Pattern Recognition Letters 133 (2020) 280–286

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Sample imbalance disease classification model based on association rule feature selection

Chenxi Huang[a], Xin Huang[b,c,*], Yu Fang[c], Jianfeng Xu[b], Yi Qu[c], Pengjun Zhai[c], Lin Fan[a], Hua Yin[d], Yilu Xu[d], Jiahang Li[d]

[a] School of informatics, Xiamen University, Xiamen 361005, China
[b] School of Software, Nanchang University, Nanchang 330047, China
[c] Department of Computer Science and Technology, Tongji University, ShangHai 201804, China
[d] Software College, Jiangxi Agricultural University, Nanchang 3300029, China

ARTICLE INFO

Article history:
Received 3 December 2019
Revised 22 February 2020
Accepted 9 March 2020
Available online 11 March 2020

MSC:
41A05
41A10
65D05
65D17

Keywords:
Association rules
Feature selection
Integrated learning
Sample imbalance

ABSTRACT

In the research of computer-aided diagnosis, the shortage of disease feature dimension curse and the imbalance of medical samples have always been the focus of research on diagnostic decision support systems. For these two problems, we propose a feature selection algorithm based on association rules and an integrated classification algorithm based on random equilibrium sampling. We extracted and cleaned the electronic medical record text obtained from the hospital to obtain a diabetes data set. The proposed algorithm was verified in this data set and the public data set UCI. Experimental results show that the feature selection algorithm based on association rules is better than the CART, ReliefF and RFE-SVM algorithms in terms of feature dimension and classification accuracy. The proposed integrated classification algorithm based on random equalization sampling is superior to the comparative SMOTE-Boost and SMOTE-RF algorithms in macro precision, macro-full rate and macro F1 value, which embodies the robustness of the algorithm.

1. Introduction

Diagnostic decision support system, as an application branch of Clinical Decision Support System (Clinical Decision Support System, CDSS) [2], has been a highline of research and application. Machine learning is widely used in medical information mining [11] and diagnostic [12,20,28,31,32] research. According to the basis of CDSS decision support, it can be divided into two categories: knowledge-based CDSS system and non-knowledge CDSS system. The knowledge-based CDSS system is essentially an expert system in the medical field. It contains a large number of medical experts' knowledge and experience. It usually consists of human-machine interface, knowledge base and inference engine. Unlike knowledge-based CDSS,non-knowledge-based CDSS is integrated with the Electronic Medical Record System(Electronic Medical Record System, EMR), and its decision support is based on algorithms such as machine learning or other statistical algorithms for pattern recognition. EMR data sets for association rule mining, classification, regression, etc. can continuously discover new knowledge to help doctors make better decisions in the disease diagnosis process. With the continuous development of data mining, pattern recognition and machine learning algorithms, and the continuous accumulation of data in electronic medical records, the research of CDSS based on machine learning has gradually become the hotspot and mainstream direction of current diagnostic decision support systems.

In the current study, two issues remain. First, it is difficult to construct a diagnostic model directly for medical cases with fewer training samples. And as the diagnostic capabilities of diagnostic models improve, the required features will continue to expand, resulting in a dimensionality of the feature matrix, resulting in more redundant and uncorrelated features, excessive computation, sparse training samples, and overfitting. Other issues that ultimately affect the classification quality of the classifier. In the disease diagnosis model, the number of diagnosed disease categories is small, and the classification accuracy rate is low, which makes the diagnostic decision model practicable. At the same time, in the electronic medical record, because there are a large number of characteristic attributes, the selection of corresponding features is

* Corresponding author.
E-mail address: 1610466@tongji.edu.cn (X. Huang).

https://doi.org/10.1016/j.patrec.2020.03.016
0167-8655/

different for different diseases. How to select the optimal features is the difficulty to improve the multi-classification task of diseases. Second, most existing diseases have the characteristics of disease distribution imbalance [5]. In the disease multi-classification problem, sample imbalance is a key factor restricting the disease multi-classification problem, because for some specific diseases In the case of the case, the number of sample sets is relatively small, which leads to sparse training samples when classifying, affecting classification accuracy and generalization performance of multi-classification tasks.

First, we searched and analyzed the association rules between diabetes and its complications and symptoms in the EMR dataset. Then we sorted the symptom feature attributes based on the two frequent sets of confidence parameters of the disease-disease, and adopted the sequence forward selection method, and the classification performance of the classifier is selected to overcome the problems of large computational complexity, training sample sparseness and over-fitting caused by the dimensionality disaster encountered in the feature matrix. Finally, the characteristics of the sample imbalance in the EMR data set are adopted. The subset of the training samples is divided into categories according to the category, and the feature vectors selected in the previous stage are used to train the unbalanced sample set, and then the random equalization sampling is performed in the iterative process, and then the training base classifier and the evaluation base classifier are passed. The F1 value is used for weighted voting, and finally an integrated classifier with the best classification performance is output, and the integrated classifier is used to complete the classification task of multiple diseases, thereby improving the quality of disease diagnosis decision.

Overall, the main contributions of our work are:

- We propose a disease feature selection algorithm based on association rules, which can screen out feature vectors for multi-disease classification and effectively improve the quality of multi-disease classification.
- We propose an integrated algorithm based on stochastic equalization, which can effectively improve the multi-disease classification of sample imbalance.

The rest of this paper is organized as follows. Section II reviews the related work for image captioning and radiology report generation. Section III shows data preprocessing and details of the data. Section IV details the design of the proposed algorithm. Section V and VI present and discuss the experimental settings and results, respectively. Finally, we draw conclusion in Section VII.

2. Related work

Feature Selection. John et al. [13] considers feature selection to be a process of reducing feature dimensions without reducing classification accuracy. Koller et al. [17] defines feature selection to select as small a feature subset as possible, while ensuring that the result class distribution is as similar as possible to the original data class distribution. Dash et al. [7] gave a comprehensive overview of the feature selection problem in the field of data mining, and gave the basic framework of feature selection. From the evaluation criteria of feature sets, feature selection methods can be divided into the following three categories: Embedded, Filter [22] and Wrapper [19]. In the embedded structure, the feature selection algorithm itself is embedded as a component in the classification algorithm, the most typical is the decision tree algorithm. Classic decision tree algorithms include Quinlan's ID3 and C4.5 [23] and Breiman's Classification and Regression Trees (CART) algorithm [2]. The evaluation criteria for filter feature selection are directly obtained from the data set, independent of the classification algorithm, and are suitable for large-scale data sets. Kira et al. [16] proposed an effective feature selection algorithm, Relief algorithm. The feature selection criterion of this algorithm is feature correlation. The biggest limitation of the Relief algorithm is that redundant features cannot be identified in the relevant feature set, and generally only binary data can be used. Kononenlko et al. [18] extended the original Relief algorithm and obtained the ReliefF algorithm, which enables it to handle multi-category, incomplete and noisy data. The wrapped feature selection algorithm has higher accuracy than the filter feature selection algorithm, but the algorithm is less efficient. Hsu et al. [10] used decision trees to perform feature selection, and used genetic algorithms to find a set of feature subsets that minimized the decision tree classification error rate. Chiang et al. [6] combined Fisher discriminant analysis with genetic algorithm to identify key variables during chemical faults and achieved good results. Guyon et al. [9] proposed the SVM-RFE (Recursive feature selection based on Support vector machine, SVM-RFE) method, also known as the support vector machine based regression feature elimination method. The algorithm considers that the RFE algorithm can guarantee the optimization of feature subsets in the process of feature sorting. When sorting features, the method uses the information in the discriminant function of support vector machine to realize feature selection and improve the performance of classification.

Unbalanced Data Sampling. Kermanidis et al. [14] used the oneside sampling technique to improve the classification accuracy of the classifier in order to solve the problem of data set imbalance. Similarly, by using the oversampling algorithm [30], it is also possible to solve the problem that the small sample data volume is insufficient to cause imbalance. In the field of oversampling technology, the SMOTE algorithm [4] is the mainstream over-sampling algorithm, but it also has certain drawbacks. Because the distribution of neighbor samples is not taken into account in the sample synthesis process, it is easy to cause the problem of sample overlap. The sample synthesis is very blind. The integration method is a common method to solve the sample imbalance problem [24,29]. Thanathamathee et al. [26] proposes a method to generate boundary cluster data using AdaBoost algorithm to deal with the unbalanced data set. The method considers the distribution of samples. In addition, the literature [21] through the Bagging method to improve the prediction performance of the two-class model under the unbalanced sample, multiple sampling and training multiple basic classifiers, and finally combined into a strong classifier, although can improve certain performance, but each classifier The separation between trainings limits the overall performance improvement. In ref. [8], the authors used three learning techniques: set learning, artificial sample generation, and by re-marking data into new sets, combining them into a new learning framework, a diversified integrated classifier, for solving sample imbalances. Learning problems under conditions. In ref. [25], the authors propose a novel integration method that first converts an unbalanced data set into multiple balanced data sets, and then builds multiple basic classifiers on these data sets using a specific classification algorithm. Combine these basic classifiers into an integrated classifier.

2.1. Data collection

The experimental data set comes from the data related to diabetes and its complications in the electronic medical record database of a Three-A hospital in Shanghai. Because there are some incomplete, noisy, and inconsistent dirty data in the medical record data, the model training cannot be directly performed. In order to ensure the integrity and accuracy of the data and improve the accuracy of the diagnostic decision model, we perform data preprocessing on data cleaning, data transformation and reduction, and

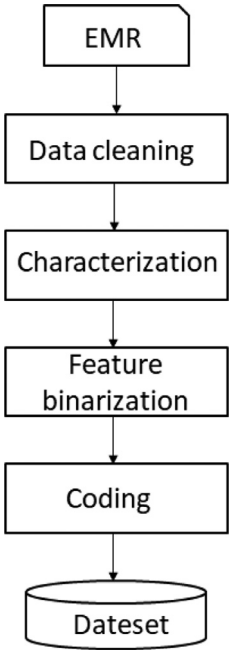

Fig. 1. Electronic medical record data extraction and cleaning flow chart.

Table 1
Diabetes and its complications.

Diabetes	Cases	Features
Type 2	2672	40
Type 1	87	22
with Gestational	91	33
with kidney	1103	55
with neurological	501	44
with ketoacidosis	317	36
with foot	173	15
with coma	63	18
with eyes	51	16
peripheral vascular	46	18

classification feature coding. The process of data preprocessing is shown in Fig. 1:

Data cleaning: delete data with gender, age, symptom, and disease attribute values is empty; delete data with special identifier.

Characterization of feature attributes: data transformation and reduction, reduction of the number of valid variables or invariance of data by data transformation or reduction.

Feature binarization: After the symptom feature attribute is normalized, the symptom feature binarization operation is performed for each piece of data.

Coding: Classification and coding of disease categories. Finally get the diabetes data set. After data integration and integration, the sample contains 5104 cases of diabetes, including 10 different disease types and 130 different disease symptom characteristics, and the sample distribution of different disease categories in the diabetes dataset is extremely uneven. The specific type name, number of cases, and characteristic number of diseases of this type are shown in Table 1.

3. Methods

3.1. Association rules features selection

In order to reduce the size of the feature subset and improve the efficiency of the feature selection algorithm without reducing the classification accuracy, we propose an ARFS) Association Rules Features Selection, ARFS) Algorithm 1. The ARFS algorithm first takes the maximum value strategy to calculate the confidence between the feature and the category, and uses the confidence value to evaluate the correlation between the feature and the category. This correlation affects the selection of the feature subset. The features are then sorted by the size of their correlation weights, and an ordered sequence of features with a large correlation between feature and category is obtained. Because the feature sequence is sorted by the correlation from large to small, the sequence forward selection of feature subsets on this basis can select the feature subset with the smallest scale without reducing the classification accuracy. Because the ARFS algorithm adopts the forward selection strategy and the ordered feature sequence, the time complexity of the search strategy can be reduced. The pseudo code of the algorithm is described as follows.

Algorithm 1 ARFS.

Input:
2 frequent sets: $L2$;
2 frequent sets of confidence: $conf$;
The number of iterations β;
Divide the step size divide length;
Classifier CART.

Output:
Feature vector: $feature\_vector$

Init $feature\_vector$;
$Vec = Sort(Max(L_2, conf))$
while $i < \beta$ **do**
  if $divide\_flag == true$ **then**
    Update $divide\_length$
  end
  $feature\_vector$ add(Vec_d)
  $F\_\max = Max(F_t(feature\_vector))$
end
Return $feature\_vector$

In the above algorithm, when the two frequent sets $L2$ are sorted according to the confidence degree, the Max strategy is adopted, and the strategy is to perform feature sorting by selecting the maximum confidence of multiple feature attribute values in a feature for the confidence degree of the feature; When the current feature subset has the highest classification accuracy rate accuracy max, the Max strategy is also adopted to select the classification accuracy of the current feature subset and the maximum value of the classification accuracy of the current optimal feature subset, and the accuracy max will follow each iteration. The maximum value of the classification accuracy is calculated continuously during the process. In the SFS(Sequence Forward Selection, SFS) method, the feature is gradually added to the feature subset from the front to the back in a certain step size according to the ordered feature sequence. The default value of the step width length is 1. In the process of iteratively selecting the optimal feature subset, the criterion for stopping the criterion is that the accuracy difference between the current feature subset and the current optimal feature subset classification is less than 0, ie $\Delta_{acc} < 0$; and must also comply with Δ The condition that the frequency of $\Delta_{acc} < 0$ is smaller than the value of the parameter β, and the default value of β is the length of the original feature vector. When the above two conditions are satisfied, the iteration can be stopped, and finally the feature subset solved by the ARFS algorithm is output.

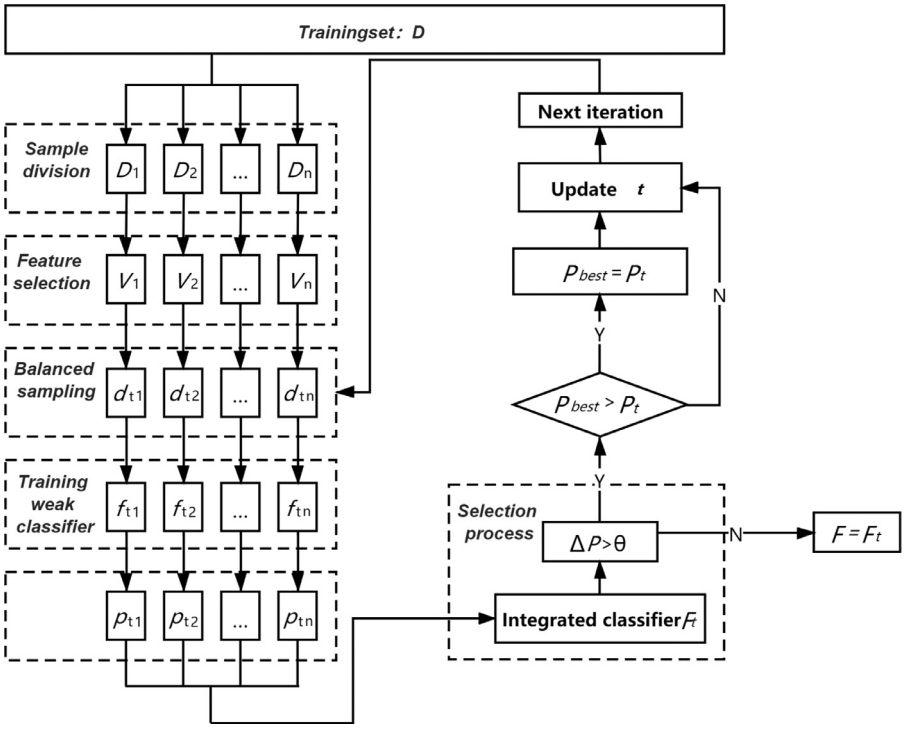

Fig. 2. Integrated Classification Algorithm Based on Random Equilibrium Sampling.

3.2. Integrated classification algorithm based on random equilibrium sampling

Although the integration of multiple base learners is usually superior to the generalization performance of a single learner, in some cases, poor results may occur. If you want to ensure that you can generate better integrated learning algorithms as much as possible, you must be able to ensure that each base learner is "good and different", that is, each base learner must have certain conditions under certain preconditions with certain accuracy. Diversity. Bagging is a common integrated learning method. Through the bootstrap sampling method, each round of random sampling process uses only about 63.2% of the original training set, while the training set is about 36.8%. Data that is not sampled into the sample set can be used as a validation set to perform "out-of-package estimation" of generalization performance [27]. We propose a RBSBagging(Balanced Sampling with Bagging, RBSBagging) algorithm based on the sample imbalance problem in the medical field or other research fields. The core idea of the algorithm is that based on the Bagging algorithm, the process of attribute selection is added similarly to RF(Random Forest, RF) [22], but the attribute selection is not random but based on the sample category. Perform attribute selection, then perform sample partitioning by category, equalize sampling, and train each sample-equalization classifier that focuses on its own category, and then continuously evaluate it by weighted voting and during iterative equalization sampling. The integrated classifier with the best classification performance is updated, and finally an integrated multi-classifier with strong classification generalization performance is obtained. The algorithm flow is shown in Fig. 2.

In the above RBSBagging algorithm, the training sample set $D = (x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ is first divided into n sample subsets D_i by category Y, where a is b groundtruth. Then, the feature selection algorithm ARFS is performed on each sample subset D_i, and the appropriate feature vector V_i is selected; and the feature vector V_i of each D_i sample subset is deduplicated and integrated to form the feature vector V of the sample training set D. Next, it enters the iterative loop. First, for each D_i sample subset, random equalization sampling is performed according to the updated sampling probability t. The essence of the random equalization sampling is to randomly adjust the imbalance degree IR of the sample subset D_i in the iterative process. The operation is that the sampling probability t takes a random number in the interval of the lower limit value of 0 bit and the upper limit value of IR. In this way, the imbalance degree IR of each sample subset D_i is randomly adjusted in the iterative process. Unbalance IR refers to the ratio of the majority of the sample in the sample data set to the sample of the minority category. Its main role is to measure the degree of imbalance in the binary data set. After each round of sampling, the base classifier fi of each D_i sample subset is trained. After the training is completed, the F1 value is calculated, recorded as p_i, and the voting weight w_i is updated. as shown by the formula 1.

$$w_i = \frac{p_i}{\frac{1}{N}\sum_{i=1}^{n} p_i} \tag{1}$$

For the t-th round, the multiple base classifiers f_i are trained, and all of them are calculated after their voting weights w_i(usually $w_i \geq 0$, $\sum_{i=1}^{n} w_i = 0$). The n-base classifier can be integrated by the weighted voting formula 2. F_i, the integrated classifier F_t that generates the current iteration round.

$$H_x = \frac{1}{T}\sum_{i=1}^{t} w_i f_i(x) \tag{2}$$

First, by calculating the macro-F1 value of its F_t, then by judging whether the difference ΔF_1 between the current macro-F1 value and the current optimal macro-F1 value is greater than the convergence threshold θ; If it is greater than, update the integrated classifier F_t of the current optimal macro-F1 value and the probability t of the equalization sample, and enter the next iteration; otherwise, end the loop and output the final integrated multi-classifier F; finally find the pair in the iteration Small sample and overall sample set classification best performance integrated classifier. The

pseudo code of the feature selection algorithm based on the association rule is as shown in Algorithm 2.

Algorithm 2 RBSBagging.

Input:

Unbalanced training set:D;
Weak learning algorithm: f;
ARFS algorithm;
Unbalance threshold:IR;
Convergence threshold:θ.

Output:

Final integrated classifier: $F(x)$

Divide training sample D into categories by sub-sample D_i by category Y;
$V_i = ARFS(D_i)$;
Integrate feature vector V_i, remove duplicate feature attributes, and get V;
Update equalization sampling probability:$Max(macro\_F1)$
Equalization sampling based on equalized sampling probability t;
Training base classifier f_i, calculate $F1$ value: p_i;
Update the base classifier voting weight according to the p_i value w_i;
Multi-classifier based on weight f_i integration;
Calculate $\Delta F1$
while $\Delta F1 < 0$ **do**
    Return $F(x) = sign(\sum_{i=0} p_i \times f_i)$ **then**
end

4. Experiments

4.1. Metrics

We use P(Precision,P), R(Recall,R), F-1 values as our basic evaluation indicators, and their calculation formulas are as follows:

$$P = \frac{TP}{TP + FP} \tag{3}$$

$$R = \frac{TP}{TP + FN} \tag{4}$$

$$F1 = \frac{2 \times P \times R}{P + R} \tag{5}$$

Where TP is True Positive, FP is False Positive, FN is False Negative, and TN is True Negative. To further comprehensively assess the accuracy, recall, and F-mearsure values of a multi-category problem, we can also use Macro-averaging to first calculate the metrics for each category and then all categories. The evaluation indicators are arithmetically averaged. The calculation formula is as follows:

$$macro_p = \frac{1}{n}\sum_{i=1}^{n} p_i \tag{6}$$

$$macro_{F_1} = \frac{2 \times macro_p \times macro_R}{macro_p + macro_R} \tag{7}$$

4.2. Details

4.2.1. ARFS Algorithm

We use the Apriori algorithm [1] to filter out some of the extraneous redundant features, which not only reduces the time complexity of the algorithm. When the minimum support min sup threshold is set to 0.001, the partial results of the disease-disorder 2 frequent sets mined by the 2 frequent set mining algorithm are shown in Table 2.

Table 2
Diabetes and its complications.

2 frequent sets	Support	Confidence
Elevated blood sugar	0.3017	0.5645
Drink more	0.0086	0.0249
Gestational week	0.0123	1
Elevated creatinine	0.0643	1
Numbness of limbs	0.0290	0.4852
Vomiting	0.0165	0.3889
Foot ulceration	0.0143	1
Unconscious	0.0069	1
Blurred vision	0.0049	0.1923
Lower limbs	0.0024	1

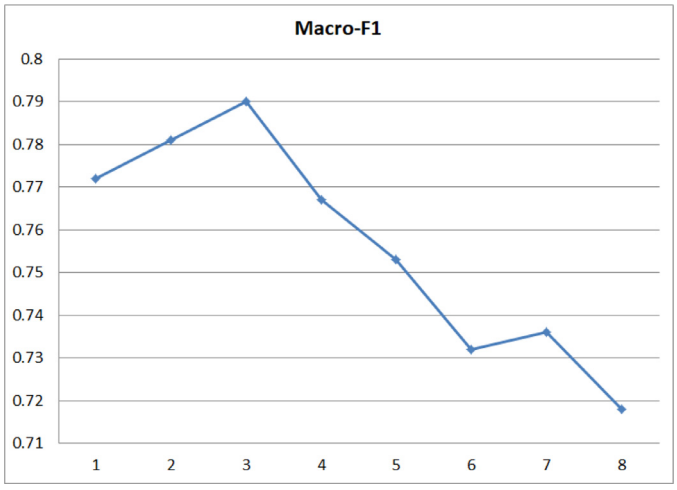

Fig. 3. Comparison of results of different IR thresholds.

For some small sample diseases such as “diabetes with coma”, because the number of cases is very small in the whole data set, the “unconscious” 2 items support value is relatively low, that is, 0.0069, However, the confidence value is 1, and the symptom “unconsciousness” is a more important feature attribute in the disease classifier. Therefore, in order to better accomplish the multi-disease classification task, the minimum support set threshold of the two frequent set mining algorithms in this paper The default setting is 0.001.

4.2.2. RBSBagging Algorithm

The base learner we chose is CART [3] with an iteration convergence threshold set to 0.0001. In the choice of IR, because different IR threshold settings have different effects on the multi-classification performance of unbalanced samples, we usea preliminary experiment tofind the IRthreshold parameters that are optimal for the classification of multiple disease imbalance samples. The results show that when the IR value is 3, the macro F1 value of the RBSBagging algorithm works best, so the IR threshold is set to 3 by default. The experimental results are shown in Fig. 3.

5. Results and discussions

To verify the robustness of algorithms, we also experimented with the unpublished diabetes dataset and the UCI machine learning database[1]. For the feature selection algorithm, we chose the cmc dataset (contraceptive method choice), monks-1, monks-2 dataset, spect-heartand and spectf-heart datasets. The monks-1

[1] http://www.ics.uci.edu/mlearn/-MLRepository.html.

Table 3
Parameter list of feature selection data sets.

Dataset	Cases	Features	Categories
monks-1	432	7	2
monks-2	432	7	2
cmc	1473	9	3
spect-heart	267	22	2
spectf-heart	267	44	2
ours	5104	130	10

Table 4
Detailed parameter list of the sample unbalanced data sets.

Dataset	Cases	Features	Categories	Category distribution(%)
spect	267	22	2	20/80
ionosphere	351	34	2	36/64
pima	768	8	2	35/65
cmc	1473	9	3	22/35/43
car	1728	6	4	4/4/22/70
glass	214	9	6	4/6/8/14/33/35
ours	5104	130	10	1/1/1/2/2/3/6/10/22/52

Table 5
The average value of the corresponding results of the ten-fold cross-validation experiment.

Datasets	Classification accuracy(%)			
	CART	ReliefF	RFE-SVM	ARFS
monks-1	85.65	85.65	93.09	93.09
monks-2	85.65	67.08	67.08	67.08
cmc	42.20	42.20	42.25	43.76
spect-heart	72.19	76.68	78.33	79.42
spectf-heart	69.32	75.97	77.14	76.14
Ours	80.55	80.92	81.40	83.46
Average	68.80	71.45	73.27	73.83

Table 6
Evaluation results of different indicators of different classification algorithms on an imbalanced sample set.

Datasets	Macro-Precision		
	SMOTE-boost	SMOTE-RF	RBSBagging
spect	**0.536**	0.511	0.531
ionosphere	0.915	**0.918**	0.912
pima	0.692	0.680	**0.707**
cmc	0.378	0.397	**0.421**
car	0.963	0.972	**0.985**
glass	**0.754**	0.739	0.749
Ours	0.733	0.721	**0.781**
Average	0.710	0.705	**0.727**
Datasets	Macro-Recall		
	SMOTE-boost	SMOTE-RF	RBSBagging
spect	0.545	0.530	**0.563**
ionosphere	**0.849**	0.847	0.832
pima	0.604	0.621	**0.681**
cmc	0.629	**0.658**	0.637
car	**0.979**	0.966	0.974
glass	0.704	0.733	**0.757**
Ours	0.753	0.767	**0.799**
Average	0.723	0.732	**0.749**
Datasets	Macro-F1		
	SMOTE-boost	SMOTE-RF	RBSBagging
spect	0.540	0.520	**0.547**
ionosphere	**0.881**	**0.881**	0.870
pima	0.645	0.649	**0.694**
cmc	0.472	0.495	**0.507**
car	0.971	0.969	**0.979**
glass	0.728	0.736	**0.753**
Ours	0.743	0.743	**0.790**
Average	0.711	0.713	**0.734**

and monks-2 data sets only use the test set; the spect-heart and spectf-heart data sets combine the test set and the training set for training. The specific parameters are shown in Table 3. For the classification algorithm, the specific parameters of the selected data set are shown in Table 4

5.1. Feature selection results

We experimented with ten-fold cross validation, which randomly divided the data set into 10 parts: $S_1, S_2, \ldots, S_{10}$, where S_i was used as the test set in the $i-th$ iteration, and the remaining subsets were used. The training classification algorithm, the mean of 10 results is used as the estimation of the accuracy of the algorithm. The classification algorithm uses the CART decisiontreeclassificationalgorithmuniformly. Theexperimental result in Table 5 is the average value of the corresponding results of the ten-fold cross-validation experiment.

According to the results in Table 5, except for the spectf-heart dataset, the classification accuracy of the ARFS algorithm is slightly worse than that of the RFE-SVM [15] feature selection algorithm. On the performance of the remaining datasets, the ARFS algorithm proposed by us has classification accuracy. Both are superior to the other three comparative experimental algorithms. The average accuracy of the classification accuracy of the ARFS feature selection algorithm on the 5 sets of public datasets and diabetes datasets is 73.83%, and the effect is optimal, indicating that the ARFS algorithm can improve the classification accuracy. The proposed feature selection algorithm ARFS based on association rules reduces the scale of feature subsets as much as possible without reducing the classification accuracy.

5.2. Classification results

For metrics, macro precision, macro-recall and macro-F1 are used to evaluate the classification effect of RBSBagging algorithm and SMOTE-BOOST algorithm and SMOTE-RF algorithm on unbalanced samples. The SMOTE-BOOST algorithm is a combination of the SMOTE algorithm and the AdaBoost algorithm. The SMOTE-RF algorithm is a combination of the SMOTE algorithm and the Random Forest algorithm. The experimental results are shown in Table 6. Our experiment uses the same ten-fold cross-validation method.

Table 6 Comparison of the evaluation results of different indicators for different classification algorithms on the unbalanced sample set. For the macro-precision index, the proposed RBSBagging algorithm has a mean-precision average of 0.727 on 6 public datasets and diabetes datasets, which is better than the SMOTE-BOOST algorithm and the SMOTE-RF algorithm, with an average height of about 2 Percentage points. For the macro-recall index, the RBSBagging algorithm has smaller macro-recall values than the SMOTE-BOOST and SMOTE-RF algorithms on the two public datasets of cmc and vehicle, respectively, and the remaining datasets include the macro-recall results of the diabetes dataset. Compared with the two comparison experiments, the average value of macro-recall is 0.749, and the effect is optimal, which is nearly 2 percentage points higher than the other two comparative experimental algorithms. For the macro-F1 index, the RBSBagging algorithm has the best macro-F1 values on the six public datasets and the diabetes dataset, and the average value of 0.734 is also the optimal value, which is 2 times higher than the other two comparative experimental algorithms. Percentage points.

6. Conclusion

First, we propose a feature selection algorithm based on association rules in view of the shortcomings of feature attribute dimension disasters in diagnosis decision support systems. Compared with CART, ReliefF, RFE-SVM and ARFS algorithm proposed in the diabetes dataset and UCI public dataset, the experimental results show that the ARFS algorithm is superior to the contrast experimental algorithm in feature dimension and classification accuracy. Secondly, based on the sample imbalance problem in medical samples, an integrated classification algorithm based on random equalization sampling is designed and implemented. The algorithm also performs three sets of comparative experiments on the private diabetes dataset and UCI public dataset. The comparison algorithms are SMOTE-Boost and SMOTE-RF respectively. The experimental results show that the RBSBagging algorithm proposed in this paper is in macro precision and macro. Both the recall rate and the macro F1 value are superior to the comparison experiment algorithm.

Declaration of Competing Interest

The authors declared that they have no conflicts of interest to this work. We declare that we do not have any commercial or associative interest that represents a conflict of interest in connection with the work submitted.

Acknowledgments

This work is supported by the National Key Research and Development Project (Nos. 2019YFB2101600) and the National Natural Science Foundation of China (Nos. 61763031).

References

[1] R. Agrawal, T. Imieliński, A. Swami, Mining association rules between sets of items in large databases, in: Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, 1993, pp. 207–216.
[2] E.S. Berner, Clinical decision support systems, 233, Springer, 2007.
[3] L. Breiman, Classification and regression trees, Routledge, 2017.
[4] N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, Smote: synthetic minority over-sampling technique, Journal of artificial intelligence research 16 (2002) 321–357.
[5] L.P. CHEN X, S.Y.Z, Research on disease prediction models based on imbalanced medical data sets, Chinese Journal of Computers 3 (2019) 596—609.
[6] L.H. Chiang, R.J. Pell, Genetic algorithms combined with discriminant analysis for key variable identification, J. Process Control 14 (2) (2004) 143–155.
[7] M. Dash, H. Liu, Feature selection for classification, Intell. Data Anal. 1 (3) (1997) 131–156.
[8] Z. Ding, Diversified ensemble classifiers for highly imbalanced data learning and their application in bioinformatics(2011).
[9] I. Guyon, J. Weston, S. Barnhill, V. Vapnik, Gene selection for cancer classification using support vector machines, Mach. Learn. 46 (1–3) (2002) 389–422.
[10] W.H. Hsu, Genetic wrappers for feature selection in decision tree induction and variable ordering in bayesian network structure learning, Inf Sci (Ny) 163 (1–3) (2004) 103–122.
[11] X. Huang, Y. Fang, M. Lu, Y. Yao, M. Li, An annotation model on end-to-end chest radiology reports, IEEE Access 7 (2019) 65757–65765.
[12] X. Jiang, Y.-D. Zhang, Chinese sign language fingerspelling via six-layer convolutional neural network with leaky rectified linear units for therapy and rehabilitation, J. Med. Imaging Health Inform. 9 (9) (2019) 2031–2090.
[13] G.H. John, R. Kohavi, K. Pfleger, Irrelevant Features and the Subset Selection Problem, in: Machine Learning Proceedings 1994, Elsevier, 1994, pp. 121–129.
[14] K. Kermanidis, M. Maragoudakis, N. Fakotakis, G. Kokkinakis, Learning greek verb complements: addressing the class imbalance, in: Proceedings of the 20th international conference on Computational Linguistics, Association for Computational Linguistics, 2004, p. 1065.
[15] K. Kira, L.A. Rendell, A Practical Approach to Feature Selection, in: Machine Learning Proceedings 1992, Elsevier, 1992, pp. 249–256.
[16] K. Kira, L.A. Rendell, et al., The feature selection problem: Traditional methods and a new algorithm, in: Aaai, 2, 1992, pp. 129–134.
[17] D. Koller, M. Sahami, Toward optimal feature selection, Technical Report, Stanford InfoLab, 1996.
[18] I. Kononenko, Estimating attributes: analysis and extensions of relief, in: European Conference on Machine Learning, Springer, 1994, pp. 171–182.
[19] P. Langley, et al., Selection of relevant features in machine learning, in: Proceedings of the AAAI Fall Symposium on Relevance, 184, 1994, pp. 245–271.
[20] Z. Li, S.-H. Wang, R.-R. Fan, G. Cao, Y.-D. Zhang, T. Guo, Teeth category classification via seven-layer deep convolutional neural network with max pooling and global average pooling, Int. J. Imaging Syst. Tech.nol 29 (4) (2019) 577–583.
[21] G. Liang, A.G. Cohn, An effective approach for imbalanced classification: unevenly balanced bagging, in: Twenty-Seventh AAAI Conference on Artificial Intelligence, 2013.
[22] H. Liu, R. Setiono, et al., A probabilistic approach to feature selection-a filter solution, in: ICML, 96, Citeseer, 1996, pp. 319–327.
[23] J.R. Quinlan, Induction of decision trees, Mach. Learn. 1 (1) (1986) 81–106.
[24] J.J. Rodríguez, J.F. Díez-Pastor, C. García-Osorio, P. Santos, Using model trees and their ensembles for imbalanced data, in: Conference of the Spanish Association for Artificial Intelligence, Springer, 2011, pp. 94–103.
[25] Z. Sun, Q. Song, X. Zhu, H. Sun, B. Xu, Y. Zhou, A novel ensemble method for classifying imbalanced data, Pattern Recognit 48 (5) (2015) 1623–1637.
[26] P. Thanathamathee, C. Lursinsap, Handling imbalanced data sets with synthetic boundary data generation using bootstrap re-sampling and adaboost techniques, Pattern Recognit. Lett. 34 (12) (2013) 1339–1347.
[27] A. Tsymbal, S. Puuronen, D.W. Patterson, Ensemble feature selection with the simple bayesian classification, Information fusion 4 (2) (2003) 87–100.
[28] S.-H. Wang, C. Tang, J. Sun, J. Yang, C. Huang, P. Phillips, Y.-D. Zhang, Multiple sclerosis identification by 14-layer convolutional neural network with batch normalization, dropout, and stochastic pooling, Front. Neurosci. 12 (2018) 818.
[29] H. Yu, J. Ni, An improved ensemble learning method for classifying high-dimensional and imbalanced biomedicine data, IEEE/ACM Trans. Comput. Biol. Bioinf. 11 (4) (2014) 657–666.
[30] X. Zhang, D. Ma, L. Gan, S. Jiang, G. Agam, Cgmos: Certainty guided minority oversampling, in: Proceedings of the 25th ACM International on Conference on Information and Knowledge Management, 2016, pp. 1623–1631.
[31] Y.-D. Zhang, V.V. Govindaraj, C. Tang, W. Zhu, J. Sun, High performance multiple sclerosis classification by data augmentation and alexnet transfer learning model, J. Med. Imaging Health Inform. 9 (9) (2019) 2012–2021.
[32] Y.-D. Zhang, C. Pan, J. Sun, C. Tang, Multiple sclerosis identification by convolutional neural network with dropout and parametric relu, J. Comput. Sci. 28 (2018) 1–10.