mean value as the evaluation indexes.

Originality. In the paper, we improved the SMOTE algorithm and combined over-sampling technology, under-sampling technology and boosting technology to solve the classification problem for imbalanced data.

Practical value. The proposed algorithm has important value in imbalanced data classification. It can be applied in the field of different kinds of imbalanced data classification, such as fault detection, intrusion detection, etc.

Keywords: *imbalanced data, ensemble learning, over sample, under sample, data classification*

Introduction. Classification problem is one of the most important in the field of data mining. Traditional classification methods have achieved good results on balanced datasets, but the actual datasets are often imbalanced. For the traditional classifier, it aims at pursuing the overall classification accuracy. The imbalance of the dataset is bound to cause the classifier to pay more attention to the majority class samples so that the classification performance of the minority class samples declines [1,2]. However, in practical application, people are more concerned about the minority class data, and the cost of the error in its classification is usually larger than that of the majority. For example, if the cancer patients were diagnosed as normal, it would delay the optimal timing of treatment, resulting in life threatening for patients. If the fault is identified as normal, it leads to failure undetected and may lead to major accidents. In network intrusion detection, if the network intrusion behavior is sentenced to normal behavior, it will have the potential danger to cause major network security incidents. Therefore, in practical application, it is more needed to improve the classification accuracy of the minority class samples. The research on imbalanced data has become one of the most popular topics [3].

The imbalanced classification is such a problem where the number of training samples in the class distribution is not balanced and the number of samples in one class is far

less than in the other one. In recent years, many scholars have proposed a variety of improved algorithms for imbalanced data classification. There are two main ways to improve the classification: one is realize dat the data level [4], the other,at the algorithm level [5−7].

On the data level, the method includes over-sampling and under sampling. It is intended to improve the imbalanced dataset by some mechanism and obtain a balanced data distribution. It is one of the important ways to deal with the imbalanced data classification because it is more advantageous to improve the overall classification performance [4].

Random over sampling is the most basic method to deal with imbalanced data. The algorithm replicates the minority class samples by random selection and adds the generated samples to the minority class. However, it is also possible to make the classifier learning appear overfitting.

Unlike the over-sampling, theunder-sampling is to remove the data from the original data. The most basic under-sampling technique is random under-sampling, which is to reduce the number of samples of the majority to make it the same as the number of the minority. However, it is also possible to lose the representative samples in the process of samples deleting.

At the algorithm level, the imbalanced data classification methods include cost sensitive learning [5], kernel method [6], integration method,etc.[7].

Ensemble classification learning is a machine learning technique. It uses a simple classification algorithm to get a number of different base classifiers that are combined in some way to receive a strong classifier. Ensemble learning plays an important role in the field of machine learning.

With the development of the integrated learning technology, more and more researchers introduce ensemble learning into the classification of imbalanced data and get many research results.

Galar, Fernandez, and Barrenechea (2013)[8] developed a new ensemble construction algorithm (EUSBoost) based on RUSBoost, one of the simplest and most accurate ensemble, which combined random under-sampling with Boosting algorithm. Khoshgoftaar, Van Hulse and Napolitano (2011) [9] compared the performance of several boosting and bagging techniques in the context of learning from imbalanced and noisy binary-class data. The experiments showed that the bagging techniques generally outperform boosting, and hence in noisy data environments, bagging was the preferred method for handling the class imbalance. Ghazikhani, Monsefi and Yazdi (2013) [10] proposed an online ensemble of neural network (NN) classifiers. The main contribution was a two-layer approach for handling class imbalance and non-stationarity. In the first layer, cost-sensitive learning was embedded in the training phase of the NNs, and in the second layer, a new method for weighting classifiers of the ensemble was proposed.

The combination of sampling technology and ensemble learning is an effective method to solve the problem of imbalanced data classification [8]. However, the existing algorithms are often unable to combine the advantages of the two methods effectively. For example, the traditional over-sampling technique blurred the boundaries of the majority and the minority. The traditional over-sampling technology leads to a large scale of data as well as low classification efficiency. It is also possible to lose some valuable data after processing the imbalanced datasets by using the under-sampling technique. In addition, the choice of integration algorithm often affects the classification accuracy of the algorithm.

To solve the above problems, the authors of the paper proposed an ensemble classification algorithm based improved SMOTE for imbalanced data (ECA-IBD), which combined the improved over-sampling technology, under-sampling technology and boosting technology to generate an efficient classifier for imbalanced data. Firstly, the existing over-sampling technology SMOTE was improved to increase the efficiency of the sampling. Secondly, the dataset was sampled by an improved over-sampling method to balance the dataset. The scale of the balanced dataset was reduced by the under-sampling technique. Thirdly, ensemble technology was used to generate a strong classifier to improve the performance of the classifier. At last, the experiment was carried out on the imbalanced dataset; the validity of the algorithm was verified by using F-value and G-mean values as the evaluation indexes.

Related works. ***SMOTE algorithm.*** SMOTE is a kind of over-sampling method that changes the balance of the dataset. It is to increase the number of minority class data and achieve a balance with the majority class.

In SMOTE, it searches for the nearest K adjacent samples in each data sample x of minority class dataset and randomly selects N samples in the nearest neighbor dataset recorded as $y_1, y_2, y_3, \ldots y_n$. The random linear interpolation operation is carried out between the minority class data x and y_i (j = 1, 2, N) to construct a new minority samples p_j. The interpolation operation is as follows

$$p_j = x + rand(0,1) * (y_j - x), j = 1,2, \ldots, N.$$

Where rand (0, 1) represents a random number in the interval (0,1) ,p_j represents new synthetic samples, x represents the sample of the minority class, y_j represents the j-th neighbor samples of x, these new synthetic minority class is merged into the original dataset to generate new training set.

K-means algorithm. The K-means algorithm is a kind of clustering algorithm based on the Euclidean distance, using distance as the similarity evaluation index. In K-means algorithm, the closer the distance between the two objects is, the greater the degree of similarity is. The cluster of the algorithm is composed of the objects, which are close to each other, so its final goal is to get a compact and independent cluster. The K-means algorithm is described as follows:

Step 1. Select k objects as the initial cluster center from N data objects.

Step 2. The distance between each object and the center object is calculated according to the mean value of each cluster, and the corresponding object is divided according to the minimum distance.

Step 3. Calculate the mean (center object) of each cluster.

Step 4. Calculate the standard measure function, when a certain condition is satisfied, the algorithm terminates. If the condition is not satisfied, return to *step 2.*

Methods. The traditional SMOTE algorithm was improved to K-SMOTE to prevent interpolation genera-

lization. Then, an integrated classification method for imbalanced dataset was proposed, which combined over-sampling and under-sampling.

The K-SMOTE algorithm based on K-means. There are two deficiencies in the SMOTE algorithm. First, the algorithm treats all the insertion location in the same way. Second, it blurred the boundaries of the majority and the minority. To demolish the defects, the improved algorithm needs to insert data items in the regional distribution and not to insert data items at the boundaries.

In K-SMOTE, the clustering operation was performed before interpolation, and the interpolation was performed in the clustering region, which could effectively prevent the interpolation generalization. At the same time, the interpolation formula was modified inK-SMOTE, and the interpolation data was on the connection between the cluster core and the original data point.

The principle of K-SMOTE algorithm was as follows: for the minority class, we first used the K-means algorithm for clustering operation. After cluster operation, the fixed K clusters were formed and the core of each cluster was recorded. The interpolation operation was performed for each cluster sample. The original sample point was interpolated by the cluster center, which was used as the original sample point. Specific steps of K-SMOTE were as follows:

1. Find the center of the minority class samples.

2. Create a new minority class. K-SMOTE improved the problem of SMOTE algorithm as follows

$$p_j = x + rand(0,1) * (X_c - x), j = 1,2,\ldots,N,$$

where $rand(0, 1)$ represents a random number in the interval (0, 1), p_j represents new synthetic samples, x represents the minority class. X_c represents the center of the minority class.

3. Replace the minority class of the original dataset with the new minority class. Then the new dataset was put into original dataset to get the final sample.

An ensemble classification algorithm based improved SMOTE for imbalanced data (ECA-IBD). An ensemble classification algorithm based improved SMOTE for imbalanced data named ECA-IBD was proposed. In ECA-IBD, K-SMOTE was used to perform over-sampling on imbalanced data to balance dataset, then, random under-sampling was carried out in equilibrium data to form multiple weak classifiers. Finally, the multiple weak classifiers were integrated to form the final strong classifier. The algorithm description of ECA-IBDis shown in Table 1.

Compared with the existing imbalanced data classification method, ECA-IBD used the K-SMOTE over-sampling technique to increase the number of the minority class and adjust the balance degree of the imbalanced dataset, to balance the data distribution. Under the condition of keeping the distribution of the whole dataset, the under-sampling was used to reduce the training data and reduce the size of the dataset, to reduce the training time of the model and improve the classification efficiency of the algorithm.

At the same time, ECA-IBD trained the weak classifier in each iteration process and used the ensemble learning method boosting technology to combine the classifier. According to the classification results,the samples were given a new weight to generate multiple weak classifiers, the final output results were obtained by the weight of the weak classifier. Therefore, the algorithm could improve the classification efficiency and increase the classification accuracy of the minority class.

Table 1

The algorithm description of ECA-IBD

Input: Dataset: $D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$;
Base classifier: C;
Over-sampling rate: M;
Under-sampling rate: N
Output: Strong classifier: $F(x)$
Process:
Step 1: The weights of the samples were initialized:

$$W(i) = 1/n$$

Step 2: The minority class P (Positive class) was sampled by K-SMOTE to form a balanced data at rate M.

Step 3: The whole dataset was under-sampled randomly at rate N under the condition of keeping the data distribution. The dataset D' was formed and its weight distribution was W.

Step 4: for $k = 1$ to K

1) Train weak classifier according to the training dataset D' and its weight distribution W', and calculate the weak hypothesis $h_t : X \times Y \to [0, 1]$

2) Calculate the pseudo loss of h_t:

$$\varepsilon_k = \sum_{(i,y):y_i \neq y} D_k(i)(1 - f_k(x_i, y_i) + f_k(x_i, y))$$

3) Calculate the weight update parameters:

$$\beta_k = \frac{\varepsilon_k}{1 - \varepsilon_k};$$

$$\omega_k = \frac{1}{2} \cdot (1 - f_k(x_i, y) + f_k(x_i, y_i))$$

4) Update weight distribution W_t:

$$W_{k+1}(i) = W_{k+1}(i)\beta_k^{\omega_k}$$

5) Normalization processing:

$$W_{k+1}(i) = \frac{W_{k+1}(i)}{\sum_i W_{k+1}(i)}$$

Step 5: The final classifier obtained by K weighted voting:

$$F(x) = \sum_{k=1}^{K} \beta_k \cdot f(x, y)$$

Experiment and result analysis. ***Evaluating indicator.*** For the classification method of balanced data, the classification accuracy is commonly used as an evaluation index. However, this evaluation index is the same for the cost of error classification of all kinds of samples, so the evaluation index is not reasonable in the imbalanced dataset.

Typically, in imbalanced datasets, the positive class (Positive) represents a minority class, and the negative class (Negative) represents the majority class. The evaluation index of imbalanced data is generally based on the confusion matrix.

As shown in Table 2, TP represents the number of positive samples that were assigned into the positive class; FP represents the number of positive samples that were assigned into the negative class; FN represents the number of negative samples that were assigned into the positive class; and TN represents negative samples that were assigned into the negative class.

Table 2

Confusion matrix

Category	Actual Positive Class	Actual Negative Class
Experimental positive class	TP	FN
Experimental negative class	FP	TN

The precision of reaction represents the ratio that actual positive sample which is classified as positive is accounted for all the actual positive class

$$Precision = \frac{TP}{TP + FP}.$$

The recall represents the ratio that actual positive sample, which is classified as positive is accounted for all the experimental positive class

$$Recall = \frac{TP}{TP + FN}.$$

Recall, Precision and F-value are the evaluation criteria for the positive class (minority class). In general, the F-value is used as the evaluation criterion for the classification of imbalanced datasets

$$F\text{-}value = \frac{(1 + l^2) \times Recall \times Precision}{l^2 \times Recall + Precision},$$

where λ expresses the relative importance of Recall and Precision, and λ is often assigned to 1.

G-mean is based on the correct classification rate of the minority class and the classification accuracy of the majority class, and it is usually used as a measure of the overall classification performance of the imbalanced dataset

$$G - mean = \sqrt{\frac{TN}{TN + FP} \times Recall}.$$

Here, F-value and G-mean were selected as the evaluation criteria to evaluate the performance of the algorithm on the imbalanced dataset.

Experiment data and results analysis. The experiment was carried out on UCI sets and the proposed algorithms were compared with the existing algorithm [4, 8] to make the effectiveness assessment.

In order to evaluate the effectiveness of the algorithm ECA-IBD for imbalanced data, 5 datasets were selected to carry out the experiment shown in Table 3. In the selected dataset, the number of minority class samples and majority class samples is not balanced.

In order to cancel the orders of magnitude difference between the dimensions of data and avoid large prediction error caused by differences in input and output, data normalization function was used here. The input feature value was normalized to [–1, 1] by data normalization function as follows

$$x_k = (x_k - x_{\min})/(x_k - x_{\max}).$$

Table 3

The dataset of theexperiment

Name of Dataset	Sample Number	Number of Minority	Proportion of Minority (%)	Attribute Number
breast-cancer	286	85	29.7	10
hepatitis	155	32	20.6	20
adult	1605	395	24.6	14
sonar	133	22	16.5	60
letter	20 000	789	3.9	16

The experiment was carried out 3 times, and the average value was taken as the final result. The experiment results are shown in Table 4 and Table 5. Table 4 shows the F-value comparison of the 4 algorithms in the 5 datasets. Table 5 shows the G-mean comparison of the 4 algorithms in the 5 datasets.

As we can see from the Table 4 and Table 5, the SVM was used to classify the imbalanced datasets directly;F-value and G-mean were relatively low. That was because it did not balance the imbalanced dataset.

We can also see that the SMOTE algorithm has been carried out to balance the partial dataset, so F-value and G-mean have been improved significantly.

As shown in Table 4 and Table 5, the AdaBoost used a number of classifiers for integration; the classification rate was higher than that of the SVM. However, the improvement was not very obvious in some data because the imbalance of the dataset was not handled. For example, in the Sonar dataset the improvement of values of F-value and G-mean was not very obvious.

Table 4

Comparison of experiment results (F-value value)

Name of Dataset	Classification Algorithm			
	SVM	SMOTE	AdaBoost	ECAIBD
breast cancer	0.508	0.695	0.522	0.755
hepatitis	0.537	0.812	0.674	0.834
adult	0.624	0.628	0.634	0.652
sonar	0.456	0.521	0.491	0.683
letter	0.732	0.894	0.864	0.934

Table 5

Comparison of experiment results (G-mean value)

Name of Dataset	Classification Algorithm			
	SVM	SMOTE	AdaBoost	ECAIBD
breastcancer	0.554	0.722	0.735	0.790
hepatitis	0.718	0.881	0.904	0.937
adult	0.688	0.691	0.689	0.703
sonar	0.555	0.655	0.568	0.706
letter	0.901	0.947	0.932	0.958

Because ECA-IBD used K-SMOTE algorithm to balance the imbalanced dataset and used the integrated technology to strengthen the classifier, so it had better classification rate, and the F-value and G-mean were higher than in the other algorithms.

At the same time, compared with AdaBoost, the modeling time of ECA-IBDwas reduced from 18.5s to 6.7 s. It was because the ECA-IBD performed under-sampling for the balanced dataset, reducing the size of the sample set and shortening the running time.

Conclusions. Based on the improved SMOTE algorithm and integration technology, we proposed an integrated classification algorithm for imbalanced data. First, the traditional SMOTE algorithm was improved to K-SMOTE, reducing the defects of the SMOTE algorithm. Then, combined with the classifier ensemble technology, an integrated classification algorithm for imbalanced data named ECA-IBD was proposed. In ECA-IBD, K-SMOTE was used to conduct over-sampling, and random under-sampling was carried out to reduce the problem scale and form a new dataset. In the new dataset, a number of weak classifiers were trained to generate, and integration techniques were used to integrate several weak classifiers to form the final strong classifier. The experiment was carried out on the UCI dataset, F-value and G-mean were used as the evaluation indexes to evaluate the proposed algorithm. The experiment result have proved the effectiveness of the new algorithm.

Acknowledgements. The work was supported by Natural Science Basic Research Plan in Shaanxi Province of China (No.2015JM6347), Science Technology Plan in Shangluo City of China (No. SK2014-01-15), Science Research Plan of Shangluo University (No.14SKY026).

References / Список літератури

1. Napierała, K. and Stefanowski, J., 2015. Addressing imbalanced data with argument based rule learning. *Expert Systems with Applications*, vol. 24, no. 24, pp. 9468–9481.
2. Ditzler, G. and Polikar, R., 2013. Incremental learning of concept drift from streaming imbalanced data. *IEEE Transactions on Knowledge & Data Engineering,* vol. 25, no. 10, pp. 2283–2301.
3. Maldonado, S. and López, J., 2014. Imbalanced data classification using second-order cone programming support vector machines. *Pattern Recognition,* vol. 47, no. 5, pp. 2070–2079.
4. Barua, S., Islam, M. M. and Yao, X., 2014. MWMOTE-majority weighted minority-oversampling technique for imbalanced dataset learning. *IEEE Transactions on Knowledge & Data Engineering*, vol. 26, no. 2, pp. 405–425.
5. Castro, C. L. and Braga, A. P., 2013. Novel cost-sensitive approach to improve the multilayer perceptron performance on imbalanced data. *IEEE Transactions on Neural Networks & Learning Systems*, vol. 24, no. 6, pp. 888–899.
6. Maratea, A., Petrosino, A. and Manzo, M., 2014.Adjusted F-measure and kernel scaling for imbalanced data learning. *Information Sciences*, vol. 257, no. 257, pp. 331–341.
7. Sun, Z., Song, Q. and Zhu, X., 2015. A novel ensemble method for classifying imbalanced data. *Pattern Recognition*, vol. 48, no. 5, pp. 1623–1637.
8. Galar, M., Fernández, A. and Barrenechea, E., 2013. EUSBoost: Enhancing ensembles for highly imbalanced datasets by evolutionary undersampling. *Pattern Recognition*, vol. 46, no. 12, pp. 460–3471.
9. Khoshgoftaar, T. M., Van Hulse, J. and Napolitano, A., 2011. Comparing boosting and bagging techniques with noisy and imbalanced data. *IEEE Transactions on Systems Man and Cybernetics – Part a Systems and Humans,* vol. 41, no. 3, pp. 552–568.
10. Ghazikhani, A., Monsefi, R. and Yazdi, H. S., 2013. Ensemble of online neural networks for non-stationary and imbalanced data streams. *Neurocomputing,* vol. 122, pp. 535–544.

Мета. У практичному застосуванні точність міноритарного класу дуже важлива, тому дослідження незбалансованих даних стало одним з найпопулярніших напрямів. З метою підвищення ефективності класифікації незбалансованих даних, у статті запропоновано алгоритм класифікації на основі вибірки даних і технології інтеграції незбалансованих даних.

Методика. По-перше, традиційний алгоритм SMOTE був поліпшений до K-SMOTE (метод збільшення числа прикладів міноритарного класу, що об'єднує стратегію семплінгу SMOTE та метод K-середніх). У K-SMOTE, набір даних підлягав кластеризації, а інтерполяція проводилася між центром кластера та точкою вихідних даних. По-друге, був запропонований алгоритм ECA-IBD (поліпшена SMOTE-стратегія класифікації незбалансованих даних на основі ансамблевого алгоритму). У ECA-IBD, збільшення числа прикладів міноритарного класу проводилося за допомогою K-SMOTE, а зменшення числа прикладів мажоритарного класу проводилося методом випадкового відбору, з метою зменшення масштабу проблеми й формування нового набору даних. Цілий ряд слабких класифікаторів і методів інтеграції було використано для формування кінцевого сильного класифікатора.

Результати. Експеримент проводився на UCI наборі незбалансованих даних. Результати показали, що запропонований алгоритм ефективний за використання F-значення та G-середнього значення в якості оціночних індексів.

Наукова новизна. Покращено алгоритм SMOTE й скомбіновані стратегії збільшення числа прикладів міноритарного класу та зменшення числа прикладів мажоритарного класу, і технологія бустінгу для вирішення задач класифікації незбалансованих даних.

Практична значимість. Запропонований алгоритм має важливе значення для класифікації незбалансованих даних. Він може застосовуватися в багатьох областях, таких як виявлення несправностей, вторгнення і т. п.

Ключові слова: *незбалансовані дані, композиційне навчання, збільшення числа прикладів міноритарного класу, зменшення числа прикладів мажоритарного класу, класифікація даних*

Цель. В практическом применении точность миноритарного класса очень важна, поэтому исследование несбалансированных данных стало одним из самых популярных направлений. С целью повышения эффективности классификации несбалансированных данных, в статье предложен алгоритм классификации

на основе выборки данных и технологии интеграции несбалансированных данных.

Методика. Во-первых, традиционный алгоритм SMOTE был улучшен до K-SMOTE (метод увеличения числа примеров миноритарного класса, объединяющий стратегию семплинга SMOTE и метод K-средних). В K-SMOTE, набор данных подлежал кластеризации, а интерполяция проводилась между центром кластера и точкой исходных данных. Во-вторых, был предложен алгоритм ECA-IBD (улучшенная SMOTE-стратегия классификации несбалансированных данных на основе ансамблевого алгоритма). В ECA-IBD, увеличение числа примеров миноритарного класса проводилось с помощью K-SMOTE, а уменьшение числа примеров мажоритарного класса проводилось методом случайного отбора, с целью уменьшения масштаба проблемы и формирования нового набора данных. Целый ряд слабых классификаторов и методов интеграции был использован для формирования конечного сильного классификатора.

Результаты. Эксперимент проводился на UCI наборе несбалансированных данных. Результаты показали, что предложенный алгоритм эффективен при использовании F-значения и G-среднего значения в качестве оценочных индексов.

Научная новизна. Улучшен алгоритм SMOTE и скомбинированы стратегии увеличения числа примеров миноритарного класса и уменьшения числа примеров мажоритарного класса, а также технология бустинга для решения задач классификации несбалансированных данных.

Практическая значимость. Предложенный алгоритм имеет важное значение для классификации несбалансированных данных. Он может быть применен во многих областях, таких как обнаружение неисправностей, обнаружение вторжения и т. п.

Ключевые слова: *несбалансированные данные, композиционное обучение, увеличение числа примеров миноритарного класса, уменьшение числа примеров мажоритарного класса, классификация данных*

Рекомендовано до публікації докт. техн. наук В. В. Гнатушенком. Дата надходження рукопису 22.04.15.

Guangbin Sun[1],
Hongqi Li[1],
Haiying Huang[2]

1 – China University of Petroleum, Beijing, China
2 – Daqing Oilfield Engineering Co,Ltd, Daqing, Heilongjiang, China

IMPROVED K-MEANS ALGORITHM AUTOMATIC ACQUISITION OF INITIAL CLUSTERING CENTER

Гуанбінь Сунь[1],
ХунціЛі[1],
Хайїн Хуан[2]

1 – Китайський університет нафти, м. Пекін, КНР
2 – Дацин Ойлфілд Інжиніринг Ко, Лтд, м. Дацин, КНР

УДОСКОНАЛЕНИЙ АЛГОРИТМ К-СЕРЕДНІХ АВТОМАТИЧНОГО ВИЗНАЧЕННЯ ПОЧАТКОВИХ ЗНАЧЕНЬ ЦЕНТРІВ КЛАСТЕРІВ

Purpose. The traditional K-means algorithm requires the K value, and it is sensitive to the initial clustering center. Different initial clustering centers often correspond to the different clustering results, and the K value is always required. Aiming at these shortcomings, the article proposes a method for getting the clustering center based on the density and max-min distance means. The selection of the clustering center and classification can be carried out simultaneously.

Methodology. According to the densities of objects, the noise was eliminated and the densest object was selected as the first clustering center. The max-min distance method was used to search the other best cluster centers, at the same time, the cluster, which the object belongs to, was decided.

Findings. Clustering results are related to the selection of parameters θ. If the sample distribution is unknown, only test method can be used through multiple test optimization. With prior knowledge for the selection of θ, it can be converged quickly. Therefore, θ should be optimized.

Originality. This article proposes the new method based on the density to get the first initial clustering center, and then the new method based on the maximum and minimum value. The improved algorithm obtained through experimental analysis insures higher and stable accuracy.

Practical value. The experiments showed that the algorithm allows for automatic obtaining of the k clustering centers and have a higher clustering accuracy in unknown datasets processing.

Keywords: *clustering, K-means clustering, max-min distance method,density*

Introduction. Clustering means that a given object is divided into several clusters based on the given definition of similarity so that the objects within a cluster can be as similar as possible and the objects of different clusters can be as different as possible. According to the clustering rules, the clustering algorithm can be divided into: based