

Relationships Between Diversity of Classification Ensembles and Single-Class Performance Measures

Shuo Wang, *Student Member, IEEE*, and Xin Yao, *Fellow, IEEE*

Abstract—In class imbalance learning problems, how to better recognize examples from the minority class is the key focus, since it is usually more important and expensive than the majority class. Quite a few ensemble solutions have been proposed in the literature with varying degrees of success. It is generally believed that diversity in an ensemble could help to improve the performance of class imbalance learning. However, no study has actually investigated diversity in depth in terms of its definitions and effects in the context of class imbalance learning. It is unclear whether diversity will have a similar or different impact on the performance of minority and majority classes. In this paper, we aim to gain a deeper understanding of if and when ensemble diversity has a positive impact on the classification of imbalanced data sets. First, we explain when and why diversity measured by Q-statistic can bring improved overall accuracy based on two classification patterns proposed by Kuncheva et al. We define and give insights into good and bad patterns in imbalanced scenarios. Then, the pattern analysis is extended to single-class performance measures, including recall, precision and F-measure, which are widely used in class imbalance learning. Six different situations of diversity’s impact on these measures are obtained through theoretical analysis. Finally, to further understand how diversity affects the single class performance and overall performance in class imbalance problems, we carry out extensive experimental studies on both artificial data sets and real-world benchmarks with highly skewed class distributions. We find strong correlations between diversity and discussed performance measures. Diversity shows a positive impact on the minority class in general. It is also beneficial to the overall performance in terms of AUC and G-mean.

Index Terms—Class imbalance learning, ensemble learning, diversity, single-class performance measures, data mining.

1 INTRODUCTION

CLASS imbalance learning refers to a classification problem, where the data set presents skewed class distributions. For a typical imbalanced data set with two classes, one class is heavily under-represented compared to the other class that contains a relatively large number of examples. Class imbalance pervasively exists in many real-world applications, such as medical diagnosis [1], fraud detection [2], risk management [3], text classification [4], etc. Rare cases in these domains suffer from higher misclassification costs than common cases. It is a promising research area that has been drawing more and more attention in data mining and machine learning, since many standard machine learning algorithms have been reported to be less effective when dealing with this kind of problems [5] [6] [7] [3]. The fundamental issue to be resolved is that they tend to ignore or overfit the minority class. Hence, great research efforts have been made on the development of a good learning model that can predict rare cases more accurately to lower down the total risk.

In recent years, ensemble approaches have become

a popular way of advancing the classification of imbalanced data, because they can be easily adapted for emphasizing the minority class regions by rebalancing the training subset from the data level [8] [9] [10] or by applying different costs from the algorithm level [11] [12] [13] [14]. In addition, the idea of combining multiple classifiers itself can reduce the probability of overfitting [15]. Particular techniques, such as oversampling and undersampling, are often used with ensembles to improve generalization of predicting the minority class. They attempt to make use of the difference of individual classifiers for better performance [16] [17] [12]. However, no study has actually investigated its effects in classifying imbalanced data sets so far.

The difference of individual learners is interpreted as “diversity” in ensemble learning. It has been proved to be one of the main reasons for the success of ensembles from both theoretical and empirical aspects [18] [19] [20]. To date, existing studies have discussed the relationship between diversity and overall accuracy. In class imbalance cases, however, the overall accuracy is not appropriate and less meaningful [21]. Some single-class performance measures are defined to evaluate how a classifier performs in the specific class we are concerned with. Recall, precision and F-measure [22] are the most widely used single-class measures in the class imbalance

The authors are with the Centre of Excellence for Research in Computational Intelligence and Applications (CERCIA), School of Computer Science, The University of Birmingham, Edgbaston, Birmingham B15 2TT, UK.

E-mail: {S.Wang, X.Yao}@cs.bham.ac.uk

learning literature. Hence, some interesting questions are raised here: *what is the relationship between diversity and the single-class performance measures? In the presence of imbalanced data, is introducing diversity beneficial to the classification of the minority/majority class?* This paper explores these issues, which provides guidelines of in which condition and aspect diversity can improve the classification of a single class. If diversity is shown to be beneficial in imbalanced scenarios, it will suggest an alternative way of handling class imbalance problems by considering diversity explicitly in the learning process.

To answer the above questions, three subtopics are studied in this paper: 1) discuss when and why diversity measured by Q-statistic [23] causes better overall accuracy based on several classification patterns of ensembles by extending Kuncheva et al.'s study [24]. We explain why diversity is not always beneficial to the overall performance. Two arguments are proposed accordingly for the minority and majority classes of a class imbalance problem respectively. 2) The pattern analysis in the first subtopic is utilized to relate Q-statistic to single-class performance measures, including recall, precision and F-measure. We show mathematically how these single-class measures behave when diversity varies. Six possible situations with different changing behaviors are obtained and analyzed. 3) To further understand how diversity affects single-class and overall performance of imbalanced data, comprehensive experiments are carried out on artificial and real-world data sets with highly skewed class distributions. We find strong correlations between Q-statistic and discussed performance measures. Diversity shows a positive impact on the minority class in terms of recall and F-measure in general, which is achieved by making the ensemble produce broader and less overfitting classification boundaries for the minority class. Besides, diversity is beneficial to the overall performance measured by AUC and G-mean. This paper focuses on binary classification problems.

The rest of this paper is organized as follows. Section 2 presents the main work about diversity in classification ensembles and briefly introduces ensemble methods and evaluation criteria in class imbalance learning. Section 3 provides a detailed explanation of the relationship between Q-statistic and majority vote accuracy, based on Kuncheva et al.'s pattern analysis [24]. Section 4 extends the mathematical links in the patterns to the single-class context. Section 5 verifies the results obtained in the previous sections empirically. Section 6 draws the conclusions and points out future research.

2 RELATED BACKGROUND

In this section, we first review the major research about diversity in classification ensembles and explain why we choose Q-statistic as our diversity measure

in this paper. Then, we introduce frequently used ensemble methods and evaluation criteria in class imbalance learning, including the reasons for considering a single class.

2.1 Diversity in Classification Ensembles

Diversity of ensembles has been a hot topic during the past few years. It is commonly agreed that the success of ensemble is attributed to diversity – the degree of disagreement within an ensemble [18] [25]. In the regression context, it has already been quantified and measured explicitly in terms of the correlation between individual learners [20]. In the classification context, it is loosely described as “making errors on different examples” [19] [26]. Clearly, a set of identical classifiers does not bring any advantages. Ali and Pazzani [26] showed empirically that an ensemble making errors in a negatively correlated manner can produce lower error rates. Ko and Sabourin [27] exploited classification diversity for ensemble selection.

However, diversity of classification ensembles has not been fully understood yet. There is no agreed definition for diversity. Quite a few pairwise and non-pairwise diversity measures were proposed in the literature [28] [29] [30], such as Q-statistic [23], double-default measure [31], entropy [32], generalized diversity [33], etc. Kuncheva and Whitaker investigated the relationship between existing diversity measures [28]. Strong positive correlations were observed. Q-statistic was particularly recommended for its simplicity and understandability compared to others [28]. Besides, it showed little relationship to the individual accuracy [29] [28]. It is a desirable property for defining diversity, especially in class imbalance scenarios. A measure depending on individual accuracy could be sensitive to imbalanced distributions and thus cause misleading results. Therefore, we choose Q-statistic as our diversity measure in this paper.

Concretely, for a two-class data set $Z = \{z_1, \dots, z_N\}$, z_j is composed of (x_j, y_j) , where y_j is the true label of example x_j and $y_j \in \{+1, -1\}$. For each classifier f_i ($i = 1, \dots, L$), we define $y_{j,i} = 1$ if f_i recognizes z_j correctly, and 0 otherwise. Similarly, we define $y_{j,ens} = 1$ if the combined model of f_1 to f_L through majority voting gives the correct label for x_j , and 0 if x_j is misclassified. Q-statistic, denoted by Q in this paper, is a pairwise similarity measure. For two classifiers f_i and f_k , it is defined as

$$Q_{i,k} = \frac{N^{11}N^{00} - N^{01}N^{10}}{N^{11}N^{00} + N^{01}N^{10}}, \quad (1)$$

where N^{ab} is the number of z_j of Z for which $y_{j,i} = a$ and $y_{j,k} = b$. Q is the average over all pairs. Higher Q-statistic indicates smaller diversity, i.e. larger similarity among classifiers.

The theoretical relationship between diversity and overall accuracy has been studied in two directions with different assumptions. When the output

of each classifier is expected to approximate the corresponding posteriori class probabilities, Tumer and Ghosh [34] [35] reformulated the error rate above the Bayes error of an ensemble involving a correlation factor. It suggests that reducing the correlation among the individual learners can bring a performance improvement. When the majority voting strategy is applied to where crisp class labels are produced, Kuncheva et al. [24] derived mathematical links between diversity measured by Q-statistic and overall accuracy under two classification patterns. They found that increasing diversity is not always beneficial. In the “best” pattern, reducing Q-statistic (i.e. increasing diversity) results in better accuracy; in the “worst” pattern, it results in worse accuracy. Further details of the patterns will be given in the next section. To answer our research question of the relationship between diversity and single-class performance, we attempt to build links between Q-statistic and single-class measures based on this pattern analysis.

2.2 Class Imbalance Learning

Finding minority class examples effectively and accurately without losing overall performance is the objective of class imbalance learning. The fundamental issue to be resolved is that the classification ability of most standard learning algorithms is significantly compromised by imbalanced class distributions. They often give high overall accuracy, but form very specific rules and exhibit poor generalization for the small class. In other words, overfitting happens to the minority class [6] [36] [37] [38] [39]. Correspondingly, the majority class is often over-generalized. Particular attention is necessary for each class. It is important to know if a performance improvement happens to both classes or just one class alone. Hence, this paper looks into single-class performance.

2.2.1 Ensemble Methods

Ensemble learning methods have become a major category of solutions for class imbalance learning, due to their flexibility and ability to improve generalization. First, an ensemble method is applicable to most classification algorithms. Second, it’s easy to combine with resampling techniques. Third, combining multiple classifiers is able to reduce the error bias/variance [40]. These attractive features lead to a variety of ensemble methods proposed to handle imbalanced data sets from the data and algorithm levels.

From the data level, sampling strategies are integrated into the training of each ensemble member. For instance, Li’s BEV [8] and Chan et al.’s combining model [41] were proposed based on the idea of Bagging [42] by undersampling the majority class examples and combining them with all the minority class examples to form balanced training subsets. SMOTEBoost [10] and DataBoost-IM [13] were

designed to alter the imbalanced distribution based on Boosting [43]. Data generation techniques are involved to emphasize the minority class examples at each iteration of Boosting. Particularly, SMOTEBoost is claimed to increase ensemble diversity.

Solutions from the algorithm level mainly include cost-sensitive methods used with Boosting, such as AdaCost [14], CSB [44], and RareBoost [11]. They modify the weight updating rule to assign higher weights to examples with larger misclassification costs. AdaCost and CSB require explicit cost items prior to learning.

To overcome the issue of overfitting small data regions, existing ensemble methods attempt to enforce the difference among the individual classifiers through different techniques [21] [45], which is concerned with ensemble diversity.

2.2.2 Performance Evaluation

Overall accuracy becomes meaningless when the learning concern is how to find rare examples effectively [21]. Other performance criteria must be considered. Because the classification interest is often the minority class, single-class measures are used to show how a classifier performs in this class. Recall, precision and F-measure [22] are the commonly accepted ones in the class imbalance learning community and discussed the most in related papers.

TABLE 1: Confusion Matrix

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

Given the confusion matrix in Table 1, the single-class measures for the positive class are defined as

$$R = \frac{TP}{TP + FN}, \quad (2)$$

$$P = \frac{TP}{TP + FP}, \quad (3)$$

$$F = \frac{(1 + \rho^2) \cdot R \cdot P}{\rho^2 \cdot R + P}, \quad (4)$$

where ‘R’, ‘P’ and ‘F’ denote recall, precision and F-measure respectively. ρ corresponds to the relative importance of precision and recall, which is usually set to 1. Recall is a measure of completeness – how many examples are classified correctly of all positive class examples. Precision is a measure of exactness – how many examples have the true positive label of all examples predicted as positive. They were claimed to be effective in evaluating classification performance in imbalanced learning scenarios [39]. F-measure incorporates both recall and precision due to the trade-off between them. It was shown to be a more favorable

measure [46] and used as a criterion for classifier selection [12] [47].

In addition to single-class performance measures, G-mean [48] [49] and AUC [50] [51] [52] are overall performance criteria widely used in class imbalance learning. They assess how well a classifier can balance the performance between classes. They are better indicators to show the performance trade-off between classes than overall accuracy for their insensitivity to class distributions [39]. G-mean is the geometric mean of recalls of the minority and majority classes. It shows to what extent accuracy on the majority class drops with the error reduction on the minority class. AUC is the area under the ROC curve, a two-dimensional graph between TP rate and FP rate. The best algorithm should produce the dominant curve, which also has the largest AUC. It has been proved that AUC is statistically consistent and more discriminating than accuracy [51].

3 RELATIONSHIP BETWEEN DIVERSITY AND OVERALL ACCURACY

In this section we explain the functional relationship between ensemble diversity and overall accuracy in several patterns by extending Kuncheva et al.'s study [24]. A classification pattern refers to the voting combinations of the individual classifiers that an ensemble can have. The accuracy is given by the majority voting method of combining classifier decisions. First, two extreme patterns are defined, which present different effects of diversity. It is shown that diversity is not always beneficial to the generalization performance. The reason is then explained in a general pattern. According to the features of the patterns, we relate them to the classification of each class of a class imbalance problem, and propose two arguments for the minority and majority classes respectively.

3.1 Notations

Here we give some additional notations following the symbols defined in section 2. For an ensemble formed by L classifiers, L is restricted to an odd number for the convenience of calculation later. Let $l = \lfloor L/2 \rfloor$. We define,

- P_{ovr} : the overall accuracy of the ensemble with the combination method of majority vote. Every classifier produces a class label and they equally contribute to the final output.
- p : the overall accuracy of the individual classifier. A classifier has probability p of giving the correct label to any input. In Kuncheva et al.'s two extreme patterns [24], L classifiers are assumed to have the same individual accuracy. All the patterns discussed in this paper are based on this assumption.
- p_{ab} : the probability of occurrence of the respective combination of correct and wrong outputs for any

pair of classifiers f_i and f_k . 'a' and 'b' belong to $\{1$ (correct label), 0 (incorrect label) $\}$.

Supposing data set Z with N examples (cardinality $|Z| = N$) is imbalanced, we let Z_{min} represent the subset of Z only containing examples with the minority class label. By convention, we treat the positive class "+1" as the minority class. The rest of the examples with label "-1" form the majority-class subset $Z_{maj} \subset Z$. We have $Z_{min} \cap Z_{maj} = \emptyset$ and $Z_{min} \cup Z_{maj} = Z$. The imbalance rate θ is defined as $|Z_{min}|/|Z|$. As a special case, a balanced data set has θ equal to 0.5. In the following, we use a "min" subscript to denote a single-class measure evaluated for the minority class, and a "maj" subscript to denote "majority". For example, R_{min} and R_{maj} indicate the recall measure of the minority class and majority class respectively.

3.2 Q-statistic and Overall Accuracy in Patterns

The relationship between Q-statistic (Q) and overall accuracy (P_{ovr}) has been discussed in two extreme situations, where explicit mathematical links between Q and P_{ovr} exist and Q exhibits a different impact on P_{ovr} [24]. To provide a deeper understanding of why accuracy behaves differently along with Q , the two extreme cases are analyzed in a more general context. Then, a good pattern and a bad pattern are defined accordingly.

3.2.1 Two Extreme Patterns

Kuncheva et al. defined and analyzed two probability distributions over the possible combinations of L votes from the ensemble members, referred to as "pattern of success" and "pattern of failure" [24]. The two patterns were claimed to have the possible characteristics of the "best" and "worst" combinations of L classifiers when they hold the same accuracy p .

(a) Pattern of success (best pattern): In this pattern, no correct votes are "wasted", which means there are exactly $l+1$ classifiers giving correct answers to every correctly predicted example, or all of the classifiers give wrong answers if the combined result is wrong. Any extra correct vote will be a waste. To give a formal definition [24],

- The probability of any combination with $l+1$ correct and l incorrect votes is α .
- The probability of all L votes being incorrect is β .
- The probability of all other combinations is 0.

Theorem 1. Under the best pattern, the expression of P_{ovr} with respect to Q is [24]

$$P_{ovr} = \frac{L}{(l+1)} \frac{(1-Q)}{(2-Q)}. \quad (5)$$

P_{ovr} is a monotone decreasing function of the pairwise dependence Q . For any $p > 2/3$, the value of Q will be -1 .

(b) Pattern of failure (worse pattern): In this pattern, correct votes are “wasted” to the maximum extent. All the classifiers give correct answers if the combined result is correct. Otherwise, there are exactly l correct votes, which produce the wrong class label. It is defined as [24],

- The probability of all L votes being correct is α .
- The probability of any combination with l correct and $l + 1$ incorrect votes is β .
- The probability of all other combinations is zero.

Theorem 2. *Under the worst pattern, the expression of P_{ovr} with respect to Q is [24]*

$$P_{ovr} = \frac{L}{(l-1)(2-Q)} - \frac{l}{(L-l)}, \quad (6)$$

P_{ovr} is a monotone increasing function of Q . For any $p > 0.5$, the value of Q is positive.

The patterns show that diversity does not always produce a positive effect on ensembles, which depends on the number of wasted correct votes from the individual classifiers. When diversity is “good”, such as the best pattern, diversity brings a performance improvement. Otherwise, there is some “bad” diversity that harms the performance. The worst pattern is one of such cases. It is a useful analysis since it finds when diversity can be beneficial or harmful, and provides the possible factor that is somehow related to the effect of diversity. However, the original paper did not explain why the number of wasted correct votes with a corresponding voting pattern of the ensemble could discriminate ensembles with different impacts of diversity.

Let’s take a closer look of those patterns and corresponding functions. In fact, the number of wasted correct votes is related to which voting combination low diversity resides in. α and β decide the ensemble accuracy. In the best pattern, low diversity exists in the voting combination that misclassifies examples, in which case each classifier gives the wrong label. Increasing diversity makes the wrong voting combination causing the low disagreement degree less likely to happen and enforces the correct voting combination taking the high disagreement degree (i.e. higher α and lower β). Therefore, the overall accuracy gets improved. On the contrary, low diversity results from the correct voting combination in the worst pattern. Encouraging diversity increases the probability of occurrence of the wrong voting combination β , and thus worse accuracy is obtained. This may provide us with some clues of the connection between the definition of the patterns and the impact of diversity.

3.2.2 General Pattern

To further understand the patterns in a general sense, we define a general pattern here by allowing more than two possible voting combinations of an ensemble. Although it is hard to derive a neat expression

for Q -statistic and accuracy under this pattern, it shows how they are related by the types of voting combinations and their possibilities. It is defined as follows,

- Given a voting combination that provides the correct class label for an input, the number of correct votes is $l + 1 + i$ with probability α_i ($i = 0, 1, \dots, l$).
- Given a voting combination that provides the incorrect class label for an input, the number of correct votes is j with probability β_j ($j = 0, 1, \dots, l$).

α_i and β_j determine the voting distribution of the ensemble. Their subscripts i and j indicate the number of wasted correct votes in the corresponding voting combination. We apply the same inferring method as in the extreme patterns [24]. P_{ovr} is the sum of the probability of each correct voting combination. There are $\binom{L}{l+1+i}$ ways of having $(l + 1 + i)$ correct out of L classifiers, each with probability α_i . The majority vote accuracy becomes

$$P_{ovr} = \sum_{i=0}^l \binom{L}{l+1+i} \alpha_i \quad (7)$$

with condition $\sum_{i=0}^l \binom{L}{l+1+i} \alpha_i + \sum_{j=0}^l \binom{L}{j} \beta_j = 1$.

TABLE 2: Pairwise table for the general pattern.

f_i	f_k	
	1	0
1	$p_{11} = \sum_{i=0}^l \binom{L-2}{l-1+i} \alpha_i + \sum_{j=2}^l \binom{L-2}{j-2} \beta_j$	$p_{10} = \sum_{i=0}^{l-1} \binom{L-2}{l+i} \alpha_i + \sum_{j=1}^l \binom{L-2}{j-1} \beta_j$
0	$p_{01} = \sum_{i=0}^{l-1} \binom{L-2}{l+i} \alpha_i + \sum_{j=1}^l \binom{L-2}{j-1} \beta_j$	$p_{00} = \sum_{i=0}^{l-2} \binom{L-2}{l+1+i} \alpha_i + \sum_{j=0}^l \binom{L-2}{j} \beta_j$

The relationship between any two classifiers f_i and f_k can be visualized using a pairwise table. In Table 2, the probability of every correct/incorrect combination is presented, where “1” stands for the correct vote and “0” stands for the incorrect vote. The entries in the table are obtained by following combinatorial reasoning. For example, when both f_i and f_k are correct, the remaining $(L - 2)$ classifiers either give $(l - 1 + i)$ correct votes with $\binom{L-2}{l-1+i}$ ways if the majority vote is correct, or give $(j - 2)$ correct votes with $\binom{L-2}{j-2}$ ways if the majority vote is wrong. Thus, the probability of having f_i and f_k both correct is $\sum_{i=0}^l \binom{L-2}{l-1+i} \alpha_i + \sum_{j=2}^l \binom{L-2}{j-2} \beta_j$. By definition, the pattern is symmetrical with respect to all classifiers, so that all pairs of individual classifiers have the same pairwise tables, and therefore the same Q . According to Eq. 1, Q -statistic can be computed by substituting the four probabilities in Eq. 8 with the expressions in Table 2,

$$Q = \frac{p_{11}p_{00} - p_{01}p_{10}}{p_{11}p_{00} + p_{01}p_{10}}. \quad (8)$$

Due to too many α and β terms contained in Q , we derive an upper bound of Q to simplify our discussions here.

Theorem 3. *Q-statistic is upper bounded by the monotone decreasing functions of p_{01} (p_{10}) under the assumption of individual classifiers having the same accuracy p .*

Proof: Due to the same individual accuracy p , $p_{01} = p_{10}$ holds. Q-statistic can be expressed as

$$Q = \frac{p_{00}p_{11} - p_{01}^2}{p_{00}p_{11} + p_{01}^2}.$$

Because the inequality $p_{00}p_{11} \leq \left(\frac{p_{00}+p_{11}}{2}\right)^2$ holds and the four probabilities satisfy $p_{00} + p_{11} + p_{01} + p_{10} = 1$ ($p_{01} < 1/2$), we obtain

$$Q \leq \begin{cases} \frac{1-4p_{01}}{4p_{01}^2} & \text{if } Q \geq 0 \\ \frac{1-4p_{01}}{4p_{01}^2 + (1-2p_{01})^2} & \text{if } Q < 0 \end{cases}$$

by eliminating p_{00} and p_{11} . The right-hand side functions of the inequality are monotone decreasing with respect to p_{01} in $[0, 1/2]$, which upper-bound Q . \square

It is easy to understand Theorem 3 intuitively. It tells us that increasing the probability of making different decisions of pairs of classifiers can make the ensemble more diverse. Now, we just concentrate on the expression of p_{01} reformulated in Eq. 9 instead of Q , and see how it relates to P_{ovr} through the voting distribution of the ensemble.

$$p_{01} = \left(\underbrace{\binom{L-2}{l-1}\alpha_0 + \binom{L-2}{l-2}\alpha_1 \dots + \binom{L-2}{0}\alpha_{l-1}}_{\text{best}} \right) + \left(\underbrace{\binom{L-2}{l-1}\beta_l + \binom{L-2}{l-2}\beta_{l-1} \dots + \binom{L-2}{0}\beta_1}_{\text{worst}} \right) \quad (9)$$

Now we consider the best and worst patterns from the view of the general pattern. According to the definition of the best pattern, p_{01} is reduced to $\binom{L-2}{l-1}\alpha_0$ as denoted in Eq. 9. Increasing p_{01} means raising the probability of making the correct prediction of the ensemble, in that only a α term exists. Similarly, $p_{01} = \binom{L-2}{l-1}\beta_l$ in the worst pattern. Increasing p_{01} leads to a higher probability of making the wrong prediction, and thus worse overall accuracy. *What does it suggest?* If α terms dominate p_{01} , then P_{ovr} is positively related to p_{01} , which means that increasing diversity brings a performance improvement. Otherwise, P_{ovr} is negatively related to p_{01} by β terms and higher diversity reduces the accuracy. Accordingly, we infer that the number of wasted correct votes of the ensemble determines which types of voting combinations would be affected more by diversity. A small amount of waste tends to make diversity have greater

positive influence on α -type voting combinations, i.e. correct ensemble outputs. The analysis here will help us to understand the impact of diversity on single-class performance for class imbalance problems next.

Before diverting our attention, we clearly define good and bad patterns for the following discussions. In general, if increasing diversity causes higher α 's and lower β 's (i.e. better accuracy), we say that the ensemble has a good voting pattern; if lower α 's and higher β 's happen (i.e. worse accuracy), it has a bad voting pattern.

3.3 Patterns and Class Imbalance Learning

In this section, we take the classification characteristics of class imbalance learning into account. We first give some insight into the class imbalance problem from the view of base learning algorithms, such as decision trees and neural networks. Skewed class distributions and different misclassification costs make the classification difficulty mainly reflect in the overfitting to the minority class and the over-generalization to the majority class, because the small class has less contribution to the classifier. Both decision trees and neural networks have been reported to be biased toward the majority concept inherently [39] [5] [7]. A tree learner can result in very specific branches for the minority class that cover very few training examples. A neural network cannot learn the minority class sufficiently, because the majority class examples overwhelm the minimization procedure of the squared error using gradient descent.

Considering an ensemble composed of many of such classifiers, each classifier tends to label most of the data as the majority class. We can imagine that the ensemble has very low diversity. As an extreme situation, all individuals misclassify any minority class example and assign the majority class label to all examples. Based on this understanding, we propose two arguments for each class from the view of patterns.

3.3.1 Minority Class and Good Pattern

Recall the general pattern in the previous section, where Q-statistic and the overall accuracy are linked through p_{01} expressed by α and β terms. For the minority class, each individual classifier has a low recognition rate. It corresponds to large β and small α values in the general pattern. Moreover, very little disagreement degree among them suggests that the number of wasted correct votes is small. Increasing p_{01} in this situation is prone to cause larger α 's and smaller β 's. We hence argue that the ensemble tends to have a good pattern over the minority class, where considering diversity can improve its classification accuracy.

3.3.2 Majority Class and Bad Pattern

The majority class contains sufficient data information for learners. Consequently, every individual tends to

make the same correct decision. Referring to the general pattern, it means large α and small β values. Besides, a lot of wasted correct votes from the ensemble exist. Increasing p_{01} in this situation is likely to reduce α 's and increase β 's. In this regard, the ensemble tends to behave in a bad pattern over the majority class, where diversity deteriorates the accuracy.

It is worth noting that the accuracy here is in the context of a single class. The two arguments reflect the fact of classifying an imbalanced data associated with the classification patterns of the ensemble. Different effects of diversity between classes are expected in imbalanced scenarios. The empirical evidence of how diversity affects the performance in each class will be given in section 5.

4 RELATIONSHIP BETWEEN DIVERSITY AND SINGLE-CLASS MEASURES

The relationship in patterns discussed earlier is concerned with the overall accuracy of majority vote. We extend it to single-class performance in this section. Three single-class measures are considered as suggested in section 2.2.2: recall, precision and F-measure. We will show how they behave in two extreme patterns as Q-statistic varies. Six possible situations are obtained through the mathematical analysis. It is worth mentioning that subscripts "min" and "maj" appearing in this section simply stand for two different classes of a data set in general without discriminating the class size, so the obtained results of the impact of diversity on single-class performance are applicable to both classes here.

For any $z_j \in Z$, two probabilities can be approximated: $p\{z_j \in Z_{min}\} = \theta$ and $p\{z_j \in Z_{maj}\} = 1 - \theta$, where θ is the proportion of Z_{min} in Z . They are constant values for a specific problem. We denote two additional probabilities as follows,

$$p_{tp} = p\{y_{j,ens} = 1 \cap z_j \in Z_{min}\}, \quad (10)$$

$$p_{fp} = p\{y_{j,ens} = 0 \cap z_j \in Z_{maj}\}. \quad (11)$$

p_{tp} indicates the probability of any minority class example predicted correctly by the ensemble. p_{fp} is the probability of any majority class example misclassified by the ensemble. According to the definitions of recall and precision, they can be expressed by

$$R_{min} = \frac{p_{tp}}{\theta}, \quad (12)$$

$$P_{min} = \frac{p_{tp}}{p_{tp} + p_{fp}}. \quad (13)$$

4.1 Recall, Precision and F-measure with Independence Assumption

We first consider a simple case. We assume that, whether an input z_j is classified correctly by the ensemble is independent of its real class label. Under

the assumption, Eq. 10 and Eq. 11 can be simplified into

$$p_{tp} = \theta P_{ovr}, \quad (14)$$

$$p_{fp} = (1 - \theta)(1 - P_{ovr}). \quad (15)$$

By substituting them into Eq. 12 and Eq. 13, we obtain

$$R_{min} = P_{ovr}, \quad (16)$$

$$P_{min} = \frac{\theta P_{ovr}}{\theta P_{ovr} + (1 - \theta)(1 - P_{ovr})}. \quad (17)$$

Eq. 16 and Eq. 17 only contain P_{ovr} and the constant θ . By substituting P_{ovr} with the monotonic functions in the best and worst patterns, we obtain clear links of recall and precision to Q-statistic. The recall measure has the same functional relation as P_{ovr} , which is monotone decreasing with respect to Q in the best pattern and monotone increasing in the worst pattern. As to precision in the best pattern,

$$P_{min} = \frac{-Q(L\theta) + L\theta}{-Q(3l\theta + \theta - l) + (2l\theta + 1)}. \quad (18)$$

It is a monotone decreasing function of Q , because its derivative is always negative. In other words, increasing diversity can improve both recall and precision in the best pattern, and lead to a better F-measure.

In the worst pattern,

$$P_{min} = \frac{QA + B}{QC + D}, \quad (19)$$

where the four constants A, B, C and D in the equation are

$$A = (l\theta)(l - 1), B = \theta(5l + 1),$$

$$C = (3l\theta + \theta - L)(l - 1),$$

$$D = 2l^2 - 2l^2\theta + 10l\theta - 5l + 4\theta - 3.$$

It is a monotone increasing function of Q . R_{min} and P_{min} will increase as Q increases in the worst pattern. Therefore, enforcing diversity leads to the reduction of recall, precision and F-measure in this case.

In summary, Q-statistic has the same impact on the three single-class measures under the independence assumption. Recall, precision and F-measure get improved or reduced simultaneously depending on "good" or "bad" Q . For class imbalance problems, however, this assumption is clearly untrue, since the minority class data are much harder to be classified correctly in general than the data belonging to the majority class. As the data set gets less imbalanced, the dependence between the class label and misclassification error should get smaller. To certain extent, we can regard this part of discussions to be suitable for balanced data sets.

4.2 Recall, Precision and F-measure without Independence Assumption

Without the above assumption, it is not easy to get such neat and separate expressions for those single-class measures. They are related to each other and can behave in different ways. Multiple situations must be considered. The overall accuracy (P_{ovr}) is utilized to associate Q-statistic with single-class measures here. P_{ovr} can be re-expressed by every two single-class measures according to their definitions through some mathematical transformations:

$$P_{ovr} = \frac{TP + TN}{|Z|} = \frac{|Z_{min}| \cdot R_{min} + |Z_{maj}| \cdot R_{maj}}{|Z|}. \quad (20)$$

Two single-class measures in Eq. 20 are from different classes. To express P_{ovr} with measures from the same class, we use Eq. 2 divided by Eq. 3,

$$\frac{R_{min}}{P_{min}} = R_{min} + \frac{|Z_{maj}|}{|Z_{min}|} (1 - R_{maj}). \quad (21)$$

Eq. 21 presents the relationship among R_{min} , P_{min} and R_{maj} . By eliminating R_{maj} from Eq. 20 and Eq. 21, we obtain

$$P_{ovr} = (1 - \theta) + \theta \left(2 - \frac{1}{P_{min}} \right) R_{min}. \quad (22)$$

Similarly,

$$P_{ovr} = 1 + 2\theta R_{min} \left(1 - \frac{1}{F_{min}} \right) \quad (23)$$

$$P_{ovr} = (1 - \theta) + \theta \left(\frac{2P_{min} - 1}{\frac{2P_{min}}{F_{min}} - 1} \right). \quad (24)$$

P_{ovr} is expressed by every two single-class measures from one class. It is monotone increasing with respect to $R_{min}/P_{min}/F_{min}$ by calculating the derivatives. The changing behaviors of the measures along with Q-statistic are concluded in Table 3 according to the above functional relations, including six possible situations in good and bad patterns. In situation (3), for example, if the ensemble performs in a good pattern (i.e. decreasing Q leads to better P_{ovr}) and R_{min} decreases, then P_{min} will increase due to Eq. 22 and F_{min} will increase due to Eq. 23. Any two single-class measures from the same class will not get worse simultaneously in the good pattern. Analogously, any two single-class measures from the same class will not increase simultaneously in the bad pattern. Table 3 is also applicable to $R_{maj}/P_{maj}/F_{maj}$. The mathematical functions of the three single-class measures obtained in section 4.1 fall into situations 1 and 6.

We have analyzed the impact of diversity on overall accuracy and extended it to single-class performance by providing mathematical links so far. Next, we would ask following questions: *how does diversity affect the classification performance of the minority and majority*

TABLE 3: The possible behaviors of P_{ovr} , R_{min} , P_{min} and F_{min} as Q-statistic decreases (i.e. increasing diversity) in “good” and “bad” patterns in terms of P_{ovr} .

$Q \downarrow$	P_{ovr}	R_{min}	P_{min}	F_{min}	Situation
Overall good pattern	\uparrow	\uparrow	\uparrow	\uparrow	(1)
		\downarrow	\uparrow	\uparrow	(2)
		\downarrow	\uparrow	\uparrow	(3)
Overall bad pattern	\downarrow	\uparrow	\uparrow	\downarrow	(4)
		\downarrow	\uparrow	\downarrow	(5)
		\downarrow	\downarrow	\downarrow	(6)

classes in real imbalanced scenarios? Which situation in Table 3 does the ensemble have over each class? According to our arguments in section 3.3, the impact of diversity on the minority and majority classes should be different. We anticipate that diversity is beneficial to the recognition of minority class examples, but can cause negative effect on the majority class.

5 DIVERSITY IN CLASS IMBALANCE LEARNING

In this section, we examine the relationship between diversity and the classification performance empirically in class imbalance scenarios, in order to verify the obtained results so far. Artificial data sets and highly imbalanced real-world benchmarks are included in our experiments.

5.1 Impact of Diversity on Classification Performance on Artificial Imbalanced Data

To clearly observe the impact of diversity on balanced and imbalanced data sets and investigate how the performance measures behave along with the diversity degree in depth, we build ensembles on several two-dimensional artificial data sets with different degrees of class imbalance. We proceed with correlation analysis and present corresponding decision boundary plots. We also provide some insight into diversity and performance measures at different levels of ensemble size.

5.1.1 Experimental Setup

The artificial data is generated from two Gaussian distributions with equal covariance and a small overlapping area close to the separating line. Three different data sizes are considered: while one class always contains 200 training points, the size of the other class is set to 10 (very imbalanced), 50 (imbalanced) and 200 (balanced) respectively. They are denoted by “200-10”, “200-50” and “200-200”. By applying the same generation method, a corresponding testing file is created with 50 points in each class.

As we know, Bagging [42] achieves diversity through resampling, where each training subset is kept different from each other. In our experiments, we apply the Bagging training strategy and manipulate

diversity by tuning the sampling rate $r\%$. A smaller r means that fewer points join the training. Each training subset is less likely to be similar to each other. Thus, the prediction becomes less stable and a larger diversity degree is expected [53]. Different from the conventional Bagging, a different sampling rate is applied to each class of the imbalanced data set in our experiments. Concretely, the majority class is randomly sampled with replacement at rate $r\%$; the minority class is randomly sampled with replacement at rate $r(1-\theta)/\theta\%$. θ is the imbalance rate as defined before. By doing so, every training subset has a balanced class distribution. This is to avoid the minority class being ignored by the classifier and affecting our experimental analysis. Especially when the training data is highly skewed, some preliminary experiments showed that the recall of the minority class always stays 0 as r varies if the training subset is not rebalanced. It means that no minority class examples are recognized. The training strategy is described in Table 4. C4.5 decision trees are used as the base learner. A simple majority vote gives the final decision.

TABLE 4: Bagging-based strategy.

<p>Training:</p> <ol style="list-style-type: none"> Given the training set Z with imbalance rate θ; resampling rate at $r\%$; number of classifiers L. Z_{min} is the minority class subset of Z. Z_{maj} is the majority class subset of Z. Construct subset Z_k by executing the following: <ol style="list-style-type: none"> Booststrap Z_{maj} at rate $r\%$ and add chosen examples into Z_k; Booststrap Z_{min} at rate $r(1-\theta)/\theta\%$ and add chosen examples into Z_k. Train a classifier using Z_k. Repeat steps 2 and 3 until k equals L. <p>Testing on a new example: (majority voting)</p> <ol style="list-style-type: none"> Collect decisions from each classifier. Return the class label that receives the most votes.

The training method is run 50 times for each setting and outputs the averages. The sampling rate $r\%$ is varied in the range of [3%, 1000%]. Every ensemble consists of 15 classifiers. For the following analysis, we compute 12 measures on the test data, including overall Q-statistic (Q), Q-statistic (Q_{min}/Q_{maj}), recall (R_{min}/R_{maj}), precision (P_{min}/P_{maj}), F-measure (F_{min}/F_{maj}) of each class, overall accuracy (P_{ovr}), G-mean and AUC. Q_{min} and Q_{maj} assess the diversity degree of an ensemble only within the minority and majority data subsets respectively.

5.1.2 Correlation Analysis

We conduct correlation analysis by computing Spearman's rank correlation coefficient between overall diversity Q and the other measures. The Spearman correlation coefficient is a non-parametric measure of statistical dependence between two variables, and

insensitive to how the measures are scaled. It ranges in $[-1, 1]$, where 1 (or -1) indicates a perfect monotone increasing (or decreasing) relationship. Table 5 presents the correlation coefficients of the single-class performance measures and the overall accuracy in two sampling ranges of r . The numbers in boldface indicate significant correlations at confidence level of 95%.

Between Q and Q_{min}/Q_{maj} , all coefficients from the three data sets are positive, which shows that ensemble diversity for each class has the same changing tendency as the overall diversity, regardless of whether the data set is balanced. On one hand, it guarantees that increasing the classification diversity over the whole data set can increase diversity over each class. On the other hand, it confirms that the diversity measure Q-statistic is not sensitive to imbalanced distributions.

For the balanced data set "200-200", the overall accuracy and most single-class measures do not present clear correlations with Q when $r \in [3, 100]$. In this range, increasing diversity does not necessarily lead to better performance. When r varies in [100, 1000], all 7 performance measures have strong negative correlations with Q , which suggests that diversity is beneficial to both classes and the overall accuracy. This observation agrees with the mathematical relations obtained in the relationship analysis with the independence assumption in section 4.1. It corresponds to the situation 1 in Table 3. In this higher range with relatively small ensemble diversity, increasing diversity improves the classification performance of both classes in the balanced case.

For imbalanced data sets "200-50" and "200-10", overall accuracy P_{ovr} gets higher in both ranges of r as the ensemble becomes more diverse. Single-class measures behave differently between classes. R_{min} and F_{min} have significant negative correlations with Q ; P_{min} has a significant positive correlation with Q . It implies that increasing diversity can find more minority class examples (i.e. better recall) but lose some classification precision. As the trade-off between recall and precision, better F-measure indicates that the improvement of recall is greater than the reduction of precision. The observation corresponds to the situation 2 in Table 3. As to the majority class, R_{maj} has a significant positive correlation with Q ; P_{maj} and F_{maj} have significant negative correlations with Q . It means that the majority-class recall gets smaller along with the increase of diversity, but precision and F-measure are improved. The measure behaviors of the majority class correspond to the situation 3 in Table 3.

Based on the results of improved F-measure of both classes and better minority-class recall by increasing ensemble diversity, we can conclude that diversity has a positive effect in classifying imbalanced data sets in general. In other words, the performance of both classes is better balanced between recall and preci-

TABLE 5: Rank correlation coefficients (in %) between overall diversity Q and single-class performance measures plus P_{ovr} for 3 artificial data sets. Numbers in boldface indicate significant correlations.

$r \in [3, 100]$	Q_{min}	Q_{maj}	R_{min}	P_{min}	F_{min}	R_{maj}	P_{maj}	F_{maj}	P_{ovr}
200-200	93	52	27	-54	-9	-50	26	-22	-16
200-50	97	98	-98	54	-98	58	-98	-98	-98
200-10	6	98	-38	33	-36	33	-36	-36	-36
$r \in [100, 1000]$	Q_{min}	Q_{maj}	R_{min}	P_{min}	F_{min}	R_{maj}	P_{maj}	F_{maj}	P_{ovr}
200-200	98	89	-79	-89	-98	-81	-88	-98	-96
200-50	100	93	-81	58	-81	58	-81	-81	-81
200-10	99	27	-81	77	-61	77	-55	-55	-61

sion with more minority class examples identified. Moreover, the different behavior of recall between the minority and majority classes agrees with the two arguments in section 3.3. Since recall is essentially the accuracy over a single class, we can say that diversity is beneficial to the accuracy of the minority class, whereas it is harmful to the accuracy of the majority class.

Table 6 summarizes the measure tendencies for both balanced and imbalanced cases along with the decrease of Q (i.e. the increase of ensemble diversity) and their corresponding situations in Table 3.

TABLE 6: Behaviors of single-class measures and overall accuracy by decreasing Q -statistic (more diverse) on artificial data sets and their corresponding situations in Table 3.

Class	P_{ovr}	R	P	F	Situation
Single-class (200-200)	↑ (good pattern)	↑	↑	↑	(1)
Minority (200-50,200-10)	↑ (good pattern)	↑	↓	↑	(2)
Majority (200-50,200-10)		↓	↑	↑	(3)

As we have explained, accuracy is not a good overall performance measure for class imbalance problems, which is strongly biased to the majority class. Although the single-class measures we have discussed so far reflect better the performance information for one class, it is still necessary to evaluate how well a classifier can balance the performance between classes. G-mean and AUC are better choices.

For a complete understanding of the effect of diversity, Table 7 presents the correlation coefficients of G-mean and AUC with respect to Q . Both measures show strong negative correlations with Q in almost all cases. It suggests that increasing diversity can lead to improved G-mean and AUC. The performance between classes is better balanced.

5.1.3 Decision Boundary Analysis

To see the radical effect of diversity visually, we produce classification boundary plots for data sets “200-200” and “200-10” at three specific sampling rates in Fig. 1: $r = 1000$ with a low diversity degree, $r = 100$

TABLE 7: Rank correlation coefficients (in %) between diversity measure Q and overall performance measures including G-mean and AUC for 3 artificial data sets. Numbers in boldface indicate significant correlations.

$r \in [3, 100]$	G-mean	AUC
200-200	-26	-39
200-50	-95	-98
200-10	-44	-51
$r \in [100, 1000]$	G-mean	AUC
200-200	-98	-99
200-50	-81	-99
200-10	-67	-96

with a moderate diversity degree, and $r = 20$ with a high diversity degree.

For the balanced case in Fig. 1(a), we can see that the main effect of diversity is to make the ensemble less overfit the training data close to boundaries. Hence, diversity improves the performance of both classes. For the imbalanced case in Fig. 1(b), in addition to a less overfitting boundary, diversity expands it towards the majority class side. A broader boundary is obtained for the minority class. It explains why more minority class examples are identified with majority-class recall reduced to some extent.

5.1.4 The Impact of Ensemble Size

Considering that the ensemble size is important to the application of an ensemble, we look into how diversity and the other performance measures change at different levels of ensemble size on the three artificial data sets “200-200”, “200-50” and “200-10”. We will show how the measures are affected by the ensemble size and the differences among the training data with different imbalance degrees.

Instead of keeping the constant size of 15 classifiers for an ensemble model, we adjust the number of decision trees from 5 to 955 with interval 50. The sampling rate for training is set to a moderate value of 100%. At each setting, we output overall diversity Q , the three single-class measures $R/P/F$ for each class, overall accuracy P_{ovr} , G-mean and AUC.

Table 8 presents the measure outputs from the ensemble with the smallest size (5 trees), the ensemble with the “optimal size” that gives the highest AUC

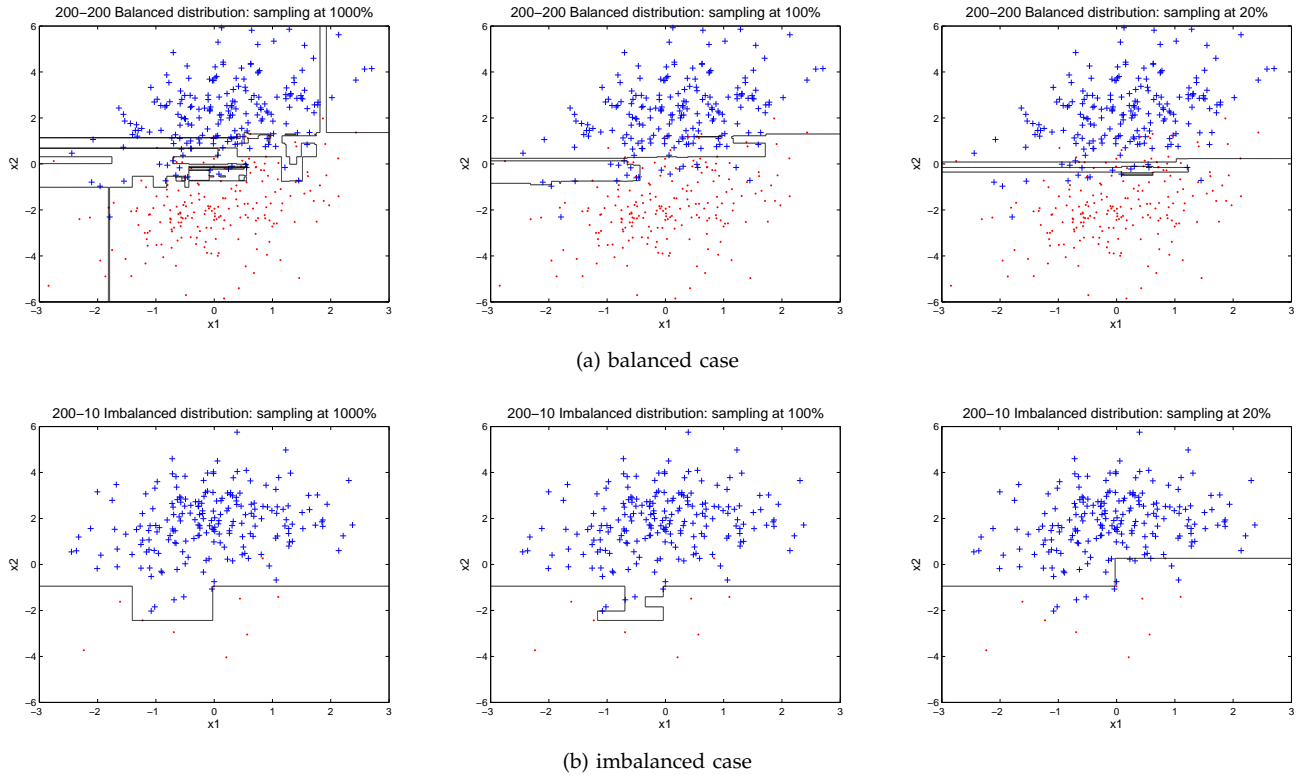


Fig. 1: Classification boundary plots produced by Bagging-based tree ensembles at sampling rates 1000%, 100%, and 20% of data sets “200-200” (balanced) and “200-10” (very imbalanced).

TABLE 8: Means and standard deviations of overall diversity Q and other performance measures produced by the ensembles with the smallest size (5 trees), the optimal size (where the highest AUC is achieved) and the largest size (955 trees) for 3 artificial data sets.

	200-200			200-50			200-10		
	Size=5	Size _{opt} =655	Size=955	Size=5	Size _{opt} =455	Size=955	Size=5	Size _{opt} =355	Size=955
Q	0.908±0.037	0.908±0.003	0.909±0.002	0.947±0.010	0.948±0.000	0.948±0.000	0.954±0.016	0.954±0.002	0.955±0.001
R_{min}	0.905±0.021	0.920±0.000	0.920±0.000	0.784±0.024	0.799±0.006	0.799±0.002	0.649±0.053	0.631±0.018	0.626±0.009
P_{min}	0.964±0.022	0.979±0.000	0.979±0.000	0.997±0.011	1.000±0.000	1.000±0.000	1.000±0.000	1.000±0.000	1.000±0.000
F_{min}	0.933±0.013	0.948±0.000	0.948±0.000	0.878±0.016	0.888±0.003	0.888±0.001	0.786±0.038	0.774±0.013	0.770±0.006
R_{maj}	0.966±0.022	0.980±0.000	0.980±0.000	0.998±0.009	1.000±0.000	1.000±0.000	1.000±0.000	1.000±0.000	1.000±0.000
P_{maj}	0.911±0.018	0.924±0.000	0.924±0.000	0.822±0.016	0.833±0.004	0.832±0.001	0.740±0.030	0.730±0.010	0.728±0.004
F_{maj}	0.937±0.013	0.951±0.000	0.951±0.000	0.901±0.011	0.908±0.002	0.908±0.001	0.851±0.019	0.844±0.006	0.842±0.003
P_{ovr}	0.935±0.013	0.950±0.000	0.950±0.000	0.891±0.013	0.900±0.003	0.899±0.001	0.824±0.026	0.815±0.009	0.813±0.004
G-mean	0.935±0.013	0.949±0.000	0.949±0.000	0.884±0.014	0.894±0.003	0.893±0.001	0.805±0.032	0.794±0.011	0.791±0.005
AUC	0.984±0.005	0.989±0.000	0.988±0.000	0.925±0.023	0.969±0.000	0.968±0.000	0.932±0.019	0.939±0.002	0.937±0.003

and the ensemble with the largest size (955 trees) for the three data sets. First, we can see that changing ensemble size is not an effective way of manipulating the diversity degree of ensembles. The Q -statistic value and most performance measures are changed very little under different ensemble sizes.

Second, comparing Q -statistic among the three data sets, we observe that Q appears to be higher in the imbalanced data sets than the one in the balanced data set. It shows that individual classifiers tend to be more similar to each other in imbalance data sets due to the difficulty in identifying rare examples, as we have explained in section 3. That is why ensemble diversity presents a more significant role in class imbalance problems.

Regarding the optimal ensemble size for each data set, a larger size ensemble seems to be more accurate when the training data is balanced or less imbalanced. The ensemble achieves the highest AUC when the ensemble size is 655 for “200-200” and 455 for “200-50”. In terms of the single-class measures and the other overall performance measures, the ensemble with the optimal size performs nearly the same as the largest size ensemble, and slightly better than the smallest size ensemble. When the training data gets more imbalanced, as shown in the case of “200-10”, a smaller size ensemble is preferable. The best AUC is obtained at the ensemble size of 355. The smallest size ensemble shows better single-class measures than the optimal size ensemble and the largest size ensemble.

The reason why a smaller size ensemble presents to be a better choice for a very imbalanced data set is that, the individual classifiers within the ensemble are so similar to each other that adding too many such classifiers only makes the combined classifier over-complex [54]. Under this scenario, a large size is not very useful. When the training data gets less imbalanced, the individual classifiers tend to be less similar to each other. So, increasing the size of the ensemble may further make use of this dissimilarity among the individuals and stabilize the prediction of the whole ensemble.

5.2 Impact of Diversity on Classification Performance on Real-World Imbalanced Data

So far, we have investigated the impact of diversity on single-class performance in depth through artificial data sets. Now we ask whether the results are applicable to real-world domains. In this section, we report the correlation results for the same research question on fifteen highly imbalanced real-world benchmarks. The data information is summarized in Table 9. Particularly, the first ten data sets come from the PROMISE data repository [55], which are collected from real software engineering projects. The task of this group of data is to predict defects existing in programming codes. There are two classes in each data set, defect and non-defect. They are highly imbalanced in nature, because defect examples are much less likely to occur than non-defect ones. The “insurance” data set comes from CoIL data mining competition [56]. The goal is to predict who would be interested in buying a specific insurance product. The last four data sets are chosen from the UCI repository [57] that are frequently discussed in the class imbalance learning literature. It is worth mentioning that the UCI data sets originally have more than two classes. We adapt them into two-class data problems by selecting a small class as the minority and merging the others as the majority.

TABLE 9: Summary of real-world data sets.

Data	Size	Attributes	Imbalance Rate
mc2	161	39	32.29%
mw1	403	37	7.69%
kc3	458	39	9.38%
cm1	498	21	9.83%
kc2	522	21	20.49%
pc1	1109	21	6.94%
pc4	1458	37	12.20%
pc3	1563	37	10.23%
kc1	2109	21	15.45%
pc2	5589	36	0.41%
insurance	5822	85	5.98%
glass	214	9	7.94%
ecoli	336	7	10.42%
balance	625	4	7.84%
car	1728	6	3.99%

The same Bagging-based training strategy is applied here as in the previous subsection (Table 4). The relationship is studied through correlation analysis using Spearman’s rank correlation coefficient. The sampling rate $r\%$ is varied in $[3\%, 100\%]$ with interval 2. At each sampling rate, 15 decision trees are built to form an ensemble. Except for “insurance” having a separate testing file, a 5-fold cross validation is run 10 times for the other data sets to compute the measures.

5.2.1 Single-Class Performance

Table 10 shows the correlations between overall diversity Q and the single-class measures plus P_{ovr} for each data set. Strong correlations are found.

Similar to the artificial cases, most coefficients between Q and Q_{min}/Q_{maj} present strong positive correlations. It confirms that ensemble diversity over each class gets larger if the overall diversity is increased by reducing the sampling rate $r\%$. Different from the artificial cases, diversity harms the overall accuracy according to its strong positive correlations with Q . Decreasing Q (i.e. larger diversity) makes P_{ovr} worse in all 15 data sets. Referring to the pattern analysis in section 3, it seems that although an ensemble performs in a good pattern in artificial scenarios, it’s more likely to behave in a bad pattern in real-world imbalanced domains.

More importantly, single-class performance should be our focus. For the minority class, recall has a very strong negative correlation with Q in all cases; precision has a very strong positive correlation with Q in 12 out of 15 cases; the coefficients of F-measure do not show a consistent relationship, where 6 cases present positive correlations and 5 cases present negative correlations. The observation suggests that more minority-class examples are identified with some loss of precision by increasing diversity. When the improvement of recall is larger than the drop of precision, F-measure gets better. Otherwise, it gets worse. Both directions could happen to F_{min} . Their behaviors correspond to the situation 4 in Table 3, except that F_{min} does not consistently follow the mathematical link in practice. For the majority class, recall and F-measure have very strong positive correlations with Q in all cases; precision has a very strong negative correlation with Q in 14 out of 15 cases. It indicates that although diversity helps to predict majority class examples more accurately, recall is sacrificed more than the improvement of precision. The result corresponds to the situation 5 in Table 3.

Generally speaking, diversity is helpful for recognizing minority class examples in real-world scenarios. A better balance between recall and precision of the minority class can be achieved in some cases. Diversity degrades the classification performance of the majority class in terms of recall and F-measure. The behaviors of recall of two classes tally with the arguments in section 3.3.

TABLE 10: Rank correlation coefficients (in %) between overall diversity Q and single-class performance measures plus P_{ovr} for real imbalanced data sets. Numbers in boldface indicate significant correlations.

data	Q_{min}	Q_{maj}	R_{min}	P_{min}	F_{min}	R_{maj}	P_{maj}	F_{maj}	P_{ovr}
mc2	96	99	-86	70	-11	89	-40	84	76
mw1	75	94	-96	86	31	98	-94	97	96
kc3	84	96	-98	96	31	99	-96	99	99
cm1	65	97	-98	59	-89	99	-97	98	98
kc2	1	96	-98	-4	-92	94	-97	59	10
pc1	97	99	-99	96	47	99	-98	98	98
pc4	97	99	-97	94	8	96	-97	93	93
pc3	97	99	-99	97	-22	99	-98	99	99
kc1	92	92	-99	97	-7	99	-99	99	98
pc2	85	93	-62	-12	-34	83	-62	82	82
insurance	100	100	-100	97	-72	100	-99	100	100
glass	95	62	-88	83	43	95	-84	93	93
ecoli	75	13	-77	77	48	80	-77	80	79
balance	97	100	-77	-77	-77	90	31	90	90
car	3	97	-84	81	66	83	-84	75	74

Table 11 summarizes the measure behaviors along with the decrease of Q (i.e. the increase of ensemble diversity) and the corresponding situation of each class in Table 3.

TABLE 11: Behaviors of single-class measures and overall accuracy by decreasing Q-statistic (more diverse) on 15 real data sets and their corresponding situations in Table 3. ‘-’ indicates an unclear changing behavior.

Class	P_{ovr}	R	P	F	Situation
Minority	↓	↑	↓	-	(4)
Majority	(bad pattern)	↓	↑	↓	(5)

5.2.2 Overall Performance

Table 12 reports the correlations of G-mean and AUC with Q-statistic. Even though Table 10 shows reduced overall accuracy as diversity increases in all 15 data sets, both G-mean and AUC present strong negative correlations with Q in most cases. It means that increasing diversity leads to better G-mean and AUC. Therefore, we can say that diversity is beneficial to the overall performance, which better balances the performance between classes. As one of our future work, it would be useful to study the theoretical links between diversity and overall performance measures including G-mean and AUC.

5.2.3 Additional Remarks

Comparing Table 6 and Table 11, we notice some differences between artificial and real-world data sets. In practice, the positive effect of diversity seems to get weaker on overall accuracy and F-measure. More complex data distributions and feature correlations could be the reasons. As diversity helps to find more minority class examples, the majority class could suffer from great accuracy reduction.

TABLE 12: Rank correlation coefficients (in %) between diversity measure Q and overall performance measures including G-mean and AUC for real imbalanced data sets. Numbers in boldface indicate significant correlations.

data	G-mean	AUC
mc2	5	60
mw1	-94	-61
kc3	-97	-76
cm1	-96	-68
kc2	-95	-96
pc1	-97	-70
pc4	-92	-57
pc3	-98	-75
kc1	-97	-81
pc2	-61	-90
insurance	-100	-98
glass	-71	-5
ecoli	-62	-56
balance	-77	-32
car	-43	14

Besides, the findings in this section are quite meaningful. Even if several authors found very little empirical relationship between overall accuracy and diversity of classification ensembles [28] [58], we observe strong correlations of single-class performance, G-mean and AUC with Q-statistic. It may suggest that different types of classification problems should be considered for better understanding the role of ensemble diversity in the future.

6 CONCLUSIONS

This paper studies the relationships between ensemble diversity and performance measures for class imbalance learning, aiming at the following questions: what is the impact of diversity on single-class performance? Does diversity have a positive effect on the classification of minority/majority class? We chose Q-statistic as the diversity measure and considered

three single-class performance measures including recall, precision and F-measure. The relationship with overall performance was also discussed empirically by examining G-mean and AUC for a complete understanding.

To answer the first question, we gave some mathematical links between Q-statistic and the single-class measures. This part of work is based on Kuncheva et al.'s pattern analysis [24]. We extended it to the single-class context under specific classification patterns of ensemble and explained why we expect diversity to have different impacts on minority and majority classes in class imbalance scenarios. Six possible behaving situations of the single-class measures with respective to Q-statistic are obtained. For the second question, we verified the measure behaviors empirically on a set of artificial and real-world imbalanced data sets. We examined the impact of diversity on each class through correlation analysis. Strong correlations are found. We show the positive effect of diversity in recognizing minority class examples and balancing recall against precision of the minority class. It degrades the classification performance of the majority class in terms of recall and F-measure on real-world data sets. Diversity is beneficial to the overall performance in terms of G-mean and AUC.

Significant and consistent correlations found in this paper encourage us to take this step further. We would like to explore in the future if and to what degree the existing class imbalance learning methods can lead to improved diversity and contribute to the classification performance. We are interested in the development of novel ensemble learning algorithms for class imbalance learning that can make best use of our diversity analysis here, so that the importance of the minority class can be better considered. It is also important in the future to consider class imbalance problems with more than two classes.

ACKNOWLEDGMENTS

Part of this research was supported by an ORS Award to the first author and an EPSRC grant (No. EP/D052785/1) to the second author.

REFERENCES

- [1] R. M. Valdovinos and J. S. Sanchez, "Class-dependant resampling for medical applications," in *Proceedings of the Fourth International Conference on Machine Learning and Applications (ICMLA'05)*, 2005, pp. 351–356.
- [2] T. Fawcett and F. Provost, "Adaptive fraud detection," *Data Mining and Knowledge Discovery*, vol. 1, no. 3, pp. 291–316, 1997.
- [3] K. J. Ezawa, M. Singh, and S. W. Norton, "Learning goal oriented bayesian networks for telecommunications risk management," in *Proceedings of the Thirteenth International Conference on Machine Learning*, 1996, pp. 139–147.
- [4] C. Cardie and N. Howe, "Improving minority class prediction using casespecific feature weights," in *Proceedings of the Fourteenth International Conference on Machine Learning*, 1997, pp. 57–65.
- [5] G. M. Weiss, "Mining with rarity: a unifying framework," *SIGKDD Explor. Newsl.*, vol. 6, no. 1, pp. 7–19, 2004.
- [6] S. Visa and A. Ralescu, "Issues in mining imbalanced data sets - a review paper," in *Proceedings of the Sixteen Midwest Artificial Intelligence and Cognitive Science Conference*, 2005, pp. 67–73.
- [7] N. Japkowicz and S. Stephen, "The class imbalance problem: A systematic study," *Intelligent Data Analysis*, vol. 6, no. 5, pp. 429 – 449, 2002.
- [8] C. Li, "Classifying imbalanced data using a bagging ensemble variation," in *ACM-SE 45: Proceedings of the 45th Annual Southeast Regional Conference*, 2007, pp. 203–208.
- [9] X.-Y. Liu, J. Wu, and Z.-H. Zhou, "Exploratory undersampling for class imbalance learning," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 39, no. 2, pp. 539–550, 2009.
- [10] N. V. Chawla, A. Lazarevic, L. O. Hall, and K. W. Bowyer, "Smoteboost: Improving prediction of the minority class in boosting," in *Knowledge Discovery in Databases: PKDD 2003*, vol. 2838, 2003, pp. 107–119.
- [11] M. V. Joshi, V. Kumar, and R. C. Agarwal, "Evaluating boosting algorithms to classify rare classes: Comparison and improvements," in *IBM Research Report*, 2001, pp. 257–264.
- [12] N. V. Chawla and J. Sylvestre, "Exploiting diversity in ensembles: Improving the performance on unbalanced datasets," *Multiple Classifier Systems*, vol. 4472, pp. 397–406, 2007.
- [13] H. Guo and H. L. Viktor, "Learning from imbalanced data sets with boosting and data generation: the databoost-im approach," *SIGKDD Explor. Newsl.*, vol. 6, no. 1, pp. 30–39, 2004.
- [14] W. Fan, S. J. Stolfo, J. Zhang, and P. K. Chan, "Adacost: Misclassification cost-sensitive boosting," in *Proc. 16th International Conf. on Machine Learning*, 1999, pp. 97–105.
- [15] M. P. Perrone and L. N. Cooper, "When networks disagree: Ensemble methods for hybrid neural networks," *Artificial Neural Networks for Speech and Vision*, pp. 126–142, 1993.
- [16] A. Estabrooks, T. Jo, and N. Japkowicz, "A multiple resampling method for learning from imbalanced data sets," in *Computational Intelligence 20*, vol. 20, no. 1, 2004, pp. 18–36.
- [17] S. B. Kotsiantis and P. E. Pintelas, "Mixture of expert agents for handling imbalanced data sets," *Annals of Mathematics, Computing and Teleinformatics*, vol. 1, no. 1, pp. 46–55, 2003.
- [18] L. I. Kuncheva, M. Skurichina, and R. P. Duin, "An experimental study on diversity for bagging and boosting with linear classifiers," *Information Fusion*, vol. 3, no. 4, pp. 245–258, 2002.
- [19] R. Polikar, "Ensemble based systems in decision making," in *IEEE Circuits and Systems Magazine*, 2006, pp. 21–45.
- [20] G. Brown, J. L. Wyatt, and P. Tino, "Managing diversity in regression ensembles," *Journal of Machine Learning Research*, vol. 6, pp. 1621–1650, 2005.
- [21] S. B. Kotsiantis, D. Kanellopoulos, and P. Pintelas, "Handling imbalanced datasets: A review," *GESTS International Transactions on Computer Science and Engineering*, vol. 30, no. 1, pp. 25–36, 2006.
- [22] C. V. Rijsbergen, *Information Retrieval*. London, Butterworths, 1979.
- [23] G. U. Yule, "On the association of attributes in statistics," *Philosophical Transactions of the Royal Society of London*, vol. A194, pp. 257–319, 1900.
- [24] L. I. Kuncheva, C. J. Whitaker, C. A. Shipp, and R. P. Duin, "Limits on the majority vote accuracy in classifier fusion," *Pattern Analysis and Applications*, vol. 6, no. 1, pp. 22–31, 2003.
- [25] G. Brown, J. Wyatt, R. Harris, and X. Yao, "Diversity creation methods: a survey and categorisation," *Information Fusion*, vol. 6, no. 1, pp. 5–20, 2004.
- [26] K. M. Ali and M. J. Pazzani, "On the link between error correlation and error reduction in decision tree ensembles," University of California, Irvine, Department of Information and Computer Science, Technical Report ICS-TR-95-38, 1995.
- [27] A. H.-R. Ko and R. Sabourin, "Compound diversity functions for ensemble selection," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 23, no. 4, pp. 659–686, 2009.
- [28] L. I. Kuncheva and C. J. Whitaker, "Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy," *Machine Learning*, vol. 51, pp. 181–207, 2003.
- [29] —, "Ten measures of diversity in classifier ensembles: limits for two classifiers," in *A DERA/IEE Workshop on Intelligent Sensor Processing (Ref. No. 2001/050)*, 2001, pp. 1–10.
- [30] K. Tang, P. N. Suganthan, and X. Yao, "An analysis of diversity measures," *Machine Learning*, vol. 65, pp. 247–271, 2006.

- [31] G. Giacinto and F. Roli, "Design of effective neural network ensembles for image classification purposes," *Image and Vision Computing*, vol. 19, no. 9-10, pp. 699-707, 2000.
- [32] P. Cunningham and J. Carney, "Diversity versus quality in classification ensembles based on feature selection," *Machine Learning: ECML 2000*, vol. 1810, pp. 109-116, 2000.
- [33] D. Partridge and W. Krzanowski, "Software diversity: practical statistics for its measurement and exploitation," *Information and Software Technology*, vol. 39, no. 10, pp. 707-717, 1997.
- [34] K. Tumer and J. Ghosh, "Linear and order statistics combiners for pattern classification," *Combining Artificial Neural Networks*, pp. 127-162, 1999.
- [35] —, "Analysis of decision boundaries in linearly combined neural classifiers," *Pattern Recognition*, vol. 29, no. 2, pp. 341-348, 1996.
- [36] Q. Gu, Z. Cai, L. Zhu, and B. Huang, "Data mining on imbalanced data sets," in *International Conference on Advanced Computer Theory and Engineering 2008*, 2008, pp. 1020-1024.
- [37] X. Guo, Y. Yin, C. Dong, G. Yang, and G. Zhou, "On the class imbalance problem," in *ICNC '08: Proceedings of the 2008 Fourth International Conference on Natural Computation*, 2008, pp. 192-201.
- [38] F. Provost, "Machine learning from imbalanced data sets 101," in *Proceedings of the AAAI'00 Workshop on Imbalanced Data Sets*, 2000.
- [39] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1263-1284, 2009.
- [40] Y. Sun, M. S. Kamel, A. K. Wong, and Y. Wang, "Cost-sensitive boosting for classification of imbalanced data," *Pattern Recognition*, vol. 40, no. 12, pp. 3358-3378, 2007.
- [41] P. K. Chan and S. J. Stolfo, "Toward scalable learning with non-uniform class and cost distributions: A case study in credit card fraud detection," in *Knowledge Discovery and Data Mining*, 1998, pp. 164-168.
- [42] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, no. 2, pp. 123-140, 1996.
- [43] Y. Freund and R. E. Schapire, "Experiments with a new boosting algorithm," in *Proc. of the 13th. Int. Conf. on Machine Learning*, 1996, pp. 148-156.
- [44] K. M. Ting, "A comparative study of cost-sensitive boosting algorithms," in *Proc. 17th International Conf. on Machine Learning*, 2000, pp. 983-990.
- [45] C. Drummond and R. C. Holte, "C4.5, class imbalance, and cost sensitivity: Why under-sampling beats over-sampling," in *Proceedings of the International Conference on Machine Learning (ICML 2003) Workshop on Learning from Imbalanced Data Sets*, 2003, pp. 1-8.
- [46] M. V. Joshi, "On evaluating performance of classifiers for rare classes," *IEEE International Conference on Data Mining*, vol. 0, pp. 641-661, 2002.
- [47] Y. Sun, M. S. Kamel, and Y. Wang, "Boosting for learning multiple classes with imbalanced class distribution," in *ICDM '06: Proceedings of the Sixth International Conference on Data Mining*, 2006, pp. 592-602.
- [48] M. Kubat, R. Holte, and S. Matwin, "Learning when negative examples abound," in *The 9th European Conference on Machine Learning*, vol. 1224, 1997, pp. 146-153.
- [49] M. Kubat and S. Matwin, "Addressing the curse of imbalanced training sets: One-sided selection," in *Proc. 14th International Conference on Machine Learning*, 1997, pp. 179-186.
- [50] A. P. Bradley, "The use of the area under the roc curve in the evaluation of machine learning algorithms," *Pattern Recognition*, vol. 30, no. 7, pp. 1145-1159, 1997.
- [51] C. X. Ling, J. Huang, and H. Zhang, "Auc: a statistically consistent and more discriminating measure than accuracy," in *Proceedings of 18th International Conference on Artificial Intelligence (IJCAI2003)*, 2003, pp. 329-341.
- [52] J. A. Swets, "Measuring the accuracy of diagnostic systems," *Science*, vol. 240, no. 4857, pp. 1285 - 1293, 1988.
- [53] M. Skurichina, L. I. Kuncheva, and R. P. Duin, "Bagging and boosting for the nearest mean classifier - effects of sample size on diversity and accuracy," *Multiple Classifier Systems*, pp. 307-311, 2002.
- [54] R. E. Schapire, Y. Freund, P. Bartlett, and W. S. Lee, "Boosting the margin: A new explanation for the effectiveness of voting methods," *The Annals of Statistics*, vol. 26, no. 5, pp. 1651-1686, 1998.
- [55] G. Boetticher, T. Menzies, and T. J. Ostrand. (2007) Promise repository of empirical software engineering data. [Online]. Available: <http://promisedata.org/repository>
- [56] P. V. D. Putten and M. V. Someren, "A bias-variance analysis of a real world learning problem: The coil challenge 2000," *Machine Learning*, vol. 57, no. 1-2, pp. 177-195, 2004.
- [57] A. Frank and A. Asuncion. (2010) UCI machine learning repository: <http://archive.ics.uci.edu/ml>.
- [58] N. Garcia-Pedrajas, C. Hervas-Martinez, and D. Ortiz-Boyer, "Cooperative coevolution of artificial neural network ensembles for pattern classification," *IEEE Transactions on Evolutionary Computation*, vol. 9, no. 3, pp. 271- 302, 2005.



Shuo Wang received the B.Sc. degree in Computer Science from the Beijing University of Technology (BJUT), China, in 2006, and was a member of Embedded Software and System Institute in BJUT in 2007. She received the Ph.D. degree in Computer Science from the University of Birmingham, U.K., in 2011, sponsored by the Overseas Research Students Award (ORSAS) from the British Government (2007). She is currently a post-doctoral Research Fellow at the Centre

of Excellence for Research in Computational Intelligence and Applications (CERCIA) in the School of Computer Science, the University of Birmingham. Her research interests include class imbalance learning, ensemble learning, machine learning and data mining.



Xin Yao (M'91-SM'96-F'03) received the B.Sc. degree from the University of Science and Technology of China (USTC), Hefei, China, in 1982, the M.Sc. degree from the North China Institute of Computing Technology, Beijing, China, in 1985, and the Ph.D. degree from USTC in 1990. He was an Associate Lecturer and Lecturer with USTC, Hefei, China, in 1985-90, a Postdoctoral Fellow at the Australian National University (Canberra) and CSIRO (Melbourne) in Australia, in 1990-92, and a Lecturer, Senior Lecturer and Associate

Professor at UNSW@ADFA, Canberra, Australia, in 1992-99. Since April 1999, he has been a Professor (Chair) of Computer Science at the University of Birmingham, U.K., where he is currently the Director of the Centre of Excellence for Research in Computational Intelligence and Applications (CERCIA). He is also a Distinguished Visiting Professor (Grand Master Professorship) of USTC. His major research interests include evolutionary computation and neural network ensembles. He has more than 350 refereed publications. He was a recipient of the 2001 IEEE Donald G. Fink Prize Paper Award, IEEE Transactions on Evolutionary Computation Outstanding 2008 Paper Award, IEEE Transactions on Neural Networks Outstanding 2009 Paper Award, and several other best paper awards. He is an IEEE Fellow, a Distinguished Lecturer of IEEE Computational Intelligence Society, an invited keynote/plenary speaker of 60 international conferences, and a former (2003-08) Editor-in-Chief of the IEEE Transactions on Evolutionary Computation.