Available online at www.sciencedirect.com
ScienceDirect
Information Sciences 178 (2008) 1235–1256
INFORMATION SCIENCES
AN INTERNATIONAL JOURNAL
www.elsevier.com/locate/ins

A weighted rough set based method developed for class imbalance learning

Jinfu Liu *, Qinghua Hu, Daren Yu

Harbin Institute of Technology, P.O. Box 458, Harbin 150001, China

Received 30 July 2006; received in revised form 26 September 2007; accepted 1 October 2007

Abstract

In this paper, we introduce weights into Pawlak rough set model to balance the class distribution of a data set and develop a weighted rough set based method to deal with the class imbalance problem. In order to develop the weighted rough set based method, we design first a weighted attribute reduction algorithm by introducing and extending Guiasu weighted entropy to measure the significance of an attribute, then a weighted rule extraction algorithm by introducing a weighted heuristic strategy into LEM2 algorithm, and finally a weighted decision algorithm by introducing several weighted factors to evaluate extracted rules. Furthermore, in order to estimate the performance of the developed method, we compare the weighted rough set based method with several popular methods used for class imbalance learning by conducting experiments with twenty UCI data sets. Comparative studies indicate that in terms of AUC and minority class accuracy, the weighted rough set based method is better than the re-sampling and filtering based methods, and is comparable to the decision tree and SVM based methods. It is therefore concluded that the weighted rough set based method is effective for class imbalance learning.

Keywords: Rough sets; Class imbalance learning; Sample weighting; Weighted entropy; Rule extraction

1. Introduction

Class imbalance exists in a large number of real-world domains, such as fraud detection [11], medical diagnosis [26] and text classification [45], and is recognized as a crucial problem in machine learning and data mining. Much work has been done to deal with the class imbalance problem [9,24,26,42,57–59]. Many international workshops were dedicated to class imbalance learning, for example, AAAI'2000 – Workshop on Learning from Imbalanced Data Sets [25], ACM SIGKDD Exploration 2004 – Special Issue on Learning from Imbalanced Data Sets [7] and ICML'2003 – Workshop on Learning from Imbalanced Data Sets [6].

When the class distribution of a data set is highly skewed, a conventional machine learning method usually has a poor classification accuracy for unseen samples from the minority class because it is strongly biased

* Corresponding author. Tel.: +86 13212935624.
E-mail address: liujinfu@hcms.hit.edu.cn (J. Liu).

0020-0255/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.ins.2007.10.002

towards the majority class [32]. A recognized solution to the class imbalance problem is to balance the class distribution of a data set at the data or algorithmic level [6,25]. At the data level, re-sampling training data is a popular solution to the class imbalance problem, and it over-samples the minority class or under-samples the majority class to balance the class distribution of a data set. A conventional machine learning method can be directly used to deal with the class imbalance problem by learning from the re-sampled data set [9,24,60]. However, previous studies show that over-sampling usually increases training time and may lead to over-fitting because it introduces some exact copies of samples into a training data set, while under-sampling may degrade the performance of a resulting classifier because it usually discards some potentially useful training samples [1]. At the algorithmic level, sample weighting is a popular solution to the class imbalance problem, and it assigns a larger weight to the minority class to balance the class distribution of a data set. By using sample weighting, some standard machine learning methods, such as decision tree [9,52] and SVM [5,8,51], have been improved for class imbalance learning. Compared to re-sampling training data, sample weighting can usually be used to achieve better performance [24,26].

Rough set theory is a powerful mathematical tool introduced by Pawlak [35,36] to deal with inexact, uncertain or vague information, and has attracted attention of many researchers to contribute to its development and applications [2,3,14,28,29,33,47,61]. The main advantage of rough set theory in data analysis is that it does not need any preliminary or additional information about data such as probability distributions in statistics, basic probability assignments in Dempster–Shafer theory, or a grade of membership in fuzzy set theory [38–41]. In Pawlak rough set based method, all samples are considered equally important, and probability $1/n$ (n is the size of a training data set) is assigned to each sample for computation of the accuracy of approximation, reduction of attributes and extraction of decision rules. When the class distribution of a data set is highly skewed, the majority class can be adequately represented but the minority class may be neglected. As a result, Pawlak rough set based method usually has a poor classification accuracy for unseen samples from the minority class. In order to improve the classification accuracy for unseen samples from the minority class, it is usually necessary for a rough set based method to introduce a priori knowledge about samples to balance the class distribution of a data set.

In order to introduce a priori knowledge about samples into rough sets, Hu et al. [23] proposed probabilistic and fuzzy probabilistic rough set models, where each sample x is associated with probability $p(x)$ instead of $1/n$. $p(x)$ can be used to take into account a priori knowledge about samples, but it is difficult to determine $p(x)$. Ma et al. [31] introduced weights into the variable precision rough set model to represent the importance of each sample, and discussed the influence of weights on attribute reduction. However, they did not establish any learning or classification algorithm. In order to deal with the class imbalance problem using a rough set based method, Stefanowski et al. [50] introduced removing and filtering techniques to process inconsistent samples from the majority class in boundary regions. The removing and filtering techniques improve the performance of a rough set based method in class imbalance learning, and the filtering technique performs better than the removing technique. However, no matter which of them is used, the a priori knowledge about samples is introduced into boundary regions rather than the whole set of samples. Consequently, these techniques can only be used to improve learning from boundary regions.

It can be seen from the reviews above that sample weighting is a good solution to the class imbalance problem, but it has not been discussed in the framework of rough sets so far. In this study, we introduce weights into Pawlak rough set model to balance the class distribution of a data set and develop a weighted rough set based method to deal with the class imbalance problem. By conducting systematic comparative experiments with twenty UCI data sets, we find that the weighted rough set based method is effective for class imbalance learning.

The remainder of this paper is organized as follows. In Section 2, we present preliminary notions related to Pawlak rough sets. In Section 3, we review Shannon entropy based uncertainty measures of knowledge. In Section 4, we introduce weights into Pawlak rough set model to represented a priori knowledge about samples and propose a weighted rough set model. In Section 5, we introduce and extend Guiasu weighted entropy to measure the uncertainty of knowledge in the weighted rough set model. In Section 6, we establish some learning and classification algorithms based on the weighted rough set model. In Section 7, we discuss performance indexes in class imbalance learning. In Section 8, we conduct systematic comparative experiments to evaluate the performance of the weighted rough set based method in class imbalance learning. Finally, in Section 9, we give the conclusions drawn from this study.

2. Preliminary notions related to Pawlak rough sets

Definition 1. $IS = \langle U, A, V, f\rangle$ is called an information system, where $U = \{x_1, \ldots, x_i, \ldots, x_n\}$ is a set of samples, $A = \{a_1, \ldots, a_j, \ldots, a_m\}$ is a set of attributes, V is the value domain of A, and $f: U \times A \to V$ is an information function.

Let $B \subseteq A$. B induces an equivalence (indiscernibility) relation on U as shown below:

$$IND(B) = \{(x, y) \in U \times U | \forall a \in B, f(x, a) = f(y, a)\}. \tag{1}$$

The family of all equivalence classes of $IND(B)$, i.e. the partition induced by B, can be denoted as

$$\Pi_B = U/B = \{[x_i]_B : x_i \in U\}, \tag{2}$$

where $[x_i]_B$ is the equivalence class containing x_i, and all elements in $[x_i]_B$ are equivalent (indiscernible) with respect to B.

Theorem 1. *Let $IS = \langle U, A, V, f\rangle$ be a given information system, $B \subseteq A$, and Π_A and Π_B be the partitions induced by A and B respectively. Then Π_A is a refinement of Π_B, i.e. $[x_i]_A \subseteq [x_i]_B$ for $\forall x_i \in U$.*

Equivalence classes are elementary sets in rough set theory, and form basic knowledge granules about U. They are used in rough set theory to approximate any subset of U.

Definition 2. Let $IS = \langle U, A, V, f\rangle$ be a given information system, $B \subseteq A$ and $X \subseteq U$. The lower and upper approximations of X with respect to B, denoted by $\underline{B}X$ and $\overline{B}X$ respectively, are defined as

$$\begin{cases} \underline{B}X = \cup\{[x_i]_B | [x_i]_B \subseteq X\}, \\ \overline{B}X = \cup\{[x_i]_B | [x_i]_B \cap X \neq \varnothing\}. \end{cases} \tag{3}$$

Lower approximation $\underline{B}X$ is the set of all samples that can be certainly classified as belonging to X using B. Upper approximation $\overline{B}X$ is the set of all samples that can be possibly classified as belonging to X using B.

$BN_B(X) = \overline{B}X - \underline{B}X$ is called the boundary region of X with respect to B. X is definable with respect to B if $BN_B(X) = \varnothing$, otherwise X is rough with respect to B. In contrast to a definable set, any rough set has a non-empty boundary region. In rough set theory, boundary regions are used to express the uncertainty of knowledge.

The accuracy of approximation of X with respect to B is defined as

$$\alpha_B(X) = |\underline{B}X|/|\overline{B}X|, \tag{4}$$

where $|\cdot|$ denotes the cardinality of a set. X is definable with respect to B if $\alpha_B(X) = 1$, otherwise X is rough with respect to B.

Definition 3. Let $IS = \langle U, A, V, f\rangle$ be a given information system, $B \subseteq A$ and $a \in B$. a is redundant in B if $U/B = U/(B - a)$, otherwise a is indispensable in B. B is independent if every $a \in B$ is indispensable in B. B is a reduct of A if $U/B = U/A$ and B is independent.

A reduct is an independent subset of attributes that preserves the indiscernibility relation induced by full attributes. There is usually more than one reduct for a given information system, and the intersection of all the reducts is called the core.

Definition 4. $IS = \langle U, A, V, f\rangle$ is called a decision table if $A = C \cup D$ and $C \cap D = \varnothing$, where C is the condition attribute set and D is the decision attribute set.

For given decision table $IS = \langle U, A = C \cup D, V, f\rangle$, partition U/D forms the classification about U. $POS_C(D) = \cup_{X \in U/D} \underline{C}X$ is called the positive region of classification U/D with respect to C, and is the set of all samples that can be certainly classified as belonging to blocks of U/D using C.

The accuracy of approximation of classification U/D with respect to C is defined as

$$\alpha_C(D) = |POS_C(D)|/\sum_{X \in U/D} |\overline{C}X| \tag{5}$$

and the quality of approximation of classification U/D with respect to C is defined as

$$\gamma_C(D) = |POS_C(D)|/|U|, \tag{6}$$

$\gamma_C(D)$ is also defined as the degree of dependency of D on C. D totally depends on C if $\gamma_C(D) = 1$, i.e. all samples in U can be uniquely classified as belonging to blocks of U/D using C, otherwise D partially depends on C.

Definition 5. Let $IS = \langle U, A = C \cup D, V, f\rangle$ be a given decision table, $B \subseteq C$ and $a \in B$. a is redundant in B with respect to D if $\gamma_{B-a}(D) = \gamma_B(D)$, otherwise a is indispensable in B with respect to D. B is independent with respect to D if every $a \in B$ is indispensable in B with respect to D. B is a D-relative reduct of C if $\gamma_B(D) = \gamma_C(D)$ and B is independent with respect to D.

A relative reduct is an independent subset of condition attributes that preserves the degree of dependency of decision attributes on full condition attributes. There is usually more than one relative reduct for a given decision table, and the intersection of all the relative reducts is called the relative core.

3. Uncertainty measures of knowledge based on Shannon entropy

Shannon entropy has been widely used in rough set theory to measure the uncertainty of knowledge [10,21,22,48,55,56], and these measures can be reviewed as shown below.

Let $IS = \langle U, A, V, f\rangle$ be a given information system, $C \subseteq A$ and $D \subseteq A$. Partitions $\Pi_C = \{X_1, \ldots, X_i, \ldots, X_n\}$ and $\Pi_D = \{Y_1, \ldots, Y_j, \ldots, Y_m\}$ induced by C and D respectively can be considered two random variables in δ-algebra. The probability distributions of Π_C and Π_D can be respectively denoted as

$$(X; P) = \begin{pmatrix} X_1 & \cdots & X_i & \cdots & X_n \\ p(X_1) & \cdots & p(X_i) & \cdots & p(X_n) \end{pmatrix} \tag{7}$$

and

$$(Y; P) = \begin{pmatrix} Y_1 & \cdots & Y_j & \cdots & Y_m \\ p(Y_1) & \cdots & p(Y_j) & \cdots & p(Y_m) \end{pmatrix}, \tag{8}$$

where $p(X_i) = \frac{|X_i|}{|U|}$ is the probability of X_i and $p(Y_j) = \frac{|Y_j|}{|U|}$ is the probability of Y_j.

Similarly, the joint probability distribution of Π_C and Π_D, i.e. the probability distribution of partition $\Pi_{C\cup D}$ jointly induced by C and D, can be denoted as

$$(X \otimes Y; P) = \begin{pmatrix} X_1 \cap Y_1 & \cdots & X_i \cap Y_j & \cdots & X_n \cap Y_m \\ p(X_1, Y_1) & \cdots & p(X_i, Y_j) & \cdots & p(X_n, Y_m) \end{pmatrix}, \tag{9}$$

where $p(X_i, Y_j) = \frac{|X_i \cap Y_j|}{|U|}$ is the joint probability of X_i and Y_j.

Definition 6. Shannon entropy of Π_C, which is also called Shannon entropy of C, is defined as

$$H(C) = -\sum_{i=1}^{n} p(X_i) \log p(X_i). \tag{10}$$

$H(C)$ measures the uncertainty of partition Π_C induced by C.

Definition 7. The joint entropy of Π_C and Π_D, which is also called the joint entropy of C and D, is defined as

$$H(C, D) = -\sum_{i=1}^{n} \sum_{j=1}^{m} p(X_i, Y_j) \log p(X_i, Y_j). \tag{11}$$

$H(C, D)$ measures the uncertainty of partition $\Pi_{C\cup D}$ jointly induced by C and D.

Definition 8. The conditional entropy of Π_D given Π_C, which is also called the conditional entropy of D given C, is defined as

$$H(D|C) = -\sum_{i=1}^{n} \sum_{j=1}^{m} p(X_i, Y_j) \log p(Y_j|X_i) = -\sum_{i=1}^{n} p(X_i) \sum_{j=1}^{m} p(Y_j|X_i) \log p(Y_j|X_i), \tag{12}$$

where $p(Y_j|X_i) = \frac{|X_i \cap Y_j|}{|X_i|}$ is the conditional probability of Y_j given X_i.

$H(D|C)$ measures the uncertainty of partition Π_D induced by D under the condition that partition Π_C induced by C has been given.

Theorem 2. *$H(C,D) = H(C) + H(D|C)$.*

Theorem 3. *Let $IS = \langle U, A, V, f\rangle$ be a given information system and $B \subseteq A$. Then*

(1) $H(A,B) = H(A)$;
(2) $H(B|A) = 0$;
(3) $H(A) \geqslant H(B)$.

Theorem 4. *Let $IS = \langle U, A, V, f\rangle$ be a given information system and $a \in A$. a is redundant in A if $H(A|A - a) = 0$, otherwise a is indispensable in A.*

Theorem 5. *Let $IS = \langle U, A, V, f\rangle$ be a given information system and $B \subseteq A$. B is a reduct of A if B satisfies*

(1) $H(A|B) = 0$;
(2) $H(B|B - a) > 0$ *for* $\forall a \in B$.

Theorem 6. *Let $IS = \langle U, A = C \cup D, V, f\rangle$ be a given decision table and $E \subseteq B \subseteq C$. Then $H(D|E) \geqslant H(D|B)$.*

Theorem 7. *Let $IS = \langle U, A = C \cup D, V, f\rangle$ be a given consistent decision table, $B \subseteq C$ and $a \in B$. a is redundant in B with respect to D if $H(D|B - a) = H(D|B)$, otherwise a is indispensable in B with respect to D. B is independent with respect to D if $H(D|B - a) > H(D|B)$ for $\forall a \in B$. B is a D-relative reduct of C if B satisfies*

(1) $H(D|B) = H(D|C)$;
(2) $H(D|B - a) > H(D|B)$ *for* $\forall a \in B$.

The detailed proofs of Theorems 2–7 can be found in [55,56].

4. Weighted rough set model

In Pawlak rough set model, all samples are considered equally important. In order to introduce a priori knowledge about samples into rough set based data analysis, we employ weights to represent the a priori knowledge, and propose a weighted rough set model to accomplish this.

Definition 9. $WIS = \langle U, A, W, V, f\rangle$ is called a weighted information system, where $U = \{x_1, \ldots, x_i, \ldots, x_n\}$ is a set of samples, $A = \{a_1, \ldots, a_j, \ldots, a_m\}$ is a set of attributes, $W = \{w(x_1), \ldots, w(x_i), \ldots, w(x_n)\}$ is a weight distribution on U, V is the value domain of A, and $f : U \times A \to V$ is an information function.

For given weighted information system $WIS = \langle U, A, W, V, f\rangle$, weight distribution W is used to represent a priori knowledge about samples. The introduced weights do not change the equivalence relations in a conventional information system, and so do not change the upper and lower approximations of arbitrary subset $X \subseteq U$. However, the introduced weights change the accuracy of approximation of X.

Definition 10. Let $WIS = \langle U, A, W, V, f\rangle$ be a given weighted information system, $B \subseteq A$, $X \subseteq U$, and $\underline{B}X$ and $\overline{B}X$ be the lower and upper approximations of X with respect to B respectively. The weighted accuracy of approximation of X with respect to B is defined as

$$\alpha_B^W(X) = |\underline{B}X|_W / |\overline{B}X|_W, \tag{13}$$

where $|\underline{B}X|_W = \sum_{x_i \in \underline{B}X} w(x_i)$ is the weighted cardinality of $\underline{B}X$ and $|\overline{B}X|_W = \sum_{x_i \in \overline{B}X} w(x_i)$ is the weighted cardinality of $\overline{B}X$.

Compared to the conventional accuracy of approximation, the weighted accuracy of approximation is defined based on the weighted cardinalities of sets. By using the weighted cardinalities of sets, a priori knowledge about samples can be calculated in the weighted accuracy of approximation. Similarly, the weighted degree of dependency of decision attributes on condition attributes can also be defined based on the weighted cardinalities of sets.

Definition 11. $WIS = \langle U, A, W, V, f\rangle$ is called a weighted decision table if $A = C \cup D$ and $C \cap D = \emptyset$, where C is the condition attribute set and D is the decision attribute set.

Let $POS_C(D)$ be the positive region of classification U/D with respect to C. The weighted accuracy of approximation of classification U/D with respect to C is defined as

$$\alpha_C^W(D) = |POS_C(D)|_W \Big/ \sum_{X \in U/D} |\overline{C}X|_W, \tag{14}$$

and the weighted quality of approximation of classification U/D with respect to C is defined as

$$\gamma_C^W(D) = |POS_C(D)|_W / |U|_W. \tag{15}$$

$\gamma_C^W(D)$ is also defined as the weighted degree of dependency of D on C. D totally depends on C if $\gamma_C^W(D) = 1$, otherwise D partially depends on C.

Definition 12. Let $WIS = \langle U, A = C \cup D, W, V, f\rangle$ be a given weighted decision table, $B \subseteq C$ and $a \in B$. a is redundant in B with respect to D if $\gamma_{B-a}^W(D) = \gamma_B^W(D)$, otherwise a is indispensable in B with respect to D. B is independent with respect to D if every $a \in B$ is indispensable in B with respect to D. B is a D-relative reduct of C if $\gamma_B^W(D) = \gamma_C^W(D)$ and B is independent with respect to D.

5. Guiasu weighted entropy based uncertainty measures of knowledge

Shannon entropy can be used in Pawlak rough set model to measure the uncertainty of knowledge, but can not be used to take into account a priori knowledge about samples. Guiasu introduced weights into Shannon entropy to represent the a priori knowledge and proposed weighted entropy to deal with this problem [15]. However, he did not propose the weighted conditional and joint entropies. In this section, we introduce and extend Guiasu weighted entropy to measure the uncertainty of knowledge in the weighted rough set model.

Definition 13. Let X and Y be two subsets of U, $p(X)$, $p(Y)$ and $p(X \cup Y)$ be the probabilities of, and $X \cup Y$ respectively, and $w(X)$, $w(Y)$ and $w(X \cup Y)$ be the weights of, and $X \cup Y$ respectively. If $X \cap Y = \emptyset$, $w(X \cup Y)$ is defined as

$$w(X \cup Y) = \frac{w(X)p(X) + w(Y)p(Y)}{p(X \cup Y)}. \tag{16}$$

Definition 14. Let X and Y be two subsets of U, and $w(X)$ and $w(Y)$ be the weights of X and Y respectively. The conditional weight of Y given X, denoted by $w(Y|X)$, is defined as

$$w(Y|X) = \frac{w(X \cap Y)}{w(X)}. \tag{17}$$

For given weighted information system $WIS = \langle U, A, W, V, f\rangle$, let $\Pi_A = \{X_1, \ldots, X_i, \ldots, X_n\}$ be the partition induced by A. The weighted probability distribution of Π_A can be denoted as

$$(X; P; W) = \begin{pmatrix} X_1 & \cdots & X_i & \cdots & X_n \\ p(X_1) & \cdots & p(X_i) & \cdots & p(X_n) \\ w(X_1) & \cdots & w(X_i) & \cdots & w(X_n) \end{pmatrix}, \tag{18}$$

where $p(X_i)$ is the probability of X_i and $w(X_i)$ is the weight of X_i.

Definition 15. Guiasu weighted entropy of Π_A, which is also called Guiasu weighted entropy of A, is defined as

$$H_W(A) = -\sum_{i=1}^{n} w(X_i)p(X_i)\log p(X_i). \tag{19}$$

Theorem 8. *If $w(X_1) = w(X_2) = \cdots = w(X_n) = w$, then $H_W(A) = wH(A)$, i.e. Guiasu weighted entropy degenerates to Shannon entropy.*

Theorem 9. *If $p(X_i) = 0$, $w(X_i) > 0$ for $\forall i \in I$, whereas $p(X_j) > 0$, $w(X_j) = 0$ for $\forall j \in J$, where $I \cup J = \{1,2,,\ldots,n\}$ and $I \cap J = \varnothing$, then $H_W(A) = 0$.*

Theorem 9 means that if some subsets of samples are interesting for applications but do not occur, whereas the others occur but are not interesting, then Guiasu weighted entropy is zero, i.e. no interesting information is obtained.

Based on the definition of Guiasu weighted entropy, the weighed conditional and joint entropies can be respectively proposed as shown below.

Definition 16. Let $WIS = \langle U, A = C \cup D, W, V, f\rangle$ be a given weighted decision table, $\Pi_C = \{X_1,\ldots,X_i,\ldots,X_n\}$ be the partition induced by C, $\Pi_D = \{Y_1,\ldots,Y_j,\ldots,Y_m\}$ be the partition induced by D, and $(X;P;W) = \begin{pmatrix} X_1 & \cdots & X_i & \cdots & X_n \\ p(X_1) & \cdots & p(X_i) & \cdots & p(X_n) \\ w(X_1) & \cdots & w(X_i) & \cdots & w(X_n) \end{pmatrix}$ and $(Y;P;W) = \begin{pmatrix} Y_1 & \cdots & Y_j & \cdots & Y_m \\ p(Y_1) & \cdots & p(Y_j) & \cdots & p(Y_m) \\ w(Y_1) & \cdots & w(Y_j) & \cdots & w(Y_m) \end{pmatrix}$ be the weighted probability distributions of Π_C and Π_D respectively. The weighted conditional entropy of Π_D given Π_C, which is also called the weighted conditional entropy of D given C, is defined as

$$\begin{aligned} H_W(D|C) &= -\sum_{i=1}^{n}\sum_{j=1}^{m} w(X_i \cap Y_j)p(X_i \cap Y_j)\log p(Y_j|X_i) \\ &= -\sum_{i=1}^{n} w(X_i)p(X_i)\sum_{j=1}^{m} w(Y_j|X_i)p(Y_j|X_i)\log p(Y_j|X_i). \end{aligned} \tag{20}$$

Definition 17. Let $WIS = \langle U, A = C \cup D, W, V, f\rangle$ be a given weighted decision table, $\Pi_C = \{X_1,\ldots,X_i,\ldots,X_n\}$ be the partition induced by C, $\Pi_D = \{Y_1,\ldots,Y_j,\ldots,Y_m\}$ be the partition induced by D, and $(X;P;W) = \begin{pmatrix} X_1 & \cdots & X_i & \cdots & X_n \\ p(X_1) & \cdots & p(X_i) & \cdots & p(X_n) \\ w(X_1) & \cdots & w(X_i) & \cdots & w(X_n) \end{pmatrix}$ and $(Y;P;W) = \begin{pmatrix} Y_1 & \cdots & Y_j & \cdots & Y_m \\ p(Y_1) & \cdots & p(Y_j) & \cdots & p(Y_m) \\ w(Y_1) & \cdots & w(Y_j) & \cdots & w(Y_m) \end{pmatrix}$ be the weighted probability distributions of Π_C and Π_D respectively. The weighted joint entropy of Π_C and Π_D, which is also called the weighted joint entropy of C and D, is defined as

$$H_W(C,D) = -\sum_{i=1}^{n}\sum_{j=1}^{m} w(X_i \cap Y_j)p(X_i \cap Y_j)\log p(X_i \cap Y_j). \tag{21}$$

Theorem 10. *Let $H_W(C)$ be Guiasu weighted entropy of C, $H_W(D|C)$ be the weighted conditional entropy of D given C, and $H_W(C,D)$ be the weighted joint entropy of C and D. Then $H_W(C,D) = H_W(C) + H_W(D|C)$.*

Proof. According to Definition 17,

$$\begin{aligned} H_W(C,D) &= -\sum_{i=1}^{n}\sum_{j=1}^{m} w(X_i \cap Y_j)p(X_i \cap Y_j)\log p(X_i \cap Y_j) \\ &= -\sum_{i=1}^{n}\sum_{j=1}^{m} w(X_i \cap Y_j)p(X_i \cap Y_j)\log(p(X_i)p(Y_j|X_i)) \\ &= -\sum_{i=1}^{n}\log p(X_i)\sum_{j=1}^{m} w(X_i \cap Y_j)p(X_i \cap Y_j) - \sum_{i=1}^{n}\sum_{j=1}^{m} w(X_i \cap Y_j)p(X_i \cap Y_j)\log p(Y_j|X_i). \end{aligned}$$

We have $\bigcup_{j=1}^{m}(X_i \cap Y_j) = X_i$ and $\bigcap_{j=1}^{m}(X_i \cap Y_j) = \varnothing$. Thus, according to Definition 13, $\sum_{j=1}^{m} w(X_i \cap Y_j) p(X_i \cap Y_j) = w(X_i)p(X_i)$. Then $H_W(C,D) = H_W(C) + H_W(D|C)$ holds. This finishes the proof. □

Theorems 3–7 also hold for Guiasu weighted entropy. They can be proved in a similar way.

6. Learning and classification based on weighted rough sets

6.1. Weighted attribute reduction

Attribute reduction is an important problem which can be solved using rough set theory, and a number of algorithms have been established for attribute reduction [3,28,29,33,47]. However, the conventional attribute reduction algorithms can not be used to take into account a priori knowledge about samples. Based on Guiasu weighted entropy, we define the weighted significance of an attribute and design a heuristic weighted attribute reduction algorithm to take into account the a priori knowledge.

Definition 18. Let $WIS = \langle U, A = C \cup D, W, V, f\rangle$ be a given weighted decision table and $B \subseteq C$. Based on Guiasu weighted entropy, the weighted significance of attribute $a \in C - B$ on the basis of B with respect to D is defined as

$$SIG_W(a, B, D) = H_W(D|B) - H_W(D|B \cup \{a\}). \tag{22}$$

Based on the weighted significance of an attribute, a heuristic weighted attribute reduction algorithm is designed as Algorithm 1. Algorithm 1 starts with an empty attribute set, and iteratively selects an attribute with the maximum weighted significance until the weighted significance of the remaining attributes is below threshold ε.

Algorithm 1. Weighted attribute reduction algorithm

Input: Weighted decision table $WIS = \langle U, A = C \cup D, W, V, f\rangle$ and threshold ε.

Output: D-relative reduct B of C.

```
1.  begin
2.    compute H_W(D|C);
3.    B ← ∅;
4.    while B ⊂ C do
5.    begin
6.      for each a ∈ C − B do
7.        compute SGF_W(a, B, D);
8.      select a such that SGF_W(a, B, D) is maximum;
9.      B ← B ∪ {a};
10.     if H_W(D|B) − H_W(D|C) < ε then exit the loop;
11.   end
12.   for each a ∈ B
13.     if H_W(D|B − {a}) − H_W(D|C) < ε then B ← B − {a};
14.   return B;
15. end
```

6.2. Weighted rule extraction

Nowadays, there are many known rule extraction algorithms inspired by rough set theory [16,19,20,27,34,46,49,53,54], and LEM2 algorithm proposed by Grzymala-Busse [16] is one of the most widely used algorithms for real-world applications. However, these algorithms can not be used to take into account a priori knowledge about samples. Based on LEM2 algorithm, we design a weighted rule extraction algorithm to deal with this problem.

The following are some preliminary descriptions about LEM2 algorithm.

In order to reduce rule extraction from inconsistent samples to that from consistent samples, a family of generalized decisions, denoted by $\widetilde{D}$, is first defined on a given decision table. A generalized decision is either a single decision or a joint decision. According to $\widetilde{D}$, all the samples in the decision table are then partitioned into a family of disjoint subsets, denoted by $\widetilde{Y}$. Each element of $\widetilde{Y}$ is either the lower approximation of a decision class, which corresponds to a single decision of $\widetilde{D}$, or one of the disjoint subsets of the boundary region of a decision class, which corresponds to a joint decision of $\widetilde{D}$. Suppose that there are three decision classes Y_1, Y_2 and Y_3 in the decision table and B is a subset of condition attributes. The boundary region of Y_1 with respect to B consists of three disjoint subsets, i.e. $BN_B(Y_1) = (\overline{B}Y_1 \cap \overline{B}Y_2 - \overline{B}Y_3) \cup (\overline{B}Y_1 \cap \overline{B}Y_3 - \overline{B}Y_2) \cup (\overline{B}Y_1 \cap \overline{B}Y_2 \cap \overline{B}Y_3)$. By using the generalized decisions, the inconsistent samples in terms of the original decisions are expressed as the consistent samples in terms of the generalized decisions. Finally, for each $K \in \widetilde{Y}$, a heuristic strategy is used in LEM2 algorithm to extract a minimum set of rules.

Let $WIS = \langle U, A = C \cup D, W, V, f \rangle$ be a given weighted decision table, $\widetilde{D}_K$ be a generalized decision, $K \in \widetilde{Y}$ be the subset of samples that corresponds to $\widetilde{D}_K$, C be an elementary condition (condition attribute-value pair) that has an expression (a, v), where $a \in C$ and $v \in V_a$, $\Phi = c_1 \wedge \cdots \wedge c_j \wedge \cdots \wedge c_q$ be the conjunction of q elementary conditions, $[\Phi]$ be the cover of Φ, i.e. the subset of samples that satisfy all elementary conditions of Φ, $[\Phi]_K^+ = [\Phi] \cap K$ be the positive cover of Φ on K, and $[\Phi]_K^- = [\Phi] \cap (U - K)$ be the negative cover of Φ on K. Then a rule, denoted by r, is described as

$$\text{if } \Phi \text{ then } \widetilde{D}_K, \tag{23}$$

where Φ is called the condition part of r, satisfying $[\Phi]_K^+ \neq \varnothing$, and $\widetilde{D}_K$ is called the decision part of r. If $\widetilde{D}_K$ is a single decision, r is called a certain rule. Otherwise, if $\widetilde{D}_K$ is a joint decision, r is called a possible rule.

Definition 19. r is discriminant if its condition part $\Phi = c_1 \wedge \cdots \wedge c_j \wedge \cdots \wedge c_q$ is

(1) consistent: $[\Phi]_K^- = \varnothing$;
(2) minimal: Φ is no longer consistent if arbitrary elementary condition c_j is removed from Φ.

Definition 20. A set of rules, denoted by R, is called a minimum set of rules if it describes generalized decision $\widetilde{D}_K$ in the following ways:

(1) every rule $r \in R$ is discriminant;
(2) $\cup_{r \in R}[\Phi] = K$;
(3) there does not exist any rule $r \in R$ such that $R - \{r\}$ satisfies conditions (1) and (2).

LEM2 algorithm has been widely used to extract a minimum set of rules from samples. In order to introduce a priori knowledge about samples into LEM2 algorithm, we design a weighted rule extraction algorithm as Algorithm 2.

Algorithm 2. Weighted rule extraction algorithm

Input: $K \in \widetilde{Y}$.

Output: Set R of rules.

```
1.  begin
2.    G ← K, R ← ∅;
3.    while G ≠ ∅ do
4.    begin
5.      Φ ← ∅;
6.      Φ_G ← {c:[c] ∩ G ≠ ∅};
7.      while (Φ = ∅) or (not([Φ] ⊆ K)) do
8.      begin
9.        select c ∈ Φ_G such that |[c] ∩ G|_W is maximum. if ties occur, select c with the smallest |[c]|_W. if
          further ties occur, select the first c from the list;
```

```
10.          Φ ← Φ ∪ {c};
11.          G ← [c] ∩ G;
12.          Φ_G ← {c:[c] ∩ G ≠ ∅};
13.          Φ_G ← Φ_G − Φ;
14.        end
15.        for each c ∈ Φ do
16.          if [Φ − {c}] ⊆ K then Φ ← Φ − {c};
17.        create rule r based on Φ;
18.        R ← R ∪ {r};
19.        G ← K − ∪_{r∈R}[r];
20.      end
21.      for each r ∈ R do
22.        if ∪_{s∈R−{r}}[S] = K then R ← R − {r};
23.    end
```

6.3. Weighted rule evaluation and weighted decision

Pawlak introduced the factors of strength, certainty and cover to evaluate extracted rules [37]. However, the factors can not be used to take into account a priori knowledge about samples. In this section, we propose several weighted factors to evaluate extracted rules, and design a weighted decision algorithm to classify an unseen sample.

Definition 21. Let $WIS = \langle U, A = C \cup D, W, V, f\rangle$ be a given weighted decision table, r be a decision rule extracted from the weighted decision table, $\widetilde{D}_K = \{d_1, \ldots, d_j, \ldots, d_n\}$ be the decision part of r, $[r]$ be the cover of r, $[\widetilde{D}_K]$ be the cover of $\widetilde{D}_K$, and $[r]_d^+ = [r] \cap [d]$ be the positive cover of r on d, where $d \in \widetilde{D}_K$. Then the factor of weighted strength of r is defined as

$$\mu_{\mathrm{str}}^W(r) = |[r]|_W / |U|_W, \tag{24}$$

the factor of weighted cover of r is defined as

$$\mu_{\mathrm{cov}}^W(r) = |[r]|_W / |[\widetilde{D}_K]|_W, \tag{25}$$

and the factor of weighted certainty of r to d is defined as

$$\mu_{\mathrm{cer}}^W(r, d) = |[r]_d^+|_W / |[r]|_W. \tag{26}$$

Extracted rules can be used to classify an unseen sample by matching the description of the sample to the condition part of each rule. This may lead to three possible cases:

(1) the sample matches exactly one rule;
(2) the sample matches more than one rule;
(3) the sample does not match any of the rules.

In case (1), if the matched rule is a certain one, it is clear that the class of the sample can be predicted using the decision of the matched rule. However, if the matched rule is a possible one, the classification is ambiguous. Similar difficulties occur in case (2). Case (3) must be also handled.

We predict the class of the sample using the most frequent class of training samples if the sample does not match any of the rules, and design a weighted decision algorithm as shown below to deal with the remaining cases.

Definition 22. Suppose that the sample matches rules $r_1, \ldots, r_i, \ldots, r_n$, and decisions $d_1, \ldots, d_j, \ldots, d_m$ are suggested. The factor of weighted strength of decision d_j is defined as

$$\mu^W_{\mathrm{str}}(d_j) = \sum_{r_i} \mu^W_{\mathrm{cer}}(r_i, d_j)\mu^W_{\mathrm{str}}(r_i). \tag{27}$$

Based on the factor of weighted strength of a decision, a weighted decision algorithm can be designed to classify the sample. The class of the sample can be predicted using decision d_j that maximizes $\mu^W_{\mathrm{str}}(d_j)$.

7. Performance indexes in class imbalance learning

The most straightforward way to evaluate the performance of a classifier is based on the confusion matrix analysis. Table 1 shows a confusion matrix for a two-class problem with positive and negative class values. In our study, the minority class is defined as the positive class and the majority class is defined as the negative class. From such a matrix it is possible to extract a number of widely used metrics to measure the performance of a classifier, such as error rate, defined as $\mathrm{Err} = \frac{\mathrm{FP+FN}}{\mathrm{TP+FN+TN+FP}}$, and overall accuracy, defined as $\mathrm{Acc} = \frac{\mathrm{TP+TN}}{\mathrm{TP+FN+TN+FP}} = 1 - \mathrm{Err}$.

However, the use of such measures may lead to misleading conclusions when the class distribution of a data set is highly skewed. Overall accuracy and error rate are particularly suspicious performance measures because they are strongly biased towards the majority class. For instance, it is straightforward to create a classifier with an overall accuracy of 99% or an error rate of 1% in a domain where the proportion of the majority class corresponds to 99% of all samples, by simply predicting every new sample as belonging to the majority class.

Another fact against the use of overall accuracy and error rate is that these measures consider different classification errors to be equally important. However, highly imbalanced problems generally have highly non-uniform error costs that often favor the minority class of primary interest. For instance, diagnosing a sick patient as healthy may be a fatal error, while diagnosing a healthy patient as sick is usually considered a much less serious error since this mistake can be corrected in future exams.

Finally, overall accuracy and error rate change as the class distribution of a data set changes even if the fundamental performance of a classifier does not change, because these measures use values from both lines of the confusion matrix.

It would be more interesting if we use a performance metric to disassociate the errors or hits that occur in each class. It is possible to derive four performance metrics from Table 1 to measure the classification performance on the positive and negative classes independently:

(1) True positive rate (minority class accuracy): $\mathrm{TP}_{\mathrm{rate}} = \mathrm{TP}/(\mathrm{TP} + \mathrm{FN})$ is the percentage of positive samples correctly classified as belonging to the positive class;
(2) True negative rate (majority class accuracy): $\mathrm{TN}_{\mathrm{rate}} = \mathrm{TN}/(\mathrm{TN} + \mathrm{FP})$ is the percentage of negative samples correctly classified as belonging to the negative class;
(3) False positive rate: $\mathrm{FP}_{\mathrm{rate}} = \mathrm{FP}/(\mathrm{FP} + \mathrm{TN})$ is the percentage of negative samples misclassified as belonging to the positive class;
(4) False negative rate: $\mathrm{FN}_{\mathrm{rate}} = \mathrm{FN}/(\mathrm{FN} + \mathrm{TP})$ is the percentage of positive samples misclassified as belonging to the negative class.

The four performance measures have the advantage of being independent of class costs and a priori probabilities. The aim of a classifier is to minimize false positive and negative rates, or similarly to maximize true negative and positive rates. Unfortunately, there is usually a tradeoff between $\mathrm{FP}_{\mathrm{rate}}$ and $\mathrm{FN}_{\mathrm{rate}}$, or similarly between $\mathrm{TN}_{\mathrm{rate}}$ and $\mathrm{TP}_{\mathrm{rate}}$. ROC (receiver operating characteristic) graphs can be used to analyze the relationship between $\mathrm{FP}_{\mathrm{rate}}$ and $\mathrm{FN}_{\mathrm{rate}}$, or similarly between $\mathrm{TN}_{\mathrm{rate}}$ and $\mathrm{TP}_{\mathrm{rate}}$ [12,43]. ROC graphs are consistent for

Table 1
Confusion matrix for a two-class problem

	Positive prediction	Negative prediction
Positive class	True positive (TP)	False negative (FN)
Negative class	False positive (FP)	True negative (TN)

a given problem even if the distribution of positive and negative samples is highly skewed. The area under the ROC curve (AUC) represents the expected performance as a single scalar. Furthermore, AUC has a known statistical meaning: it is equivalent to the Wilconxon test of ranks, and is also equivalent to several other statistical measures for evaluating classification and ranking models [17].

In our study, we employ minority class accuracy, majority class accuracy, overall accuracy and AUC as the performance indexes to evaluate the performance of a learning method in class imbalance learning. For a multi-class problem, the minimum class accuracy is defined as the minority class accuracy, the maximum class accuracy is defined as the majority class accuracy, and the AUC is computed using the method proposed in [18].

8. Experimental evaluation

In order to evaluate the performance of the weighted rough set based method in class imbalance learning, systematic comparative experiments are conducted in this section. The methods employed for the comparison comprise eight rough set based methods, two decision tree based methods and six SVM based methods. All the methods are described and configured as shown below:

(a) Eight rough set based methods (RS is the conventional method, and the others are the methods for class imbalance learning).
   (1) WRS: the weighted rough set based method. Algorithm 1 is employed to perform weighted attribute reduction, Algorithm 2 is employed to perform weighted rule extraction, the weighted factors given in Definition 21 are employed to evaluate extracted rules and the weighted decision algorithm based on the weighted factors is employed to classify an unseen sample. An inverse class probability weight is assigned to each sample for class imbalance learning. Suppose that there are $n_1, \ldots, n_i, \ldots, n_l$ samples in decision classes $Y_1, \ldots, Y_i, \ldots, Y_l$ respectively. Then the inverse class probability weight of each sample from decision class Y_i is $1/n_i$.
   (2) RS: Pawlak rough set based method. RS can be considered WRS with equal weighting. Instead of an inverse class probability weight, an equal weight is assigned to each sample.
   (3) WAR: WRS only with weighted attribute reduction. An inverse class probability weight is assigned to each sample for weighted attribute reduction, while an equal weight is assigned to each sample for weighted rule extraction and weighted decision.
   (4) WRE: WRS only with weighted rule extraction and weighted decision. An inverse class probability weight is assigned to each sample for weighted rule extraction and weighted decision, while an equal weight is assigned to each sample for weighted attribute reduction.
   (5) FILTER: the filtering method proposed by Stefanowski et al. [50]. The inconsistent samples from the majority class in boundary regions are relabeled as belonging to the minority class, and then RS is used to perform learning and classification.
   (6) OS: the random over-sampling method. The ith class is randomly over-sampled until the size of the ith class is equal to the size of the maximum class, and then RS is used to perform learning and classification. The random over-sampling method is used because it is simple and is competitive with other complicated over-sampling methods in performance [1].
   (7) US: the random under-sampling method. The ith class is randomly under-sampled until the size of the ith class is equal to the size of the minimum class, and then RS is used to perform learning and classification.
   (8) MS: the random middle-sampling method. The ith class is randomly over-sampled or under-sampled until the size of the ith class is equal to the mean size of the maximum and minimum classes, and then RS is used to perform learning and classification.

(b) Two decision tree based methods (C4.5 is the conventional method, and C4.5_CS is the method for class imbalance learning).
   (1) C4.5: C4.5 decision tree proposed by Quinlan [44].
   (2) C4.5_CS: cost-sensitive (weighted) C4.5 decision tree proposed by Ting [52]. An inverse class probability weight is assigned to each sample for class imbalance learning.

(c) Six SVM based methods (SVM_R1, SVM_R100 and SVM_L100 are the conventional methods, and the others are the methods for class imbalance learning).
  (1) SVM_R1: SVM with Rbf kernel function and parameter $C = 1$.
  (2) WSVM_R1: weighted SVM [5] with Rbf kernel function and parameter $C = 1$. An inverse class probability weight is assigned to each class for class imbalance learning.
  (3) SVM_R100: SVM with Rbf kernel function and parameter $C = 100$.
  (4) WSVM_R100: weighted SVM with Rbf kernel function and parameter $C = 100$. An inverse class probability weight is assigned to each class for class imbalance learning.
  (5) SVM_L100: SVM with linear kernel function and parameter $C = 100$.
  (6) WSVM_L100: weighted SVM with linear kernel function and parameter $C = 100$. An inverse class probability weight is assigned to each class for class imbalance learning.

Twenty UCI data sets [4], which consist of 10 two-class data sets and 10 multi-class data sets, are used in our experiments. All the data sets are described in Table 2. It can be seen from Table 2 that the class distribution of each data set is skewed. Concretely, the ratio of the majority class to the minority class in size ranges from 1.25 to 3.84 for the two-class data sets, and the ratio of the maximum class to the minimum class in size ranges from 1.48 to 85.5 for the multi-class data sets. Moreover, the size of the minimum class is below 10 for most multi-class data sets.

The comparative experiments are performed using 10-fold cross validation. The missing values in each data set are filled with mean values for continuous attributes and majority values for nominal attributes. Moreover, all the continuous attributes are discretized using the recursive minimal entropy partitioning method proposed by Fayyad and Irani [13,30] when a rough set based method is used.

The minority class accuracy, majority class accuracy, overall accuracy and AUC achieved by each method on each data set are listed in Tables 3–10. It can be seen from the experimental results that almost all the methods for class imbalance learning improve minority class accuracy and AUC, and decrease majority class accuracy and overall accuracy compared to the conventional methods. Class imbalance learning aims at improving minority class accuracy but not decreasing majority class accuracy too much. AUC can be used to analyze the relationship between minority class accuracy and majority class accuracy. Therefore, we employ AUC as the primary performance index and minority class accuracy as the secondary performance index to evaluate the performance of each method in class imbalance learning.

Three strategies, i.e. weighting, re-sampling and filtering, are used in the rough set based methods for class imbalance learning. Almost all the methods based on these strategies improve AUC and minority class accuracy compared to Pawlak rough set based method. This means that all the strategies are effective for class imbalance learning. The detailed analysis of each strategy is given as shown below:

(1) The methods based on the strategy of weighting comprise WRS, WAR and WRE. In terms of AUC, WRS has an average increase of 0.0332, WAR 0.011 and WRE 0.0191 compared to RS. In terms of minority class accuracy, WRS has an average increase of 0.1376, WAR 0.0092 and WRE 0.0931 compared to RS. It can be seen that WRS achieves the best performance. Moreover, WRE achieves better performance than WAR, and this means that weighted rule extraction and weighted decision have greater influence on the performance of the weighted rough set based method than weighted attribute reduction.
(2) The methods based on the strategy of re-sampling comprise OS, MS and US. In terms of AUC, OS achieves the best performance and US achieves the worst performance, while in terms of minority class accuracy, MS achieves the best performance and US achieves the worst performance. This may be explained by the fact that US usually discards some potentially useful training samples, and so the performance of a resulting classifier is usually degraded [1]. Moreover, compared to WRS, OS has an average decrease of 0.0222 in terms of AUC and MS has an average decrease of 0.0525 in terms of minority class accuracy. This means that WRS is better than the re-sampling based methods.
(3) As the method based on the strategy of filtering, FILTER can not be used to remarkably improve AUC and minority class accuracy compared to RS. This may be explained by the fact that FILTER introduces a priori knowledge about samples into boundary regions rather than the whole set of samples, and so it can only be used to improve learning from boundary regions.

Among the rough set based methods for class imbalance learning, WRS achieves the best performance. Moreover, among the three strategies for class imbalance learning, weighting is the best, and filtering is the worst.

The decision tree based methods achieve satisfactory performance in class imbalance learning. Compared to RS, C4.5 has an average increase of 0.0323 in terms of AUC and 0.0981 in terms of minority class accuracy. By sample weighting, compared to C4.5, C4.5_CS has an average increase of 0.0055 in terms of AUC and 0.0086 in terms of minority class accuracy. It can be seen from the comparison between C4.5 and C4.5_CS

Table 2
Description of data sets (C: Continuous, N: Nominal)

	Data set	Size	Attribute		Class	Class distribution
1	Echocardiogram	131	6C	1N	2	43/88
2	Hepatitis	155	6C	13N	2	32/123
3	Heart_s	270	6C	7N	2	120/150
4	Breast	286		9N	2	85/201
5	Horse	368	7C	15N	2	136/232
6	Votes	435		16N	2	168/267
7	Credit	690	6C	9N	2	307/383
8	Breast_w	699	9C		2	241/458
9	Tic	958		9N	2	332/626
10	German	1000	24C		2	300/700
11	Zoo	101		16N	7	4/5/8/10/13/20/41
12	Lymphography	148		18N	4	2/4/61/81
13	Wine	178	13C		3	48/59/71
14	Machine	209	7C		8	2/4*2/7/13/27/31/121
15	Glass	214	9C		6	9/13/17/29/70/76
16	Audiology	226		69N	24	1*5/2*7/3/4*3/6/8/9/20/22*2/48/57
17	Heart	303	6C	7N	5	13/35/36/55/164
18	Solar	323		10N	3	7/29/287
19	Soybean	683		35N	19	8/14/15/16/20*9/44*2/88/91*2/92
20	Anneal	898	6C	32N	5	8/40/67/99/684

Table 3
Minority class accuracy achieved by rough set based methods

Data set	RS	WRS	WAR	WRE	FILTER	OS	US	MS
Echocardiogram	0.2450	0.7350	0.6100	0.6900	0.2900	0.6150	0.2850	0.5900
Hepatitis	0.6333	0.8583	0.7917	0.6500	0.6333	0.6667	0.8500	0.7250
Heart_s	0.7333	0.7500	0.7583	0.7333	0.7333	0.7333	0.7500	0.7167
Breast	0.2222	0.3514	0.2333	0.3514	0.2681	0.4097	0.5639	0.4444
Horse	0.9484	0.9412	0.9341	0.9555	0.9484	0.9555	0.9637	0.9341
Votes	0.9761	0.9761	0.9761	0.9761	0.9761	0.9699	0.9515	0.9585
Credit	0.8173	0.8308	0.8046	0.8373	0.8109	0.8176	0.8148	0.8374
Breast_w	0.8958	0.9292	0.9208	0.9167	0.8917	0.9250	0.9375	0.9417
Tic	0.7766	0.8250	0.7766	0.8250	0.7887	0.7715	0.8854	0.8127
German	0.0267	0.9067	0.1133	0.8100	0.0333	0.6067	0.4500	0.6933
Zoo	0.7000	0.7000	0.7000	0.7000	0.7000	0.7000	0.7000	0.7000
Lymphography	0.3333	0.7190	0.4024	0.3333	0.3333	0.7190	0.3476	0.7024
Wine	0.8650	0.9350	0.9350	0.8900	0.8650	0.9150	0.9150	0.9350
Machine	0.6000	0.8000	0.3000	0.7000	0.6000	0.5000	0.3000	0.4000
Glass	0.6000	0.7000	0.3000	0.6000	0.6000	0.6000	0.6000	0.5000
Audiology	0.7000	0.5000	0.7000	0.7000	0.8000	0.5000	0.3000	0.6000
Heart	0	0.3000	0	0.2000	0	0.1500	0.3000	0.2500
Solar	0.0333	0.1000	0.0333	0.1000	0.0333	0.1000	0.1000	0.0667
Soybean	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
Anneal	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
Average	0.6053	0.7429	0.6145	0.6984	0.6153	0.6827	0.6507	0.6904

Table 4
Minority class accuracy achieved by decision tree and SVM based methods

Data set	C4.5	C4.5 _CS	SVM _R1	WSVM _R1	SVM _R100	WSVM _R100	SVM _L100	WSVM _L100
Echocardiogram	0.5050	0.4600	0.0950	0.6700	0.2150	0.2150	0.4750	0.6900
Hepatitis	0.7583	0.6917	0	0	0	0	0.6583	0.5667
Heart_s	0.8267	0.8067	0.0083	1.0000	0.0250	0.0250	0.7917	0.8667
Breast	0.3167	0.5042	0.1556	0.5681	0.4333	0.4236	0.3653	0.5569
Horse	0.9264	0.9489	0.3176	0.3538	0.3907	0.3907	0.8747	0.8896
Votes	0.9765	0.9824	0.9882	0.9824	0.9393	0.9574	0.9699	0.9816
Credit	0.8435	0.8502	0.0424	0.9514	0.0652	0.0652	–	–
Breast_w	0.9458	0.9128	0.9875	0.9875	0.9917	0.9917	0.9625	0.9625
Tic	0.8915	0.9307	0.7196	0.7767	0.9880	0.9880	0.5357	0.6804
German	0.4833	0.5400	0.1467	0.1967	0.2967	0.2967	0.4900	0.6133
Zoo	0.7000	0.8000	0.6000	0.3000	0.8000	0.8000	0.7000	0.8000
Lymphography	0.7500	0.7857	0.8000	0.8000	0.9000	0.8000	1.0000	0.8000
Wine	0.9100	0.9500	0.0900	1.0000	0.1300	0.1300	0.9800	0.9600
Machine	0.6846	0.6929	0.8000	0.8000	0.8000	0.8000	–	–
Glass	0.6000	0.7000	0.2000	0.1000	0.7000	0.7000	0.6000	0.7000
Audiology	0.8500	0.5500	0.8000	0.8000	0.8000	0.8000	0.8000	0.8000
Heart	0.1000	0	0	0	0	0	0.1000	0.4000
Solar	0	0.1333	0.3000	0.3000	0.3000	0.5000	0.3000	0.6000
Soybean	1.0000	1.0000	1.0000	0.2000	1.0000	1.0000	1.0000	1.0000
Anneal	1.0000	1.0000	0.7000	0.2000	0.7000	0.7000	–	–
Average	0.7034	0.7120	0.4375	0.5493	0.5237	0.5292	0.6825	0.7569

Table 5
Majority class accuracy achieved by rough set based methods

Data set	RS	WRS	WAR	WRE	FILTER	OS	US	MS
Echocardiogram	0.8639	0.6833	0.7958	0.6236	0.8639	0.6569	0.6194	0.6125
Hepatitis	0.9006	0.9340	0.9506	0.9006	0.9256	0.8865	0.8192	0.8596
Heart_s	0.8000	0.8200	0.8267	0.8200	0.8200	0.8000	0.8067	0.8000
Breast	0.8257	0.7810	0.8257	0.7810	0.8207	0.7860	0.6319	0.7314
Horse	0.9783	0.9783	0.9783	0.9783	0.9783	0.9696	0.9565	0.9739
Votes	0.9661	0.9624	0.9624	0.9624	0.9698	0.9625	0.9437	0.9624
Credit	0.8382	0.8330	0.8304	0.8225	0.8433	0.8174	0.8226	0.8096
Breast_w	0.9672	0.9650	0.9650	0.9672	0.9672	0.9629	0.9542	0.9453
Tic	0.9137	0.8962	0.9137	0.8962	0.9314	0.9042	0.8642	0.8882
German	0.9843	0.3043	0.9400	0.4686	0.9843	0.6500	0.2886	0.5343
Zoo	1.0000	0.9000	0.9750	1.0000	1.0000	0.8750	0.8000	0.8500
Lymphography	0.9028	0.8028	0.7903	0.8903	0.8903	0.7917	0.6069	0.7417
Wine	0.9714	0.9143	0.9429	0.9571	0.9714	0.9018	0.8714	0.8732
Machine	0.8263	0.7103	0.7936	0.6179	0.8263	0.7679	0.4250	0.6519
Glass	0.6679	0.6125	0.6857	0.6000	0.6679	0.6589	0.1714	0.5482
Audiology	0.9500	0.9133	0.9133	0.9500	0.9667	0.8633	0.1800	0.6933
Heart	0.8246	0.8118	0.8728	0.7324	0.8371	0.7504	0.5188	0.6713
Solar	0.9583	0.8192	0.9548	0.8225	0.9583	0.7628	0.4080	0.7494
Soybean	0.8289	0.7956	0.7956	0.8178	0.8411	0.8056	0.6311	0.8489
Anneal	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
Average	0.8984	0.8219	0.8856	0.8304	0.9032	0.8287	0.6660	0.7873

that a decision tree based method is not sensitive to the class distribution of a data set. Furthermore, compared to WRS, C4.5_CS has an average increase of 0.0046 in terms of AUC and an average decrease of 0.0309 in terms of minority class accuracy. This means that C4.5_CS is just a litter better than WRS in terms of AUC, while WRS is much better than C4.5_CS in terms of minority class accuracy.

Table 6
Majority class accuracy achieved by decision tree and SVM based methods

Data set	C4.5	C4.5 _CS	SVM _R1	WSVM _R1	SVM _R100	WSVM _R100	SVM _L100	WSVM _L100
Echocardiogram	0.7389	0.7069	0.9528	0.5569	0.8292	0.8292	0.8306	0.7611
Hepatitis	0.9109	0.8942	1.0000	1.0000	1.0000	1.0000	0.9346	0.9263
Heart_s	0.7167	0.6833	1.0000	0	1.0000	1.0000	0.8800	0.8000
Breast	0.8707	0.6914	0.9302	0.6667	0.7069	0.7114	0.8755	0.6960
Horse	0.9913	0.9870	0.9748	0.9707	0.9489	0.9489	0.9397	0.9141
Votes	0.9738	0.9738	0.9477	0.9440	0.9664	0.9664	0.9625	0.9662
Credit	0.8668	0.8695	0.9662	0.1596	0.9557	0.9557	–	–
Breast_w	0.9672	0.9606	0.9454	0.9454	0.9476	0.9476	0.9716	0.9716
Tic	0.9761	0.9488	0.9984	0.8994	1.0000	1.0000	0.8499	0.7045
German	0.8457	0.7743	0.9757	0.9429	0.8829	0.8829	0.8886	0.8143
Zoo	1.0000	1.0000	1.0000	0.7750	1.0000	1.0000	1.0000	1.0000
Lymphography	0.7750	0.8250	0.9264	0	0.8653	0.6667	0.8514	0.4444
Wine	0.9286	0.9429	1.0000	0	1.0000	1.0000	0.9286	0.9446
Machine	0.3500	0.4000	0.9750	1.0000	0.9750	0.9750	–	–
Glass	0.6446	0.7000	0.7268	0.8196	0.6946	0.8982	0.6607	0.8821
Audiology	0.8300	0.5933	0.8833	0.1000	0.9667	0.2767	0.9833	0.5867
Heart	0.8434	0.7147	1.0000	0.2000	1.0000	1.0000	0.9018	0.8107
Solar	1.0000	0.7313	0.9964	0.6000	0.9583	0.6200	1.0000	0.5372
Soybean	0.9700	0.9600	0.9789	0.2356	0.9344	0.9233	0.9567	0.9233
Anneal	1.0000	1.0000	0.9971	1.0000	0.9971	0.9971	–	–
Average	0.8600	0.8179	0.9588	0.5908	0.9314	0.8800	0.9068	0.8049

Table 7
Overall accuracy achieved by rough set based methods

Data set	RS	WRS	WAR	WRE	FILTER	OS	US	MS
Echocardiogram	0.6566	0.7038	0.7335	0.6484	0.6714	0.6484	0.5077	0.6093
Hepatitis	0.8442	0.9163	0.9167	0.8504	0.8642	0.8383	0.8242	0.8313
Heart_s	0.7704	0.7889	0.7963	0.7815	0.7815	0.7704	0.7815	0.7630
Breast	0.6470	0.6538	0.6505	0.6538	0.6573	0.6749	0.6127	0.6472
Horse	0.9672	0.9645	0.9618	0.9699	0.9672	0.9645	0.9590	0.9592
Votes	0.9702	0.9679	0.9679	0.9679	0.9725	0.9656	0.9472	0.9610
Credit	0.8290	0.8319	0.8188	0.8290	0.8290	0.8174	0.8188	0.8217
Breast_w	0.9428	0.9528	0.9499	0.9499	0.9413	0.9499	0.9485	0.9441
Tic	0.8664	0.8717	0.8664	0.8717	0.8821	0.8581	0.8716	0.8623
German	0.6970	0.4850	0.6920	0.5710	0.6990	0.6370	0.3370	0.5820
Zoo	0.9300	0.9000	0.9300	0.9300	0.9300	0.8900	0.8209	0.8600
Lymphography	0.8110	0.7771	0.7433	0.8110	0.8043	0.7438	0.4824	0.7033
Wine	0.9317	0.9327	0.9382	0.9435	0.9317	0.9098	0.9157	0.9268
Machine	0.6598	0.6652	0.6505	0.6124	0.6743	0.6790	0.4645	0.6167
Glass	0.6905	0.6255	0.6074	0.6385	0.6905	0.6223	0.4143	0.6126
Audiology	0.7524	0.6377	0.6460	0.7700	0.7700	0.6067	0.1721	0.5834
Heart	0.5149	0.5576	0.5742	0.4783	0.5578	0.4851	0.3730	0.4585
Solar	0.8641	0.7494	0.8610	0.7495	0.8641	0.7027	0.3848	0.6874
Soybean	0.8317	0.8858	0.8800	0.8259	0.8361	0.8888	0.6676	0.8814
Anneal	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9922	1.0000
Average	0.8088	0.7934	0.8092	0.7926	0.8162	0.7826	0.6648	0.7656

Among the SVM based methods, all the methods with Rbf kernel function, i.e. SVM_R1, WSVM_R1, SVM_R100 and WSVM_R100, have low values of AUC and minority class accuracy. Sample weighting can not help a SVM based method with Rbf kernel function remarkably improve AUC and minority class accuracy. A big value of parameter C is helpful for improving the performance of a SVM based method,

Table 8
Overall accuracy achieved by decision tree and SVM based methods

Data set	C4.5	C4.5 _CS	SVM _R1	WSVM _R1	SVM _R100	WSVM _R100	SVM _L100	WSVM _L100
Echocardiogram	0.6637	0.6258	0.6720	0.5962	0.6275	0.6275	0.7170	0.7401
Hepatitis	0.8775	0.8513	0.7937	0.7937	0.7937	0.7937	0.8779	0.8513
Heart_s	0.7778	0.7519	0.5593	0.4444	0.5667	0.5667	0.8407	0.8296
Breast	0.7067	0.6365	0.6995	0.6367	0.6261	0.6259	0.7241	0.6546
Horse	0.9673	0.9727	0.7311	0.7419	0.7419	0.7419	0.9155	0.9048
Votes	0.9749	0.9772	0.9634	0.9588	0.9563	0.9633	0.9656	0.9725
Credit	0.8565	0.8609	0.5551	0.5116	0.5594	0.5594	–	–
Breast_w	0.9599	0.9442	0.9600	0.9600	0.9628	0.9628	0.9685	0.9685
Tic	0.9468	0.9425	0.9019	0.8570	0.9958	0.9958	0.7412	0.6963
German	0.7370	0.7040	0.7270	0.7190	0.7070	0.7070	0.7690	0.7540
Zoo	0.9200	0.9600	0.9400	0.7418	0.9600	0.9600	0.9500	0.9600
Lymphography	0.7767	0.7833	0.7843	0.4124	0.8305	0.7505	0.8310	0.6343
Wine	0.9376	0.9487	0.4448	0.2693	0.4784	0.4784	0.9605	0.9546
Machine	0.6748	0.6840	0.6031	0.5790	0.6031	0.5936	–	–
Glass	0.6680	0.7093	0.6675	0.4210	0.7000	0.5690	0.6636	0.4857
Audiology	0.7972	0.7879	0.5010	0.1684	0.8237	0.3721	0.8372	0.5854
Heart	0.5411	0.4945	0.5412	0.2056	0.5412	0.5412	0.5908	0.5476
Solar	0.8886	0.6872	0.8855	0.5699	0.8640	0.5973	0.8886	0.5172
Soybean	0.9326	0.9312	0.9400	0.1685	0.9355	0.9487	0.9458	0.9458
Anneal	1.0000	1.0000	0.9254	0.7617	0.9599	0.8619	–	–
Average	0.8302	0.8127	0.7398	0.5758	0.7617	0.7108	0.8345	0.7648

Table 9
AUC achieved by rough set based methods

Data set	RS	WRS	WAR	WRE	FILTER	OS	US	MS
Echocardiogram	0.5544	0.7092	0.7029	0.6568	0.5769	0.6360	0.4522	0.6013
Hepatitis	0.7670	0.8962	0.8712	0.7753	0.7795	0.7766	0.8346	0.7923
Heart_s	0.7667	0.7850	0.7925	0.7767	0.7767	0.7667	0.7783	0.7583
Breast	0.5240	0.5662	0.5295	0.5662	0.5444	0.5978	0.5979	0.5879
Horse	0.9633	0.9597	0.9562	0.9669	0.9633	0.9625	0.9601	0.9540
Votes	0.9711	0.9692	0.9692	0.9692	0.9730	0.9662	0.9476	0.9604
Credit	0.8278	0.8319	0.8175	0.8299	0.8271	0.8175	0.8187	0.8235
Breast_w	0.9315	0.9471	0.9429	0.9420	0.9295	0.9439	0.9459	0.9435
Tic	0.8452	0.8606	0.8452	0.8606	0.8600	0.8378	0.8748	0.8505
German	0.5055	0.6055	0.5267	0.6393	0.5088	0.6283	0.3693	0.6138
Zoo	0.9054	0.9135	0.9264	0.9112	0.9054	0.9110	0.8424	0.8821
Lymphography	0.7358	0.7984	0.7109	0.7208	0.7326	0.7699	0.5583	0.7470
Wine	0.9424	0.9498	0.9528	0.9535	0.9424	0.9334	0.9383	0.9479
Machine	0.6930	0.7650	0.6812	0.7341	0.7110	0.7059	0.6326	0.7129
Glass	0.7868	0.7801	0.7063	0.7824	0.7868	0.7261	0.7014	0.7543
Audiology	0.7829	0.7272	0.7259	0.7908	0.7939	0.7060	0.5523	0.7152
Heart	0.5422	0.6165	0.5760	0.5684	0.5773	0.5544	0.5513	0.5700
Solar	0.5479	0.5381	0.5470	0.5306	0.5479	0.5324	0.4520	0.5207
Soybean	0.9076	0.9447	0.9392	0.9079	0.9064	0.9469	0.8516	0.9400
Anneal	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9904	1.0000
Average	0.7750	0.8082	0.7860	0.7941	0.7821	0.7860	0.7325	0.7838

but can not enable a SVM based method with Rbf kernel function to achieve the performance better than RS in class imbalance learning. It can therefore be concluded that the SVM based methods with Rbf kernel function are not suitable for class imbalance learning. However, our experiments indicate that the SVM based methods with Linear kernel function are effective for class imbalance learning. Compared to RS, SVM_L100

Table 10
AUC achieved by decision tree and SVM based methods

Data set	C4.5	C4.5 _CS	SVM _R1	WSVM _R1	SVM _R100	WSVM _R100	SVM _L100	WSVM _L100
Echocardiogram	0.6219	0.5835	0.5239	0.6135	0.5221	0.5221	0.6528	0.7256
Hepatitis	0.8346	0.7929	0.5000	0.5000	0.5000	0.5000	0.7965	0.7465
Heart_s	0.7717	0.7450	0.5042	0.5000	0.5125	0.5125	0.8358	0.8333
Breast	0.5937	0.5978	0.5429	0.6174	0.5701	0.5675	0.6204	0.6264
Horse	0.9588	0.9679	0.6462	0.6622	0.6698	0.6698	0.9072	0.9018
Votes	0.9751	0.9781	0.9680	0.9632	0.9529	0.9619	0.9662	0.9739
Credit	0.8552	0.8599	0.5043	0.5555	0.5104	0.5104	–	–
Breast_w	0.9565	0.9367	0.9665	0.9665	0.9696	0.9696	0.9670	0.9670
Tic	0.9338	0.9398	0.8590	0.8381	0.9940	0.9940	0.6928	0.6924
German	0.6645	0.6571	0.5612	0.5698	0.5898	0.5898	0.6893	0.7138
Zoo	0.9137	0.9600	0.9500	0.8396	0.9667	0.9667	0.9583	0.9667
Lymphography	0.7799	0.8100	0.8322	0.7333	0.8970	0.8611	0.9252	0.8407
Wine	0.9513	0.9607	0.5392	0.5000	0.5667	0.5667	0.9730	0.9678
Machine	0.6690	0.7128	0.6798	0.6643	0.6798	0.6762	–	–
Glass	0.7895	0.8176	0.7129	0.5886	0.7897	0.7665	0.7570	0.6863
Audiology	0.8262	0.8315	0.8394	0.8107	0.9384	0.8765	0.9415	0.9100
Heart	0.5727	0.5587	0.5000	0.5062	0.5000	0.5000	0.5944	0.6192
Solar	0.5000	0.5667	0.5741	0.5750	0.5979	0.6425	0.5750	0.6176
Soybean	0.9780	0.9799	0.9713	0.5295	0.9800	0.9850	0.9822	0.9833
Anneal	1.0000	1.0000	0.8441	0.5250	0.8918	0.7807	–	–
Average	0.8073	0.8128	0.7009	0.6529	0.7300	0.7210	0.8138	0.8102

has an average increase of 0.0388 in terms of AUC and 0.0772 in terms of minority class accuracy. By sample weighting, compared to SVM_L100, WSVM_L100 has an average decrease of 0.0036 in terms of AUC and an average increase of 0.0744 in terms of minority class accuracy. It can be seen from the comparison between SVM_L100 and WSVM_L100 that a SVM based method with Linear kernel function is not sensitive to the class distribution of a data set in terms of AUC, but sample weighting can help it greatly improve minority class accuracy. Furthermore, compared to WRS, WSVM_L100 has an average increase of 0.002 in terms of AUC and 0.014 in terms of minority class accuracy. This means that WSVM_L100 is a little better than WRS. However, the SVM based methods with Linear kernel function are quite time-consuming. In our experiments, they can not generate experimental results on three data sets (i.e. credit, machine and anneal) in an acceptable period of time.

It can be seen from these comparative experiments that in terms of AUC and minority class accuracy, the weighted rough set based method is better than the re-sampling and filtering based methods, and is comparable to the decision tree and SVM based methods. It can therefore be concluded that the weighted rough set based method is effective for class imbalance learning. Furthermore, compared to a decision tree or SVM based method, a rough set based method is much sensitive to the class distribution of a data set. When the class distribution of a data set is highly skewed, it is necessary for a rough set based method to employ some techniques for class imbalance learning.

In the experiments above, an inverse class probability weight is assigned to each sample for class imbalance learning. In order to verify whether the inverse class probability weighting is the optimal solution to the class imbalance problem, we conduct further experiments to analyze the influence of weights on the performance of a learning method in class imbalance learning.

Definition 23. Suppose that n_1 and n_2 are the sizes of the minority and majority classes respectively, and w_1 and w_2 are the weights of the minority and majority classes respectively. Then the probability weight function of the minority class (the positive class) is defined as

$$PWF(+) = \frac{n_1 \times w_1}{n_1 \times w_1 + n_2 \times w_2}. \tag{28}$$

$PWF(+)$ increases as the weight of the minority class increases. When $PWF(+) = 0.5$, an inverse class probability weight is assigned to each class and the class distribution of a data set is completely balanced.

The experiments are conducted on 10 two-class data sets using the weighted rough set based method. By changing weights w_1 and w_2 such that $PWF(+) = \{0.05, 0.1, 0.15, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.85, 0.9, 0.95\}$, we obtain the variation of performance indexes versus $PWF(+)$ as shown in Fig. 1. It can be seen from Fig. 1 that weights have significant influence on the performance of the weighted rough set based method in class imbalance learning. On almost all the data sets, with the increase of $PWF(+)$, minority class accuracy

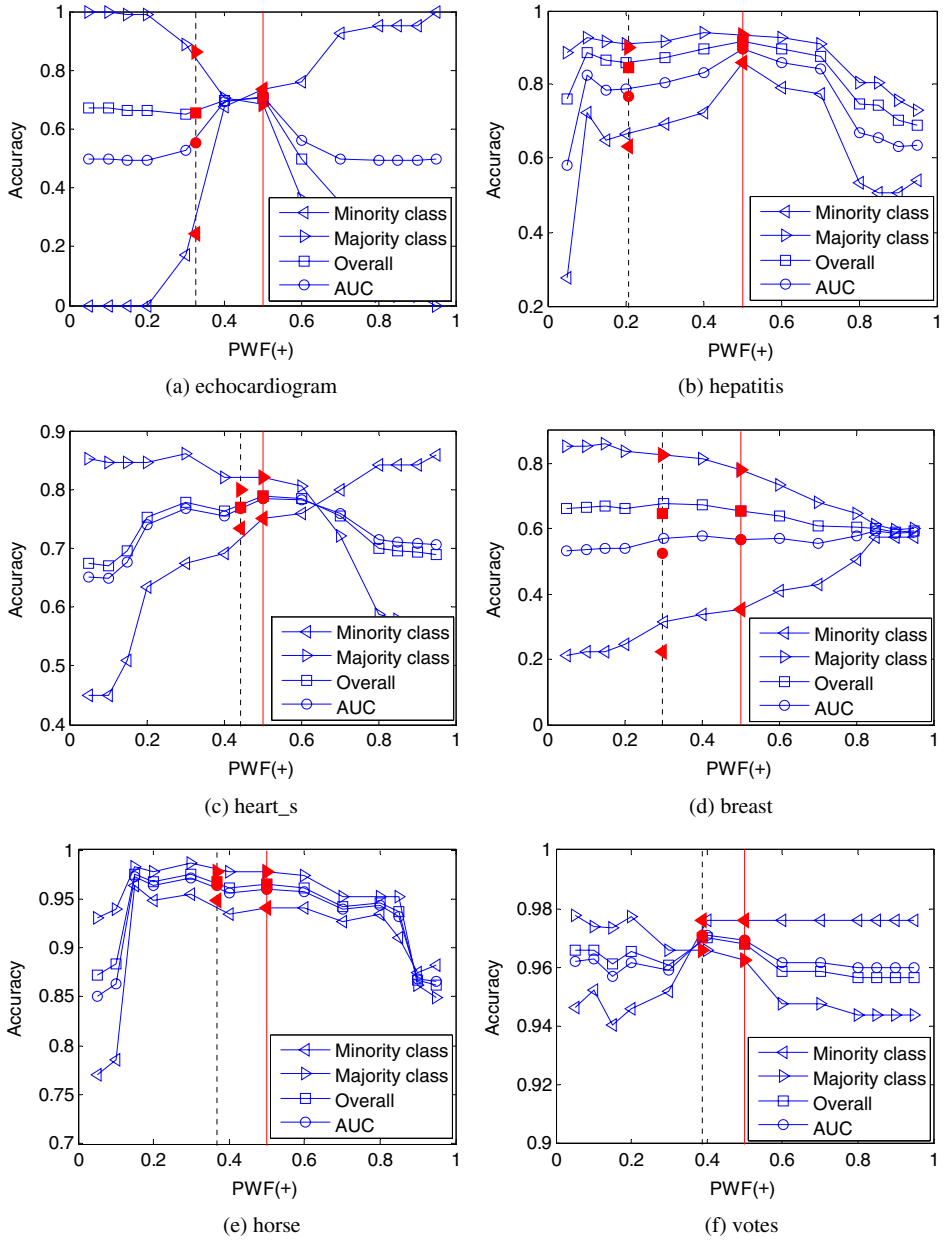
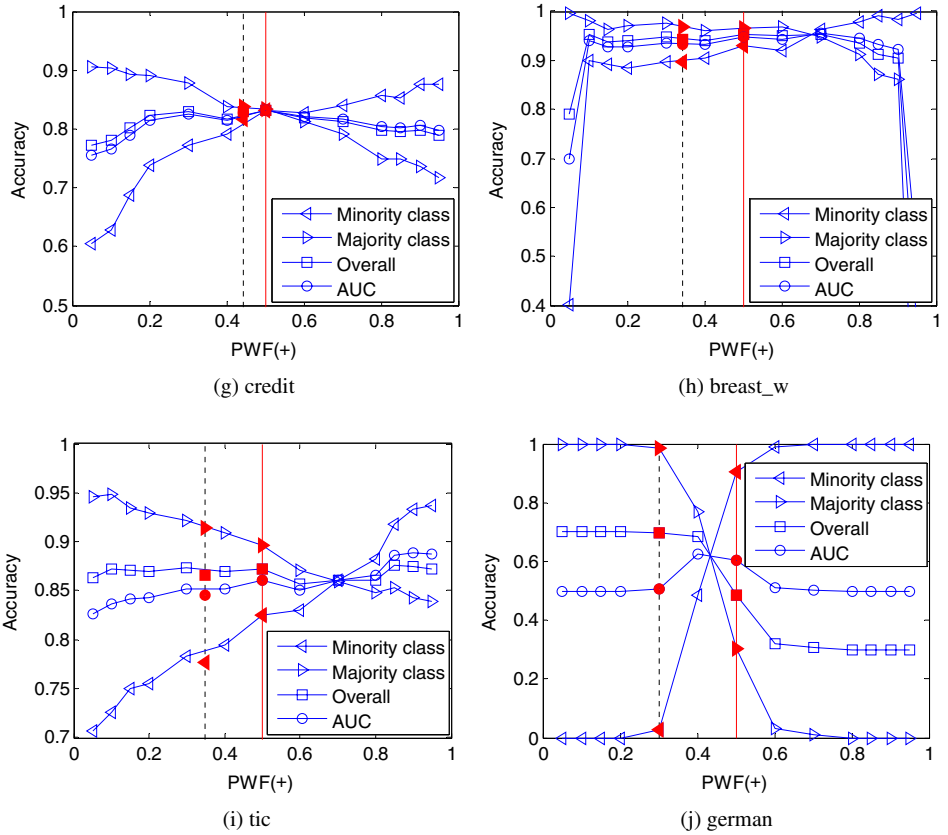


Fig. 1. Performance indexes versus $PWF(+)$ on 10 two-class data sets. The vertical real line represents the case of a completely balanced class distribution obtained using the inverse class probability weighting. The vertical dashed line represents the case of an original class distribution. AUC is usually optimal or suboptimal in the case of a completely balanced class distribution, and this means that the inverse class probability weighting is a simple and effective solution to the class imbalance problem.



(g) credit (h) breast_w

(i) tic (j) german

Fig. 1 (*continued*)

increases, majority class accuracy decreases, and overall accuracy and AUC first increase and then decrease. AUC is usually optimal or suboptimal in the case of $PWF(+) = 0.5$, and this means that the inverse class probability weighting is a simple and effective solution to the class imbalance problem.

9. Conclusion

We introduce weights into Pawlak rough set model to balance the class distribution of a data set and develop a weighted rough set based method to deal with the class imbalance problem. In order to estimate the performance of the developed method, we compare the weighted rough set based method with several popular methods used for class imbalance learning by conducting experiments with twenty UCI data sets. Comparative studies indicate that in terms of AUC and minority class accuracy, the weighted rough set based method is better than the re-sampling and filtering based methods, and is comparable to the decision tree and SVM based methods. It is therefore concluded that the weighted rough set based method is effective for class imbalance learning.

Furthermore, we also find from the experimental results that: (1) weighted rule extraction and weighted decision have greater influence on the performance of the weighted rough set based method than weighted attribute reduction, and this will guide us to improve the weighted rough set based method; (2) compared to a decision tree or SVM based method, a rough set based method is much sensitive to the class distribution of a data set, and so it is necessary for a rough set based method to employ some techniques for class imbalance learning; (3) AUC is usually optimal or suboptimal in the case of a completely balanced class distribution, and so the inverse class probability weighting is a simple and effective solution to the class imbalance problem.

Acknowledgements

We would like to thank the anonymous reviewers and Prof. X.F. Wen for their valuable comments and suggestions to improve this paper. This work is supported by National Natural Science Foundation of China under Grant 60703013.

References

[1] G. Batista, R.C. Prati, M.C. Monard, A study of the behavior of several methods for balancing machine learning training data, SIGKDD Explorations 6 (1) (2004) 20–29.
[2] J. Bazan, A comparison of dynamic and non-dynamic rough set methods for extracting laws from decision tables, in: L. Polkowski, A. Skowron (Eds.), Rough Sets in Knowledge Discovery, Phisica-Verlag, Heidelberg, 1998, pp. 321–365.
[3] M. Beynon, Reducts within the variable precision rough sets model: a further investigation, European Journal of Operational Research 134 (2001) 592–605.
[4] C. Blake, E. Keogh, C.J. Merz, UCI Repository of Machine Learning Databases, Dept. of Information and Computer Science, Univ. of California, Irvine, 1998. Available from: <http://www.ics.uci.edu/~mlearn/MLRepository.html>.
[5] U. Brefeld, P. Geibel, F. Wysotzki, Support vector machines with example dependent costs, in: N. Lavrac, D. Gamberger, L. Todorovski, H. Blockeel (Eds.), Proc. 14th European Conf. Machine Learning, ECML'03, Cavtat, Croatia, 2003, pp. 23–34.
[6] N. Chawla, N. Japkowicz, A. Kolcz (Eds.), Proc. ICML'03 Workshop on Learning from Imbalanced Data Sets (II), Washington DC, USA, 2003. Available from: <http://www.site.uottawa.ca/~nat/Workshop2003/workshop2003.html>.
[7] N. Chawla, N. Japkowicz, A. Kolcz (Eds.), Special Issues on Learning from Imbalanced Data SetsSIGKDD Explorations 6 (1) (2004) 1–6.
[8] N. Cristianini, J. Shawe-Taylor, An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods, Cambridge University Press, Cambridge, 2000.
[9] C. Drummond, R.C. Holte, C4.5, class imbalance, and cost sensitivity: why under-sampling beats over-sampling, in: N. Chawla, N. Japkowicz, A. Kolcz (Eds.), Proc. ICML'03 Workshop on Learning from Imbalanced Data Sets (II), Washington DC, USA, 2003. Available from: <http://www.site.uottawa.ca/~nat/Workshop2003/imbalance03.tar.gz>.
[10] I. Duntsch, G. Gediga, Uncertainty measures of rough set prediction, Artificial Intelligence 106 (1) (1998) 109–137.
[11] R.E. Fawcett, F. Provost, Adaptive fraud detection, Data Mining and Knowledge Discovery 3 (1) (1997) 291–316.
[12] T. Fawcett, An introduction to ROC analysis, Pattern Recognition Letters 27 (2006) 861–874.
[13] U.M. Fayyad, K.B. Irani, Multi-interval discretization of continuous-valued attributes for classification learning, in: R. Bajcsy (Eds.), Proc. 13th Int. Joint Conf. Artificial Intelligence, IJCAI'93, Chambery, France, 1993, pp. 1022–1027.
[14] S. Greco, B. Matarazzo, R. Slowinski, Rough set theory for multicriteria decision analysis, European Journal of Operational Research 129 (2001) 1–47.
[15] S. Guiasu, Information Theory with Applications, McGraw-Hill, International Book Company, New York, 1977.
[16] J.W. Grzymala-Busse, LERS – a system for learning from examples based on rough sets, in: R. Slowinski (Ed.), Intelligent Decision Support: Handbook of Applications and Advances of the Rough Sets Theory, Kluwer Academic Publishers, Dordrecht, 1992, pp. 3–18.
[17] D.J. Hand, Construction and Assessment of Classification Rules, John Wiley and Sons, New York, 1997.
[18] D.J. Hand, R.J. Till, A simple generalization of the area under the ROC curve to multiple class classification problems, Machine Learning 45 (2) (2001) 171–186.
[19] X.-H. Hu, N. Cercone, Data mining via discretization, generalization and rough set feature selection, Knowledge and Information Systems 1 (1) (1999) 33–60.
[20] Q.-H. Hu, X.-D. Li, D.-R. Yu, Analysis on classification performance of rough set based reducts, in: Q. Yang, G. Webb (Eds.), Proc. 9th Pacific Rim Int. Conf. Artificial Intelligence, PRICAI'06, LNAI 4099, Springer-Verlag, Heidelberg, 2006, pp. 423–433.
[21] Q.-H. Hu, D.-R. Yu, Entropies of fuzzy indiscernibility relation and its operations, International Journal of Uncertainty Fuzziness and Knowledge-Based Systems 12 (5) (2004) 575–589.
[22] Q.-H. Hu, D.-R. Yu, Z.-X. Xie, Information-preserving hybrid data reduction based on fuzzy-rough techniques, Pattern Recognition Letters 27 (5) (2006) 414–423.
[23] Q.-H. Hu, D.-R. Yu, Z.-X. Xie, J.-F. Liu, Fuzzy probabilistic approximation spaces and their information measures, IEEE Transactions on Fuzzy Systems 14 (2) (2006) 191–201.
[24] N. Japkowicz, Learning from imbalanced data sets: a comparison of various strategies, in: N. Japkowicz (Eds.), Proc. AAAI'00 Workshop on Learning from Imbalanced Data Sets, Technical Report WS-00-05, AAAI Press, Menlo Park, CA, 2000, pp. 10–15.
[25] N. Japkowicz (Eds.), Proc. AAAI'00 Workshop on Learning from Imbalanced Data Sets, Technical Report WS-00-05, AAAI Press, Menlo Park, CA, 2000.
[26] N. Japkowicz, S. Stephen, The class imbalance problem: A systematic study, Intelligent Data Analysis 6 (5) (2002) 429–450.
[27] M. Kryszkiewicz, Rules in incomplete information systems, Information Sciences 113 (3–4) (1999) 271–292.
[28] M. Kryszkiewicz, Comparative study of alternative type of knowledge reduction in inconsistent systems, International Journal of Intelligent Systems 16 (2001) 105–120.
[29] J.Y. Liang, Z.B. Xu, The algorithm on knowledge reduction in incomplete information systems, International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 10 (2002) 95–103.

[30] H. Liu, F. Hussain, C.L. Tan, M. Dash, Discretization: an enabling technique, Data Mining and Knowledge Discovery 6 (2002) 393–423.
[31] T.-H. Ma, M.-L. Tang, Weighted rough set model, in: Y. Chen, A. Abraham (Eds.), Proc. 6th Int. Conf. Intelligent Systems Design and Applications, ISDA'06, Jinan, Shandong, China, 2006, pp. 481–485.
[32] M.A. Maloof, Learning when data sets are imbalanced and when costs are unequal and unknown, in: N. Chawla, N. Japkowicz, A. Kolcz (Eds.), Proc. ICML'03 Workshop on Learning from Imbalanced Data Sets (II), Washington DC, USA, 2003. Available from: <http://www.site.uottawa.ca/~nat/Workshop2003/imbalance03.tar.gz>.
[33] J.-S Mi, W.-Z. Wu, W.-X. Zhang, Approaches to knowledge reduction based on variable precision rough set model, Information Sciences 159 (2004) 255–272.
[34] R.S. Michalski, A theory and methodology of inductive learning, in: R.S. Michalski, J.G. Carbonell, T.M. Mitchell (Eds.), Machine Learning: an Artificial Intelligence Approach, Morgan Kaufmann, San Mateo, CA, 1983, pp. 83–134.
[35] Z. Pawlak, Rough sets, International Journal of Computer and Information Sciences 11 (1982) 341–356.
[36] Z. Pawlak, Rough Sets: Theoretical Aspects of Reasoning About Data, Kluwer Academic Publishers, Dordrecht, 1991.
[37] Z. Pawlak, Rough sets and intelligent data analysis, Information Sciences 147 (2002) 1–12.
[38] Z. Pawlak, J.W. Grzymala-Busse, R. Slowinski, W. Ziarko, Rough sets, Comm. ACM 38 (11) (1995) 89–95.
[39] Z. Pawlak, A. Skowron, Rough sets: some extensions, Information Sciences 177 (1) (2007) 28–40.
[40] Z. Pawlak, A. Skowron, Rudiments of rough sets, Information Sciences 177 (1) (2007) 3–27.
[41] Z. Pawlak, A. Skowron, Rough sets and boolean reasoning, Information Sciences 177 (1) (2007) 41–73.
[42] R.C. Prati, G.E.A.P.A. Batista, M.C. Monard, Class imbalances versus class overlapping: an analysis of a learning system behavior, in: R. Monroy, G. Arroyo, L.E. Sucar, H. Sossa (Eds.), Proc. 3rd Mexican Int. Conf. Artificial Intelligence, MICAI'04, LNAI 2972, Springer-Verlag, Heidelberg, 2004, pp. 312–321.
[43] F.J. Provost, T. Fawcett, Analysis and visualization of classifier performance: comparison under imprecise class and cost distributions, in: D. Heckerman, H. Mannila, D. Pregibon (Eds.), Proc. 3rd Int. Conf. Knowledge Discovery and Data Mining, KDD'97, AAAI Press, Menlo Park, CA, 1997, pp. 43–48.
[44] J.R. Quinlan, C4.5 Programs for Machine Learning, Morgan Kaufman, San Mateo, CA, 1993.
[45] G. Sakkis, I. Androutsopoulos, G. Paliouras, V. Karkaletsis, C. Spyropoulos, P. atamatopoulos, Stacking classifiers for anti-spam filtering of E-mail, in: L. Lee, D. Harman (Eds.), Proc. 6th Conf. Empirical Methods in Natural Language Processing, EMNLP'01, Carnegie Mellon University, Pittsburgh, PA, USA, 2001, pp. 44–50.
[46] Q. Shen, A. Chouchoulas, A rough-fuzzy approach for generating classification rules, Pattern Recognition 35 (11) (2002) 2425–2438.
[47] D. Slezak, Approximate reducts in decision tables, in: B. Bouchon-Meunier, M. Delgado, J.L. Verdegay, M.A. Vila, R.R. Yager (Eds.), Proc. 6th Int. Conf. Information Processing and Management of Uncertainty in Knowledge-based System, IPMU'96, Granada, Spain, 1996, pp. 1159–1164.
[48] D. Slezak, Approximate entropy reducts, Fundamenta Informaticae 53 (3–4) (2002) 365–390.
[49] J. Stefanowski, On rough set based approaches to induction of decision rules, in: L. Polkowski, A. Skowron (Eds.), Rough Sets in Knowledge Discovery, Physica-Verlag, Heidelberg, 1998, pp. 501–529.
[50] J. Stefanowski, S. Wilk, Rough sets for handling imbalanced data: combining filtering and rule-based classifiers, Fundamenta Informaticae 72 (1) (2006) 379–391.
[51] Q. Tao, G.-W. Wu, F.-Y. Wang, J. Wang, Posterior probability support vector machines for unbalanced data, IEEE Transactions on Neural Networks 16 (6) (2005) 1561–1573.
[52] K.M. Ting, An instance-weighting method to induce cost-sensitive trees, IEEE Transactions on Knowledge and Data Engineering 14 (3) (2002) 659–665.
[53] S. Tsumoto, Automated extraction of medical expert system rules from clinical databases based on rough set theory, Information Sciences 112 (1–4) (1998) 67–84.
[54] S. Tsumoto, Mining diagnostic rules from clinical databases using rough sets and medical diagnostic model, Information Sciences 162 (2) (2004) 65–80.
[55] G.Y. Wang, Rough reduction: in algebra view and information view, International Journal of Intelligent Systems 18 (6) (2003) 679–688.
[56] G.Y. Wang, J. Zhao, J.J. An, et al., A comparative study of algebra viewpoint and information viewpoint in attribute reduction, Fundamenta Informaticae 68 (3) (2005) 289–301.
[57] G.M. Weiss, Mining with rarity: problems and solutions: a unifying framework, SIGKDD Explorations 6 (1) (2004) 7–19.
[58] G.M. Weiss, F. Provost, The effect of class distribution on classifier learning: an empirical study, Technical Report ML-TR-44, Dept. of Computer Science, Rutgers University, New Brunswick, NJ, 2001.
[59] B. Zadrozny, C. Elkan, Learning and making decisions when costs and probabilities are both unknown, in: F. Provost, R. Srikant (Eds.), Proc. 7th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, KDD'01, San Francisco, CA, USA, 2001, pp. 204–213.
[60] Z.-H. Zhou, X.-Y. Liu, Training cost-sensitive neural networks with methods addressing the class imbalance problem, IEEE Transaction on Knowledge and Data Engineering 18 (1) (2006) 63–77.
[61] W. Ziarko, Variable precision rough set model, Journal of Computer and System Sciences 46 (1993) 39–59.