Sample Reduction Algorithm Based on Classification Contribution

Zheng Chai, Yanying Li*, Jiaoni Zhang, Xialin Wang, Wen Li, Yucong Jiang

***Abstract*—The KNN algorithm takes exponentially growth of time to process dataset containing a large number of samples and has low classification performance. To address this problem, this paper proposes a sample reduction method based on classification contribution ranking (SRCCR). First, SRCCR performs a denoising process to expand the smoothing decision boundary by removing the noise sample in the initial training dataset; next, the denoised samples are sorted in ascending order according to the classification contribution strategy; finally, representative boundary samples and center samples are selected based on the local set to form the final subset. SRCCR reduces storage requirement and execution time, and significantly improves the classification performance of the KNN algorithm. To verify the effectiveness of the proposed method, we conduct comparative experiments on 31 real datasets from the UCI and KEEL databases. Compared with several classical instance selection algorithms, the proposed SRCCR algorithm has advantages in terms of accuracy and reduction rate. The results of the study on the two-dimensional dataset ”Banana” show that the SRCCR algorithm not only selects more representative boundary and center samples, but preserves the distribution of the original dataset.**

***Index Terms*—Sample reduction, K nearest neighbor, Classification contribution, local set**

I. Introduction

WIth the explosion of data, a large number of imbalanced datasets are flooding into the field of machine learning and data mining, which makes it difficult for models to handle them. To solve this problem, scholars have proposed many learning algorithms, among which a well-known example of sample-based classification algorithm is the K-nearest neighbor rule (KNN) [1] (cover and hart1967). KNN is comparing the query sample with each sample in the training set, so the training process requires a huge amount of storage space to store the entire training instances. As with KNN, case-based learning algorithms are widely used in pattern recognition, image processing, and data mining, but they still face problems such as high storage requirement, low predictive performance, and noise sensitivity. Therefore, sample reduction [2-3] is introduced to better deal with the imbalanced problem. As a popular data processing method, it removes noise samples from the dataset to improve the classification efficiency, and maintains the classification accuracy.

The sample reduction technique is used as pre-processing step for the dataset. The sample reduction technique finds the best subset S by filtering out redundant and noise samples from the original dataset T. It aims to reduce the size of the training set, allowing the runtime and prediction accuracy of the classifier on a subset to be improved. Due to these advantages and features, sample reduction techniques have attracted a lot of attention [4-6]. Reduction technique is divided into three types: edition method, condensation method and hybrid method. The purpose of the edition method is to eliminate the noise instances and the condensation method to search the consistent subset of the training set. By eliminating the instances in the training set that do not affect the classification accuracy of the whole training set, a consistent subset is formed. The hybrid method combines the features of the edition and condensation methods while eliminating noise and redundant samples.

The Edit nearest neighbor method (ENN) (Wilson, 1972) [11] is a classical sample reduction method. As the simplest example of the edition method, ENN removes those points that are not correctly classified by its K nearest neighbors. ENN effectively removes the noise samples from the dataset and smoothes the decision boundary. Famous ENN variants include Multiple edition [12] and Nearest Centroid Neighbor (NCNEdit) [13]. The classification performance of these edition methods is mainly affected by two problems: parameter dependence and noise sensitivity.

Another classic algorithm of sample reduction technology is condensed nearest neighbor method (CNN) [7] that initially attempts to reduce the number of samples in the training set. By removing instances using the condensed nearest neighbor rule, a consistent subset is obtained that does not affect the performance of the training set. Since the goal of the CNN algorithm is to retain boundary samples, the algorithm is very sensitive to noise samples and is susceptible to the order of samples in the dataset. Most of the condensation methods try to improve the effect of CNN algorithm by generating smaller consistent subset. For example, Reduced nearest neighbor algorithm (RNN) [8], Minimum

Manuscript received January 12, 2023; revised June 4, 2023. This work was supported by Natural Science Basic Research Program of Shaanxi (2023-JC-YB-632), the National Natural Science Foundation of China (61573266), the Foundation of the Doctoral Scientific Research of Baoji University of Arts and Sciences (209040122), the Foundation of Baoji City Guidance plan for scientific and technological innovation (2017JH2-23), the Scientific Research Program Funded by Shaanxi Provincial Education Department(18JK0045), the Horizontal project (2023-KJHX002, 2021-KJHX033), the Special project of philosophy and Social Sciences in Baoji (BJSKZX-202314).

Zheng Chai is Postgraduate of the School of Mathematics and Information Science, Baoji University of Arts and Sciences, Baoji 721013, Shaanxi,China. He is Lecturer of Public Course Teaching Department,Jiangxi University of Engineering. (e-mail:15735159302@163.com).

Yanying Li is Associate Professor of the School of Mathematics and Information Science, Baoji University of Arts and Sciences, Baoji 721013, Shaanxi, China. (Corresponding author, e-mail: liyanying2021@163.com).

Jiaoni Zhang is Postgraduate of the School of Mathematics and Information Science, Baoji University of Arts and Sciences, Baoji 721013, Shaanxi, China. (e-mails: 2108651612@qq.com).

Xialin Wang is Postgraduate of the School of Mathematics and Information Science, Baoji University of Arts and Sciences, Baoji 721013, Shaanxi, China. (e-mails: 939203103@qq.com).

Wen Li is Postgraduate of the School of Mathematics and Information Science, Baoji University of Arts and Sciences, Baoji 721013, Shaanxi, China. (e-mails: 2388088943@qq.com).

Yucong Jiang is Postgraduate of the School of Mathematics and Information Science, Baoji University of Arts and Sciences, Baoji 721013, Shaanxi, China. (e-mails: 1253294705@qq.com).

consistent set algorithm (MCS) [9] and Fast nearest neighbor condensed algorithm (FCNN) [10]. In fact, the condensation method greatly reduces the number of instances and rarely changes the prediction accuracy, while the edition method moderately reduces the number of instances and may change the prediction accuracy to a great extent. In recent years, the hybrid method [14-22], which combines the advantages of edition method and condensation method, has attracted extensive attention. In our work, we mainly focus on hybrid method.

Hybrid methods typically use edition method to eliminate noise, and then use condensation method to retain representative sample. Classical hybrid methods include Class Boundary Preserving algorithm (CBP) [14] and the Adaptive Threshold-based Instance Selection algorithm (ATISA) [16]. Many modified hybrid methods [15,18-22,27-30] have also been developed by using new models, innovative theories and optimization algorithms. However, most existing hybrid methods [14-18] rely heavily on parameters and are relatively time-consuming, leading to unstable performance. Although parameter-free hybrid methods have been proposed [19,20], it is still difficult to achieve high accuracy and high simplicity at the same time.

To solve the above issues, we propose a new hybrid algorithm Sample Reduction based on Classification Contribution Ranking (SRCCR), which introduces the concept of classification contribution to assign scores for samples. We differentiate between the center and boundary samples based on the scores. The higher the score, the closer the sample is to the center of the sample; the lower the score, the closer the sample is to the border.

SRCCR consists of three steps: the first step is to filter noise sample: ENN is used to remove the noise samples from the training set. The second step is to assign score to sample: each sample is assigned a score based on the classification contribution of the current sample to the rest of the training samples, and then the samples are ranked according to the score. The third step is to eliminate redundant samples. The redundant samples are removed based on the local set strategy to obtain the final subset. The experiments show that the proposed method effectively reduces the number of samples and significantly improves the prediction accuracy. In our experiment, we compare SRCCR with KNN and two classical sample reduction methods on 31 real datasets. The experimental results show that SRCCR has better performance in improving the prediction accuracy and reduction rate of KNN, which is better than the classical hybrid method. The main contributions of this paper are as follows:

(a) We propose a new ranking-based sample reduction method, SRCCR. Experimental results show that SRCCR is very competitive in terms of classification performance and reduction rate.

(b) SRCCR scores each sample based its classification contribution, and the score is used to distinguish whether the sample is near the center or the boundary.

(c) Based on the concept of local set, we remove the redundant sample so that the reduced sample distribution is as consistent as possible with the original dataset.

In the rest of our paper, the arrangement is as follows. The related work is listed in Section II, Section III is basic definition, Section IV is the proposed method SRCCR, Section V is experiment results and discussion, Section VI is conclusion.

II. Related work

In this section, we describe some hybrid methods in detail. These hybrid methods first denoise the initial training sample set, then partition the dataset with tricks to divide the training samples into boundary samples and center samples, and finally select the final subset based on certain preferences. CBP [14] is a typical hybrid method based on segmentation. Firstly, the training set is denoised using ENN; secondly, the concept of reachable set is used to distinguish between boundary and non-boundary samples; finally representative boundary samples are selected to form the final subset. Spectral Instance Reduction (SIR) [15] utilizes spectrogram theory to distinguish between boundary and interior samples. ATISA [16] is another hybrid method based on partitioning that uses ENN to filter noise and defines a threshold for each sample to divide the dataset based on this threshold. Among them, ATISA2 focuses on retaining the center samples, ATISA3 focuses on retaining the boundary samples. In addition, scholars have proposed the Local Set of Instance Reduction (LSIR) [17] and Binary Nearest Neighbor Tree (BNNT) based sample selection algorithms [18]. LSCO selects the center sample by the idea of clustering, while LSBO focuses on the selection of boundary samples. BNNT establishes a binary nearest neighbor tree to evaluate the status of node samples in the tree. When the binary tree is located in a class, all internal samples in the tree are replaced by newly generated samples; when the binary tree is located at the boundary of different classes, the connected samples of different classes are retained, and finally all the retained samples are used as the classification of new test samples. Recently, two sample reduction methods based on natural neighbors have been proposed, constrained nearest neighbor based sample reduction (CNNIR) [19] and non-parametric hybrid sample selection algorithm based on local set of natural neighbors (LSNANIS) [20]. CNNIR uses constrained nearest neighbor chains to find boundary samples and natural neighbors to find core samples. LSNANIS consists of three sub-algorithms LSEditing, LSBorder and LSCore. LSEditing is used to eliminate internal noise and smooth class boundary. LSBorder divides the samples of the denoised dataset into center samples and representative boundary samples. LSCore is used to compress the center samples. Finally, the compressed center samples and representative boundary samples are combined into the final subset. In the process of generating the final subset, the above two algorithms select center and boundary samples at the same time.

In addition, there are some hybrid algorithms based on ranking, such as the sample selection algorithm based on boundary ranking (IRB) [21] and the sorting-based classification sample selection algorithm (RIS) [22]. IRB uses the similarity measure based on the nearest enemy and sorts the samples by the size of the similarity measure. In contrast, RIS calculates score based on the relationship between each sample and other samples, and ranks the samples according to the scores.

However, the existing hybrid methods [14-22] still have some drawbacks, such as parameter dependence, high time complexity, and difficulty in achieving high accuracy and high reduction rate simultaneously.

III. Basic definition

Let $T = \{x_i \in R^d\}_{i=1}^N$ be a training set with N training samples in d-dimensional feature space. There are m class labels, each sample $x_i = (x_{i1}, x_{i2}, \cdots, x_{id})$ corresponds to its class label y_i, where $y_i \in \{c_1, c_2, \cdots, c_m\}$. Let l: $x \to C$ be a function that maps sample x_i to its class label $l(x_i)$.

For the simplicity, the following concepts and terms are used in this paper:

- T: Training set,
- $NN_K(x_i, T)$: The K nearest neighbors of x_i in T;
- $d(x_i, x_j)$: Euclidean distance of samples x_i and x_j;
- $NE(x_i, T)$: The nearest neighbor between x_i and other classes in T;
- $r(x_i) = d(x_i, NE(x_i, T))$: The radius of x_i in set T;
- $s(x_i)$: Classification contribution score of sample x_i;
- $IS(x_i, T)$: Sample classification contribution scoring function, return the classification contribution score of sample x_i;
- R: Final subset after reduction;
- $L(T)$: Class label set of dataset T.

A. Sample classification contribution definition

Inspired by the training dataset cleaning method [23] based on the classification ability ranking of KNN classifier, we describe the location distribution of samples in the dataset by their classification contribution to other samples in the dataset.

For a given initial training set T and selected sample x_i, we obtain the classification contribution score $s(x_i)$ of sample x_i by the following strategy.

The sample x_j ($x_j \neq x_i$) in T is recorded as the test sample, and the rest of the sample set T^* is recorded as the training set. The KNN algorithm is used to predict the class label of sample x_j. This process is repeated $N-1$ times, and five cases occur in each classification process:

case 1: $x_i \notin NN_K(x_j, T)$;
case 2: $x_i \in NN_K(x_j, T)$, $y_i \neq l(x_j) \wedge y_i \neq y_j$;
case 3: $x_i \in NN_K(x_j, T)$, $y_i \neq l(x_j) \wedge y_i = y_j$;
case 4: $x_i \in NN_K(x_j, T)$, $y_i = l(x_j) \wedge y_i \neq y_j$;
case 5: $x_i \in NN_K(x_j, T)$, $y_i = l(x_j) \wedge y_i = y_j$.

It can be seen that in case 1, sample x_i does not belong to the nearest neighbors of x_j, and we consider that sample x_i does not contribute to the classification result of x_j. In case 2, x_i has confusing effect on the correct classification of x_j. Although x_i is one of the K nearest neighbors of x_j, it has a different class label than most of the samples in the K nearest neighbors of x_j. In cases 3 and 5, x_i has positive effect on the correct classification of x_j. This is because x_i is not only a neighbor of x_j, but the class label is the same. In particular, in case 5, the class to which x_i belongs is the more dominant class among the K nearest neighbors of x_j, which plays a decisive role in the correct classification of x_j. On the contrary, in case 4, x_i plays a negative role in the classification of x_j. x_i is the majority of class samples in the nearest neighbors of x_j, and their different class labels directly lead to the misclassification of x_j.

The classification ability of x_i depends on its contribution to the correct classification of x_j. The higher the contribution, the higher the score. To quantify the classification contribution of the sample x_i, we set five scores in the above five different cases. In $i = 1, 2, \cdots, 5$, the scores of the corresponding cases are denoted by a_i, $a_1 = 0$, $a_2 = 0.2$, $a_3 = 0.4$, $a_4 = -0.6$, and $a_5 = 1$.

In the $N-1$ classification process, the number of occurrence of different cases is recorded and denoted by n_1, n_2, n_3, n_4 and n_5 respectively. It is obvious that $n_1+n_2+n_3+n_4+n_5 = N-1$. Finally, the classification contribution score of sample x_i is $s(x_i)$, which can be obtained from Eq. 1.

$$s(x_i) = \begin{cases} \frac{n_2 \times a_2 + n_3 \times a_3 + n_4 \times a_4 + n_5 \times a_5}{n_2 + n_3 + n_4 + n_5}, & n_1 \neq N-1, \\ 0, & n_1 = N-1. \end{cases} \tag{1}$$

The classification contribution score of the sample shows that the higher score of the sample, the more frequently case 5 appears. This means that the higher score, the sample is closer to the center of the class. The lower score of the sample, the more frequently case 4 appears. This means that the sample is closer to the boundary of the class. The pseudo code of sample classification contribution scoring algorithm is shown in Algorithm 1.

Algorithm 1: Sample classification contribution scoring algorithm.

```
Input:
  training dataset $T$, Sample to be graded $x_i$ ($x_i \in T$),
  the neighborhood size $K$.
Output:
  $s(x_i)$: Classification contribution score of sample $x_i$.
1. for $1 \leq l \leq 5$ do
     $n_l = 0$
   end for
2. for $x_j \in T, x_j \neq x_i$ do
     find $NN_K(x_j, T)$ and $l(x_j) = Class_{max}(NN_K(x_j, T))$
     if case 1 is satisfied, then $n_1 = n_1 + 1$
     else if case 2 is satisfied, then $n_2 = n_2 + 1$
     else if case 3 is satisfied, then $n_3 = n_3 + 1$
     else if case 4 is satisfied, then $n_4 = n_4 + 1$
     else if case 5 is satisfied, then $n_5 = n_5 + 1$
   end for
3. if $n_1 = N - 1$
     $s(x_i) = 0$
   else if
 $s(x_i) = \begin{cases} \frac{n_2 \times a_2 + n_3 \times a_3 + n_4 \times a_4 + n_5 \times a_5}{n_2 + n_3 + n_4 + n_5}, & n_1 \neq N-1, \\ 0, & n_1 = N-1. \end{cases}$
```

IV. The proposed SRCCR method

A. motivation

After analyzing many different sample reduction algorithms, we realize that most algorithms focus on selecting boundary samples and that samples near the boundary have much information to describe the decision boundary accurately.In contrast, center samples far from the class boundary have less impact on classification accuracy, but retaining proper core samples also improve classification accuracy. In addition, noise samples near the class boundary have significant impact on the classification performance of the algorithm. Therefore, our proposed SRCCR algorithm removes a great number of noise and redundant points through three steps: noise filtering, contribution ranking of samples, and selection of the final subset, while retaining representative boundary samples and small number of center samples. The pseudo code of SRCCR is detailed in Algorithm 2.

Algorithm 2: The proposed SRCCR method

```
Input:
  $T$: training dataset.
  $K$: the number of nearest neighbors.
Output:
  $R$: The reduced final subset.
Initialization: $T^* = T, R = \emptyset$.
1. for $x_i \in T$ do
     if $y_i \neq Class_{max}(NN_3(x_i, T))$ then
       $T^* = T^* \setminus x_i$
     end if
   for
2. for $x_i \in T^*$ do
     $r(x_i) = d(x_i, NE(x_i, T^*))$
     $s(x_i) = IS(x_i, T^*)$
   end for
   $T^* = sortAscByS(T^*)$
3. for $x_i \in T^*$ d0
     $x_j = NN_1(x_i, R)$
     if $y_i \notin L(R)$ or $d(x_i, x_j) > r(x_j)$
       $R = R \cup x_i$
     end if
   end for
```

The presence of noise samples significantly weakens the classification performance of algorithm, so SRCCR uses ENN to denoise the sample set in the first step. Existing papers show that ENN works best when the K value is set to 3. Therefore, in this step, we also set the parameter K to 3. The second step is to calculate the radius and classification contribution score for each sample in the dataset T respectively, and then use $T^* = sortAscByS(T^*)$ to sort the samples in T in descending order by score. The third step is to selectively retain the sorted samples. If the class label of sample x_i is not in the final subset R or the distance between x_i and the nearest $NN_1(x_i, R)$ in R is greater than the radius of $NN_1(x_i, R)$, then x_i is added to the final subset R.

V. Experiments

A. Data information

In this section, we briefly present information about the datasets that are used in the experiments. The 31 real datasets are collected from the UCI database [24] and the Keel database [25]. All data information, including sample size, number of attributes and number of classes, is displayed in Table I. In these datasets, the maximum and minimum sample sizes are 19020 and 106, respectively; the maximum and minimum attribute sizes are 41 and 2, respectively. Among them, there are 13 datasets with sample sizes larger than 1000.

B. Experiments on real datasets

The experiment is divided into two parts. The first part shows the execution results of SRCCR intuitively on the two-dimensional data ‘Banana’; in the second part, we compare SRCCR algorithm with several classical algorithms on the above datasets to verify the performance of SRCCR. All algorithms are written in Python and use Euclidean distance norm as distance measure. Using the 10-fold cross-validation scheme, we divide the dataset into 10 subsets. Nine subsets are used as training sets to train the model, and the remaining one subset is used as test set to test the performance of the model. For each algorithm, we perform 5 times 10-fold cross-validation on each dataset shown in Table I and use the average of the classification accuracy, reduction rate (Re) obtained from these 5 times experiments to evaluate the performance of algorithms.

TABLE I: The details of datasets.

Data	Samples	Attributes	Classes
Appendicitis	106	7	2
Balance	625	4	2
Banna	5300	2	2
BanKnote	1372	4	2
Biopsy	683	9	2
Breast_caner	569	30	2
Bupa	347	6	2
Column	310	6	2
Contraceptive	1473	9	3
Forestypes	523	27	4
German	1000	24	2
Haberman	306	3	2
Heart_failure	299	12	2
Indian_liver	579	10	2
Ionosphere	351	34	2
Messidor_features	1151	19	2
Pima	532	7	2
QsaK	1055	41	2
Satimage	6435	36	7
Thyroid	7200	21	3
Titanic	2201	3	2
Transfusion	748	4	2
Wirelsee	2000	7	4
Yeast	1334	8	5
Plrx	180	18	2
Phoneme	5404	4	2
Mammographic	827	5	2
Heart	462	9	2
Saheart	270	13	2
Page_blocks	5473	10	5
Magic	19020	10	2

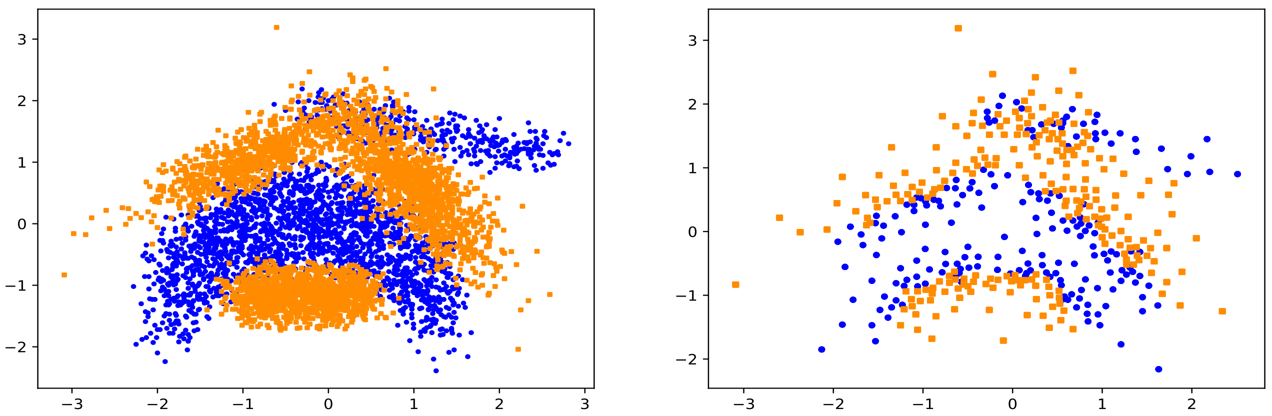
Sample reduction is actually a double objective optimization problem. When we compare the performance of the two algorithms, we need to pay attention to all the objectives. To verify the effectiveness of the algorithm, this paper uses two metrics to evaluate classification performance of the algorithms.

$$Re = 100 \times (1 - \frac{|R|}{|T|}) \quad (2)$$

where $|R|$ is the number of samples of the final subset after reduction, $|T|$ is the number of samples of original dataset.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}. \quad (3)$$

where TP is true positive, TN is true negative, FP is false positive and FN is false negative.



1 Banana

2 The reduced set of Banana

Fig. 1: The result of SRCCR on the Banana.

C. *Experimental results and analysis*

To visually demonstrate the reduction effect of the SRCCR algorithm, we use a two-dimensional dataset ”Banana” from the KEEL dataset, as shown in Fig. 1. This data contains two classes and 5300 samples. From the left figure it is seen that many samples are overlapping. From the right figure it is seen that the SRCCR algorithm removes a major number of redundant samples and retains small number of samples. We find that more samples located at the decision boundary are retained, and these samples are important samples on the decision boundary. As can be seen from the Fig. 1, as the noise samples are removed, the decision boundary becomes smoother. Although the SRCCR algorithm retains small number of samples, these samples perfectly outline the distribution of original dataset.

The classification accuracy and reduction ratio of SRCCR algorithm on ’Banana’ dataset are listed in Table II and Table III. In terms of classification accuracy achieved on ‘Banana’ dataset, SRCCR (90%) is higher than CNN (86.71%), ATISA (88.96%) and KNN (88.48%); in terms of reduction ratio, SRCCR (90.42%) is also higher than by ATISA (89.55%) and CNN (83.5%).

The experimental results are shown in Table II and Table III. In order to evaluate whether the differences in classification accuracy, reduction rates obtained by these algorithms are significant, we use the nonparametric Wilcoxon signed-rank test with a confidence of 0.05 [26]. In order to facilitate reading, this chapter uses ‘ + ’, ‘ - ’, ‘ $\sim$ ’ symbols to indicate that SRCCR algorithm is significantly better, worse or equivalent than other algorithms.

TABLE II: The classification accuracy rates(%) of every algorithm.

Data	KNN	CNN	ATISA	SRCCR
Appendicitis	83.95	80.18	79.32	**87.51**
Balance	80.96	71.39	84.16	**85.19**
Banana	88.48	86.71	88.96	**90.00**
BanKnote	**99.94**	69.70	98.45	99.21
Biopsy	**96.70**	95.25	96.44	96.50
Breast_caner	92.68	92.23	94.36	**96.24**
Bupa	63.15	59.17	60.28	**67.00**
Column	81.48	65.90	72.29	**82.13**
Contraceptive	49.45	43.16	51.53	**52.39**
Forestypes	88.77	79.80	87.12	**89.14**
German	68.63	62.96	69.20	**71.28**
Haberman	70.16	65.83	71.34	**74.70**
Heart_failure	58.89	63.67	64.78	**66.92**
Indian_liver	65.90	64.59	63.89	**68.28**
Ionosphere	84.70	81.46	83.97	**84.96**
Messidor_features	63.95	58.27	61.87	**65.86**
Pima	73.64	70.20	72.45	**75.07**
QSAR	82.16	70.46	**82.47**	80.80
Satimage	**91.08**	87.76	90.05	90.00
Thyroid	93.87	90.07	92.68	**93.92**
Titanic	67.70	57.73	59.78	**70.05**
Transfusion	74.57	61.93	75.34	**76.28**
Wirelsee	98.37	71.73	97.60	**98.86**
Yeast	57.44	36.40	56.70	**60.72**
Plrx	63.99	54.30	63.61	**69.28**
Phoneme	**89.11**	82.79	86.86	85.82
Mammographic	**79.64**	76.40	77.69	78.55
Heart	64.14	60.85	63.40	**65.26**
Saheart	57.08	59.67	**67.01**	66.14
Page_blocks	95.68	94.03	95.53	**95.73**
Magic	80.04	35.16	**83.16**	80.88
Mean	77.61	69.35	77.17	**79.51**
Wilcoxon	$+$	$+$	$+$	N/A

Table II shows the classification accuracy of the KNN of the three reduction algorithms and the original data without reduction, and the best performance obtained on each dataset is displayed in bold. The results show that the classification accuracy of SRCCR algorithm is the best in 23 of 31 datasets, and the average classification effect is also significantly better than that of other methods. According to Wilcoxon test, SRCCR is significantly better than KNN, ATISA and CNN. Table III shows the reduction ratio of the three reduction methods. Although the reduction ratio of SRCCR method is superior to other methods in most data, it can be seen from the Wilcoxon test results that there is no significant difference between them in reduction ratio.To sum up, the SRCCR method can largely reduce the sample size of the

TABLE III: The reduction rates(%) of every algorithm.

Data	CNN	ATISA	SRCCR
Appendicitis	**94.04**	88.90	86.20
Balance	68.95	77.23	**81.89**
Banana	83.50	89.55	**90.42**
BanKnote	**99.76**	94.67	95.58
Biopsy	**94.10**	91.96	92.86
Breast_caner	88.83	78.52	**92.60**
Bupa	**58.26**	53.72	57.12
Column	**93.79**	67.53	69.18
Contraceptive	**81.40**	69.25	62.45
Forestypes	**77.80**	69.67	74.24
German	60.71	64.82	**70.07**
Haberman	62.29	78.43	**80.93**
Heart_failure	63.00	70.60	**75.04**
Indian_liver	62.78	63.60	**70.70**
Ionosphere	**82.27**	73.55	73.75
Messidor_features	58.48	63.85	**66.84**
Pima	61.97	**76.21**	73.79
QsaK	**93.70**	69.08	68.53
Satimage	**86.26**	78.84	81.28
Thyroid	85.94	91.05	**92.38**
Titanic	65.64	99.15	**99.49**
Transfusion	65.76	81.14	**84.82**
Wirelsee	**99.42**	85.99	88.59
Yeast	**91.19**	68.81	59.31
Plrx	54.48	**65.38**	54.54
Phoneme	81.31	79.36	**82.48**
Mammographic	67.27	82.02	**83.77**
Heart	54.23	**68.30**	66.69
Saheart	57.99	71.87	**73.68**
Page_blocks	92.68	93.29	**94.99**
Magic	**99.84**	79.79	78.82
Mean	77.00	76.96	**78.15**
Wilcoxon	∼	∼	N/A

dataset and improve the classification performance. It is a competitive sample reduction method.

TABLE IV: The comparison of classifiers on 31 datasets using Friedman test.

Method	CNN	ATISA	KNN	SRCCR
Mean rank	3.87	2.55	2.23	1.35

Then, significance test is performed using Friedman test for KNN, CNN, ATISA and SRCCR. Based on the classification performance obtained by these methods on each dataset, we sort these algorithms in descending order. After that, the best performing algorithm is recorded as 1, the second algorithm is recorded as 2, if two algorithms have the same classification performance, these two algorithms get equal ranking, and so on. Let R_i^j be the ranking of the j-th algorithm on the i-th dataset, and the mean ranking of j-th algorithm is computed as $R_j = \frac{1}{n}\sum_i R_i^j$. The Friedman test is distributed in accordance χ_F^2 with L-1 degrees of freedom when $n > 10$ and $L \geq 5$.

$$\chi_F^2 = \frac{12n}{L(L+1)}[\sum_j R_j^2 - \frac{L(L+1)^2}{4}]. \tag{4}$$

where n is the number of datasets and L is the number of algorithms.

In order to evaluate the performance of SRCCR, we compare the classification accuracy obtained by SRCCR with the other three algorithms. The average ranking of the obtained results is shown in Table IV. According to Eq. (4), χ_F^2(=60.79). If all algorithms have the same accuracy on each dataset, the average ranking $\overline{R}$ is 2.5. From Table IV, we find that the average ranking $\underline{R_j}$ computed by each algorithm is significantly different from $\overline{R}$ and χ_F^2 is greater than $(\chi_F^2)_{0.01}$(=11.34). Therefore, we obtain that there is a significant difference between these algorithms.

We use Friedman test to prove that there is a significant difference between the proposed method and the other methods, but we need to further distinguish between the algorithms. We test these four methods with the Holm post hoc test, calculated as follows

$$Z = (R_i - R_0)/SE \tag{5}$$

where R_0 and R_i are the average rankings of SRCCR and the i-th method in CNN, ATISA and KNN, respectively. SE is obtained from formula $\sqrt{\frac{L(L+1)}{6n}}(= \sqrt{\frac{4\times(4+1)}{6\times 31}} = 0.327)$. Based on the calculated Z-value, it is compared with α(=0.05). Table V shows the Z-values obtained for each comparison method as well as the p values. Based on the results in the table, we find that the obtained p values are much smaller than the α values. Therefore, we reject the original hypothesis (i.e., there is no significant difference between the proposed method and the other methods). The results indicate that the SRCCR method is superior to the other methods.

TABLE V: The p values are obtained using Holm's post hoc test.

i	Method	$Z = (R_i - R_0)/SE$	p value	$\alpha/(k-i)$
1	CNN	6.047	1.5×10^{-9}	0.0167
2	ATISA	4	6.3×10^{-5}	0.025
3	KNN	3.341	6×10^{-4}	0.05

Finally, according to the classification accuracy obtained by each algorithm, we use T-test to compare SRCCR with CNN, ATISA, and KNN, and the results are shown in Table VI. According to the experimental results, we find that the p value obtained from the pairwise test between the proposed method and other methods is less than 0.01, so the proposed method SRCCR is significantly different from CNN, ATISA, and KNN.

In a word, the proposed SRCCR classification performance is widely verified in real datasets. The results can be summarized as follows:

- In the SRCCR algorithm, the method of dividing the data based on the sample classification contribution

TABLE VI: The pairwise contrasts of SRCCR with other algorithms on 31 datasets using T-test ('yes' indicate the dramatic difference between two algorithms)

Pairwise comparison	p-value	Significant
CNN VS. SRCCR	2.88×10^{-6}	yes
ATISA VS. SRCCR	2.14×10^{-4}	yes
KNN VS. SRCCR	4.02×10^{-4}	yes

identifies well whether the sample is located at the class center or at the boundary.

- The experimental part intuitively shows the reduction effect of the algorithm on two-dimensional data, which shows that SRCCR ensure that the distribution of the reduced subset is consistent with the initial dataset.
- Compared with other algorithms, SRCCR effectively improves the classification effectiveness of the algorithm while ensuring the reduction rate, which indicates that the samples selected by SRCCR contain more information useful for classification and better represent the original dataset.
- Experimental results on 31 real datasets show that SRCCR is competitive in terms of classification performance and reduction rate.

VI. Conclusion

This paper introduces a new sample reduction algorithm SRCCR based on classification contribution ranking. SRCCR mainly includes three steps. The first step is to filter out the noisy samples. The second step is to sort the samples in the training set in descending order according to the classification contribution rate. The third step is to select representative boundary samples and center samples to obtain an optimized final subset. SRCCR significantly improves the prediction accuracy of KNN and reduces the storage requirement. Through experimental analysis, SRCCR achieves high accuracy and high reduction rate to improve the classification performance of KNN compared with the existing classical methods. In future work, we plan to find a suitable sample selection process with deeper filtering of the final subset to further improve the classification performance.

References

[1] Cover T M, Hart P E. Nearest neighbor pattern classification[J]. IEEE Trans.inf.theory, 1953, 13(1):21-27.

[2] Zhai J, Qi J, Zhang S. An instance selection algorithm for fuzzy K-nearest neighbor[J]. Journal of Intelligent and Fuzzy Systems, 2020, 40(1):1-13.

[3] Kim Y, EnKe D. Instance selection using genetic algorithms for an intelligent ensemble trading system[J]. Procedia computer science, 2017, 114: 465-472.

[4] Yang L, Zhu Q, Huang J, et al. Natural neighborhood graph-based instance reduction algorithm without parameters[J]. Applied Soft Computing, 2018, 70: 279-287.

[5] Tsai C F, Lin W C, Hu Y H, et al. Under-sampling class imbalanced datasets by combining clustering analysis and instance selection[J]. Information Sciences, 2019, 477: 47-54.

[6] Pang X, Xu C, Xu Y. Scaling KNN multi-class twin support vector machine via safe instance reduction[J]. Knowledge-Based Systems, 2018, 148: 17-30.

[7] Hart P E. The Condensed Nearest Neighbor Rule[J]. IEEE Transactions on Information Theory, 1968, 14(3):515-516.

[8] Gates G. The reduced nearest neighbor rule (corresp.)[J]. IEEE transactions on information theory, 1972, 18(3): 431-433.

[9] Dasarathy B V. Minimal consistent set (MCS) identification for optimal nearest neighbor decision systems design[J]. Systems Man & Cybernetics IEEE Transactions on, 1994, 24(3):511-517.

[10] Angiulli F. Fast Nearest Neighbor Condensation for Large Data Sets Classification[J]. IEEE Transactions on Knowledge and Data Engineering, 2007, 19(11):1450-1464.

[11] Wilson, Dennis L. Asymptotic Properties of Nearest Neighbor Rules Using Edited Data[J]. IEEE Transactions on Systems Man and Cybernetics, 2007, 2(3):408-421.

[12] Ferri F J, Albert J V, Vidal E. Considerations about sample-size sensitivity of a family of edited nearest-neighbor rules[J]. IEEE Transactions on Systems Man & Cybernetics Part B Cybernetics A Publication of the IEEE Systems Man & Cybernetics Society, 1999, 29(5):667.

[13] Sanchez J S, Barandela R, Marques A I, et al. Analysis of new techniques to obtain quality training sets[J]. Pattern Recognition Letters, 2003, 24(7):1015-1022.

[14] NiKolaidis K, Goulermas J Y, Wu Q H. A class boundary preserving algorithm for data Condensation[J]. Pattern Recognition, 2011, 44(3):704-715.

[15] NiKolaidis K, Rodriguez-Martinez E, Goulermas J Y, et al. Spectral Graph Optimization for Instance Reduction[J]. IEEE Transactions on Neural NetworKs and Learning Systems, 2012, 23(7):1169-1175.

[16] Cavalcanti G, Ren T I, Pereira C L. ATISA: Adaptive Threshold-based Instance Selection Algorithm[J]. Expert Systems with Applications, 2013, 40(17):6894-6900.

[17] Leyva E, Gonzalez A, Perez R. Three new instances selection methods based on local sets: A comparative study with several approaches from bi-objective perspective. 2015.

[18] Li J, Wang Y. A new fast reduction technique based on binary nearest neighbor tree[J]. Neurocomputing, 2015, 149(pt.c):1647-1657.

[19] Yang L, Zhu Q, J Huang, et al. Constraint nearest neighbor for instance reduction[J]. Soft Computing, 2019, 23(11).

[20] Li J, Zhu Q, Wu Q. A parameter-free hybrid instance selection algorithm based on local sets with natural neighbors[J]. Applied Intelligence, 2020, 50(5):1527-1541.

[21] Hernandez-Leal P, Carrasco-Ochoa J A, JF Martínez-Trinidad, et al. InstanceRank based on borders for instance selection[J]. Pattern Recognition, 2013, 46(1):365-375.

[22] Cavalcanti G, Soares R. RanKing-based Instance Selection for Pattern Classification[J]. Expert Systems with Applications, 2020, 150:113269.

[23] Wang Y, Pan Z, Pan Y. A Training Data Set Cleaning Method by Classification Ability RanKing for the K-Nearest Neighbor Classifier[J]. IEEE Transactions on Neural NetworKs and Learning Systems, 2019, PP(99):1-13.

[24] Blake C. UCI Repository of machine learning databases[J]. http://www.ics.uci.edu/mlearn/MLRepository.html, 1998.

[25] Alcala-Fdez J, Fernandez A, Luengo J, et al. KEEL Data-Mining Software Tool: Data Set Repository, Integration of Algorithms and Experimental Analysis FrameworK[J]. Journal of Multiple-Valued Logic & Soft Computing, 2011, 17(2-3):255-287.

[26] Wilcoxon F. Individual Comparisons by RanKing Methods[J]. Biometrics Bulletin, 1945, 1(6):80-83.

[27] Mayabadi S, Saadatfar H. Two density-based sampling approaches for imbalanced and overlapping data[J]. Knowledge-Based Systems, 2022, 241: 108217.

[28] Lin W C, Tsai C F, Hu Y H, et al. Clustering-based undersampling in class-imbalanced data[J]. Information Sciences, 2017, 409: 17-26.

[29] Tsai C F, Lin W C, Hu Y H, et al. Under-sampling class imbalanced datasets by combining clustering analysis and instance selection[J]. Information Sciences, 2019, 477: 47-54.

[30] Dai Q, Liu J, Shi Y. Class-overlap undersampling based on Schur decomposition for Class-imbalance problems[J]. Expert Systems with Applications, 2023, 221: 119735.