Abstract

This report presents the work done since the progress report 4, the current research, and a brief research plan for the 3rd year PhD study. During the past several months, we studied the relationship between diversity of classification ensembles and single-class measures theoretically and empirically for class imbalance learning. The goal is to find out if and when ensemble diversity can improve the classification performance on the minority class. We analyzed the issue on imbalanced data domains with two classes. Currently, in order to introduce diversity into classifier’s training procedure explicitly, a novel ensemble algorithm based on negative correlation learning (NCL) is in progress. It aims to overcome the limitations of the classical NCL algorithm that have been identified. For the 3rd year research, we plan to extend the ensemble learning over imbalanced data sets to domains with multiple classes (more than two classes). This report also introduces some additional literature studies. Finally, the timetable for remainder of the research is proposed.
## Contents

1 Introduction .................................................. 3

2 Work Done Thus Far – The Relationship Between Diversity and Single-Class Measures (Two Classes) ........... 4

2.1 Why Can Diversity Improve the Classification on Imbalanced Data Sets? ........................................ 4

2.2 Additional Literature Review on Diversity in Classification Ensembles ........................................ 5

2.2.1 Tumer and Ghosh’s “added” error [16] ........................................................................................................ 5

2.2.2 Kuncheva et al.’s pattern analysis [6] ........................................................................................................ 6

2.3 The Relationship Between Diversity and Single-Class Measures ....................................................... 7

2.3.1 Recall, precision, F-measure vs. Q with independence assumption .................................................... 7

2.3.2 Recall, precision, F-measure vs. Q without independence assumption .............................................. 8

2.4 Evaluate Good/Bad Diversity .................................................................................................................. 9

3 Current Research ................................................... 10

3.1 Additional Study of Cost-Sensitive Boosting ......................................................................................... 11

3.2 Proposed Algorithm ............................................................................................................................... 11

4 Future Studies – Class Imbalance Learning with Multiple Classes ...................................................... 12

5 Updated Timetable .................................................. 13

6 Appendix: Published Paper in Workshop of ICDM2009 ................................................................. 14

6.1 Theoretical Study of the Relationship Between Diversity and Single-Class Measures for Class Imbalance Learning .............................................................................................................. 17


1 Introduction

Before the last report, we empirically explored the impact of diversity of classification ensembles for class imbalance learning. We found that diversity improves recall over the minority class significantly, and single-class performance measures (recall, precision, F-measure) present different behaviors between the minority class and the majority class [20]. Then, we considered NCL, a diversity-encouraging ensemble algorithm, can be a convenient way to introduce diversity into ensemble system for identifying more examples from the minority class, since classifiers tend to overfit a small amount of minority examples and ignore the others. The results show that it is quite competitive with some other state-of-art solutions [19]. We are intrigued to know the theoretical reasons that how diversity impacts single-class performance measures over either the minority class or the majority class, in order to find out if and when ensemble’s diversity can improve the classification performance on the minority class. Hence, our research is continued around the following questions: what is the impact of diversity on single-class measures? Is it different from the impact on multi-class (overall) measures? If we increase the diversity degree of ensemble on minority class, does it give better performance? To address those questions, we build the link between Q-statistic and widely used single-class measures (including recall, precision, and F-measure), by utilizing the relationship between Q-statistic and majority vote accuracy [6].

Furthermore, imbalanced data sets are not limited to have two classes. A large number of real-world problems fall into multiple categories, where multiple minority or majority classes exist. It is interesting and worthwhile to extend the ensemble analysis on two-class data sets to multi-class data sets. Little work has been done on this topic yet. This will be the main work for my future research.

Since diversity has positive impact on the classification of the minority class, it could be a good idea to introduce diversity explicitly into classifier training procedure. We explored NCL for class imbalance learning. However, we notice that NCL suffers some obvious limitations, including low flexibility and long training time. No good solution has been proposed yet. We get some ideas from AdaBoost [3] and cost-sensitive learning methods [2] [13] [15], and propose a novel ensemble algorithm AdaBoost.NC to overcome the shortcomings of NCL. This part of work is in progress.

In this report, section 2 presents the work done since May 2009. More details can be found in “Appendix” section. Section 3 introduces some additional literature study and explains our new method to handle the problems of NCL. Section 4 gives some directions for my future study. Finally, section 5 presents the updated timetable for remainder of the research.
2 Work Done Thus Far – The Relationship Between Diversity and Single-Class Measures (Two Classes)

This section describes the work that has been done since progress report 4. Our previous research did diversity analysis experimentally on some imbalanced data sets. We find that high diversity degree always leads to better recall on the minority class, but there is no clear pattern for the other performance measures. Single-class measures on the minority and majority class behave towards two different directions. Proper diversity degree tends to improve the classification performance on the minority class, but degrade the performance on the majority class. In order to get a deeper understanding of diversity impact on the minority class, we are interested in the theoretical reasons that how those single-class measures are affected by the diversity in an ensemble system. If diversity is proved to be capable of advancing the classification of imbalanced data from both theoretical and empirical aspects, it will be a good idea to introduce diversity explicitly into learning process. Hence, we have been working on the relationship between diversity and single-class performance measures theoretically and empirically. Before that, we firstly explain why we believe diversity can improve the classification on imbalanced data sets briefly. Then, we give a simple review of some important literature. Finally, we conclude the main work we have done.

2.1 Why Can Diversity Improve the Classification on Imbalanced Data Sets?

Learning from imbalanced data sets usually produces biased classifiers that have a higher accuracy over the majority class, but poorer accuracy over the minority class. The classification boundary is too “tight” towards the minority class and causes overfitting. Intuitively, if we increase diversity of the ensemble, classification “uncertainty” will be introduced. The examples from the minority class thus get higher probability to be identified.

In the light of bias-variance-covariance decomposition [17] and ambiguity decomposition [5], diversity issue is reformulated as a three-way trade-off. The fundamental problem is how to balance diversity against individual accuracy. Sometimes, diversity is harmful to generalization performance, because it is difficult to maximize diversity without affecting the other parts of error. For a classifier learnt from an imbalanced data set, however, it is always biased towards the majority class and overfits the minority class. Even if the error bias is very low for the training data, the error bias for the unseen data is very high. Therefore, increasing diversity can cause obvious accuracy degradation over the training data, but may not happen over the unseen data, and it may even help rectify the biased classification boundary.

Tang et al. [14] analyzed the ineffectiveness of the diversity measures for classification
ensembles by relating them to the concept of margin \[1\] in 2006. They explained that achieving a larger margin on the training set results in an improved bound on the generalization error of the ensemble, and the ensemble with the largest minimum margin will have the best generalization error bound. They also proved that seeking diversity in an ensemble can be viewed as an implicit way to maximize the minimum margin of the ensemble. According to their analysis, margin of the ensemble can be calculated by,

\[
m_i = \sum_{j=1}^{L} w_j O_{ij}
\]

(1)

\(m_i\) is the margin on example \(x_i\). \(O_{ij} = 1\) if training example \(x_i\) is classified correctly by base classifier \(h_j\). Otherwise, \(O_{ij} = -1\). For an imbalanced data set, the minimum margin is usually caused by the examples belonging to the minority class, since more negative \(O_{ij}\) exist when classifying difficult examples. If we introduce diversity to the minority class, the minimum margin will get improved. Therefore, better generalization performance is achieved.

Based on the discussion, we conjecture that increasing diversity explicitly can be a good way to deal with imbalanced data classification. We thus have the following diversity analysis on the single-class measures. We aim to understand the impact of diversity on a single class, and prove if and when diversity is beneficial to the classification over the minority class.

2.2 Additional Literature Review on Diversity in Classification Ensembles

Diversity in regression ensembles has already been proved and quantified theoretically \[1\]. When problem domain produces crisp class labels, the decomposition of generalization error is not as easy as the ones in regression ensembles. So far, the theoretical studies of the relationship between diversity and classification performance has been discussed mainly in two ways.

2.2.1 Tumer and Ghosh’s “added” error \[16\]

They consider the classifier whose outputs are expected to approximate the posteriori class probabilities if it is reasonably well trained. When multiple such classifiers are linearly combined, they derive an explicit relationship between the correlation among classifier errors and the error reduction due to combining. Under the assumption that \(L\) combined classifiers are unbiased, the added error region of the combining result is given by,
\[
E_{add}^{ave} = E_{add} \left( \frac{1 + \delta (L - 1)}{L} \right)
\]  \tag{2}

\( \delta \) is the overall correlation term among classifiers, which is a part of combining error variance. If the errors are independent, then the second part of the reduction term vanishes and the combined error is reduced by \( L \). If the error of each classifier has correlation 1, then the combiner does not bring improvement. If the errors are negatively correlated, consistent with the conclusions in regression ensembles, classifier combining causes the largest error reduction comparing to the first two cases.

2.2.2 Kuncheva et al.’s pattern analysis \[6\]

Kuncheva’s study builds a connection between Q-statistic (denoted as \( Q \)) and overall accuracy produced by majority vote (denoted as \( P_{maj} \)). It has two main contributions. Firstly, they gave several synthetic examples with fixed settings, and illustrated that there is no clear pattern of relationship between \( P_{maj} \) and the \( Q \)'s. However, the \( Q \)'s with high \( P_{maj} \) tend to achieve negative values. Motivated by the first part of study, they 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”. The two patterns are assumed to be the “best case” and the “worst case”. Under the assumptions of each pattern, the functional link between Q-statistic and \( P_{maj} \) was found.

Pattern of success (best pattern) Under the pattern of success, no correct votes are wasted. The expression of \( P_{maj} \) with respect to \( Q \) is,

\[
P_{maj} = \frac{L}{l + 1} \frac{(1 - Q)}{(2 - Q)}
\]  \tag{3}

where \( l = \lfloor L/2 \rfloor \) for the convenience. \( P_{maj} \) is a monotone decreasing function of the pairwise dependence \( Q \). For any \( p > 2/3 \) (individual accuracy), the value of \( Q \) will be -1.

Pattern of failure (worst pattern) Under the pattern of failure, correct votes are wasted to the greatest extent. The expression of \( P_{maj} \) with respect to \( Q \) is,

\[
P_{maj} = \frac{L}{(l - 1)} \frac{1}{(2 - Q)} - \frac{l}{(L - l)}
\]  \tag{4}

\( P_{maj} \) is a monotone increasing function of \( Q \). For any \( p > 0.5 \), the value of \( Q \) is positive.
2.3 The Relationship Between Diversity and Single-Class Measures

Considering that \( P_{maj} \), recall, precision and F-measure are correlated with each other, we discuss the relationship between diversity and single-class measures by extending Kuncheva’s pattern analysis. We firstly build the functional link for Q-statistic and single-class measures in the “good” and “bad” patterns. “Good” pattern has good diversity status that is beneficial to classification accuracy on a set of data, such as “pattern of success”. Otherwise, it is “bad” pattern, such as “pattern of failure”. Then, we analyze the impact of diversity on the minority and majority class by considering the classification characteristics caused by the imbalanced data distribution.

2.3.1 Recall, precision, F-measure vs. Q with independence assumption

To simplify the relationship problem, we firstly derive the expressions for recall (Rec), precision (Acc), and F-measure (F) with respect to diversity measure Q under some assumption. Here, it is assumed that the class label of an example voted by the majority classifiers is independent with its real class label. This assumption is quite strong and especially not appropriate for imbalanced data sets, because data from the minority class are harder to be recognized than data from the majority class. As data sets get less imbalanced, however, the dependence on the class label will get ambiguous. The assumption is likely to be the case for balanced data sets, and can be regarded as an extreme situation. Under the assumption, we make the following conclusions.

**Best pattern** Both Rec and Acc are monotone decreasing functions with respect to Q.

\[
Rec = \frac{L}{l+1} \frac{1-(Q)}{(2-Q)}
\]

\[
Acc = \frac{L \theta (1-Q)}{l(1-\theta)(2-Q) + L(2\theta-1)(1-Q)}
\]

where \( \theta \) is the proportion of the class of data. This result supports both recall and precision values will get smaller when \( Q \) increases, and therefore F-measure will get smaller. In other words, diversity results in performance improvement on single class under this pattern.

**Worst pattern** Both Rec and Acc are increasing functions with respect to Q.

\[
Rec = \frac{L}{(l-1)(2-Q)} - \frac{l}{(L-l)}
\]

\[
Acc = \frac{L \theta (1-Q)}{L(1-2\theta)(1-Q) + L(L-l) \theta (2-Q)}
\]
When Q gets larger, both recall and precision values will get larger, and therefore F-measure will improve. Diversity leads to performance deterioration on single class under this pattern.

2.3.2 Recall, precision, F-measure vs. Q without independence assumption

The independence assumption removes the trade-off between recall and precision. Without the assumption, it is not easy to get such neat and separate expressions for those measures. Their relations make them behave in correlated ways. Majority vote accuracy ($P_{maj}$) is therefore utilized to connect single-class measures and Q-statistic. $P_{maj}$ is reformulated in several ways, each of which is expressed by measures from the same class.

\[
P_{maj} = (1 - \theta) + \theta \left( 2 - \frac{1}{\text{Acc}} \right) \text{Rec} \tag{9}
\]

\[
P_{maj} = (1 - \theta) + \theta \left( \frac{2\text{Acc} - 1}{2\frac{F}{Rec} - 1} \right) \tag{10}
\]

\[
P_{maj} = (1 - \theta) + \theta \left( 2\text{Rec} - \eta \frac{\theta}{\eta} \right) \tag{11}
\]

\[
P_{maj} = 1 - (1 - F) (\theta + \eta) \tag{12}
\]

where $\eta$ is the probability of predicting an example as the corresponding class, within which the single-class measures are calculated. To further discuss the role of diversity on the minority and majority class, we make use of the above links among Q-statistic, $P_{maj}$, and single-class measures, and take the classification characteristics brought by imbalanced data distribution into consideration. Intuitively, two situations can be imagined for the minority and majority class respectively.

**Hypothesis 1:** Learners always have poor ability of recognizing data from minority class. Traditional algorithms tend to ignore them or overfit the training data. For an ensemble system, most of the examples from the minority class are misclassified, and very few of them are identified by over half of the individuals reluctantly. There should not be many wasted correct votes. Therefore, this class of data is close to the “best pattern”, at least in good pattern. The ensemble has low diversity degree. According to our analysis, increasing diversity on minority class should have positive influence, particularly on recall.

**Hypothesis 2:** Data from the majority class are capable of providing sufficient information to learners. Consequently, every ensemble member tends to make the same correct decision. There are a lot of wasted votes that give the correct answers. In this situation, the majority class is likely to belong to the “worst pattern”. Increasing diversity may produce negative effect.
The intention of the hypotheses is to map minority/majority class to a proper pattern according to their performance characteristics caused by imbalanced data distribution. From the hypotheses and experimental verifications, we obtain the impact of diversity on the minority and majority class, which is different from the case of balanced data sets. As Q-statistic increases,

**Better** $P_{maj}$: For the minority class, if $P_{maj}$ is enhanced, the most obvious result is the reduction of recall. Precision tends to increase. How F-measure changes is still unclear. In this case, diversity enhancement means increasing the probability of classifying an example into the minority class, but sacrificing the classification precision. Therefore, it may or may not be helpful. Given the hypotheses and similar analysis, measures on majority class will behave in the opposite way.

**Worse** $P_{maj}$: For the minority class, if $P_{maj}$ is reduced, both recall and F-measure suffers degradation, but the change of precision is unsure. In this case, diversity enhancement results in the improvement of both recall and F-measure. Therefore, introducing diversity could be a good choice to deal with imbalanced data.

More details can be found in Appendix section. All the above hypotheses and conclusions are verified experimentally on imbalanced data sets and balanced data sets.

### 2.4 Evaluate Good/Bad Diversity

In Kuncheva’s pattern analysis on overall accuracy and our diversity analysis on single-class measures, “good/bad” diversity is kept mentioned. The best pattern has good diversity, since increasing diversity is beneficial to classification accuracy over a set of data. Similarly, the worst pattern has bad diversity, because diversity is harmful to classification accuracy. However, the literature does not give any deep study of it. The key problem is how to judge the current diversity degree if it is good or bad. Diversity measure itself cannot tell us if diversity will bring performance improvement or deterioration into the current ensemble. In the best pattern, Q’s possible range is $[-1, 1]$. In the worst pattern, Q varies in $[2 - L/l, 1]$. Therefore, we need a measure to evaluate the current diversity “quality”.

During the study of the relationship between diversity and overall/single-class measures in patterns, we find two factors relevant to the current diversity status, the number of wasted correct votes from the classifier individuals and the combination probabilities of majority voting. We therefore define a new measure – wasted vote (denoted as $N_{waste}$) based on the factors. It is expected to be an evaluation criterion for current diversity degree and predict performance behavior caused by diversity.

$$N_{waste} = (a - l - 1) \binom{L}{a} \alpha + b \binom{L}{b} \beta$$  \hspace{1cm} (13)

The first term represents $(a - l - 1)$ correct votes are wasted with probability $\alpha$ when
an input is classified correctly. The second term means \( b \) correct votes are wasted with probability \( \beta \) when an input is misclassified. In the best pattern, \( N_{\text{waste}} = 0 \). In the worst pattern, \( N_{\text{waste}} = l \). \( l/2 \) is the critical value to discriminate between good and bad diversity.

It is validated on four imbalanced data sets. It shows that the ideal critical value of \( N_{\text{waste}} \) is hard to achieve in real data. The practical threshold is usually higher than \( l/2 \). We observe the \( N_{\text{waste}} \) values from each class. The \( N_{\text{waste}} \) of the minority class is always lower than the \( N_{\text{waste}} \) of the majority class. The \( N_{\text{waste}} \) of the minority class will get smaller as diversity degree becomes lower. It further proves that if the ensemble “overfits” the minority data more, less correct votes will be wasted. Diversity tends to be good to the minority class, and bad to the majority class. Generally, if \( Q \) is high and \( N_{\text{waste}} \) is low, diversity of current ensemble is in good pattern; if \( Q \) is high and \( N_{\text{waste}} \) is high, diversity is in bad pattern.

This part of work is discussed in our journal paper, which will be finished within the next two months. Besides, we will further give an AUC/G-mean discussion to see if diversity is also beneficial to overall performance. AUC and G-mean are usually used in class imbalance learning, because they are not sensitive to imbalanced data distribution, and they describe the performance trade-off between the minority class and the majority class.

### 3 Current Research

During the research of ensemble learning and class imbalance learning, we find that increasing diversity can generally improve the classification performance on the minority class in most of the time. Therefore, we consider introducing diversity explicitly into ensemble training procedure. All existing ensemble solutions achieve diversity implicitly, such as under-sampling the data from the majority class, generating new data for the minority class, etc. NCL algorithm [9] and the related techniques have been studied. NCL is a successful ensemble learning technique of neural networks (NN). Unlike Bagging and Boosting, NCL encourages diversity explicitly by adding correlation penalty term to NN’s error function. Our previous results show that NCL is quite competitive with some state-of-art ensemble methods to deal with imbalanced data sets [10]. However, it suffers some known drawbacks, which hinder it from being more widely used. It is only applicable to NN-based ensemble that uses back propagation training strategy. The pattern-by-pattern weight updating method makes the training procedure much slower than training multiple NNs independently (e.g. Bagging) or sequentially (e.g. Boosting).

To overcome those problems, we propose a new ensemble algorithm designed for classification data domains. Currently, we are implementing the algorithm and doing comparisons with the classical NCL algorithm and other state-of-art ensemble methods in class imbalance learning area. Our idea is inspired by cost-sensitive Boosting.
3.1 Additional Study of Cost-Sensitive Boosting

AdaBoost algorithm is a breakthrough in ensemble learning, proposed by Freund and Schapire in 1996 [3]. AdaBoost builds base learners sequentially by emphasizing the hardest examples. A set of weights is maintained over the training data. Misclassified examples by previous learner get their weights increased for the next iteration. Therefore, harder examples possess larger weights that have higher possibility to be selected for the next training.

A widely used variation of AdaBoost is cost-sensitive Boosting, mainly aiming for improving classification of imbalanced data and data suffering different misclassification costs. Some state-of-art algorithms include AdaCost [2], CSB1/CSB2 [15], and RareBoost [4]. RareBoost [4] is claimed to be a “cost-insensitive” method by the authors. Actually, it uses recall and precision information of the rare class as “cost”. AdaBoost is easy to be adapted for advancing the classification on minority class, because the weighting strategy makes it possible to concentrate on some specific training examples. Besides, it is algorithm-independent. Any base learning method suitable to the classification problem can be applied. Particularly, Sun [13] proposed a generic framework of cost-sensitive Boosting algorithms.

3.2 Proposed Algorithm

Considering the advantages of AdaBoost, its weight updating strategy allows us to balance accuracy and diversity at the same time. We can make AdaBoost “cost-sensitive” by using diversity as cost term. Hence, we propose AdaBoost.NC algorithm (the negative correlation version of AdaBoost). Table 1 presents the basic idea of AdaBoost.NC algorithm.

AdaBoost.NC is independent of the selection of base learning algorithms, and the weight-updating happens on the ensemble level, instead of the pattern level in the classical NCL method. It makes negative correlation learning simpler and faster. AdaBoost.NC could be a good way to overcome the disadvantages of NCL.

Some algorithm-related issues existing in AdaBoost.NC need further discussion and experimental validation. Some key points that we are working on are listed in the following,

- The choice of penalty term $p_t$ as the cost of diversity in step 3;
- The form of weight updating rule in step 5; (How should we introduce penalty term into the weight updating rule? Different forms result in different ways of weight adjustment.)
- The formula for calculating the weight updating parameter $\alpha_t$ to bound the training error of the final classifier in step 4;
Table 1: AdaBoost.NC Algorithm

<table>
<thead>
<tr>
<th>Given training data set $Z = {(x_1, y_1), \ldots, (x_i, y_i), \ldots, (x_N, y_N)}$, $y_i \in {-1, +1}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initialize data weights $D_{1(i)} = 1/N$, penalty term $p_{1(i)} = 1$.</td>
</tr>
<tr>
<td>For training epoch $t = 1, 2, \ldots, T$:</td>
</tr>
<tr>
<td>(1) Train weak learning $h_t$ using distribution $D_t$.</td>
</tr>
<tr>
<td>(2) Get weak hypothesis $h_t$.</td>
</tr>
<tr>
<td>(3) Calculate penalty term for every example $x_i$: $p_t(i)$</td>
</tr>
<tr>
<td>(4) Calculate $h_t$’s weight $\alpha_t$ by error and penalty.</td>
</tr>
<tr>
<td>(5) Update data weights $D_t$ and obtain new weights $D_{t+1}$:</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>$D_{t+1(i)} = \left(\frac{p_{t(i)}}{\lambda} \right)^\lambda \frac{D_t(i) \exp(-\alpha_t y_i h_t(x_i))}{Z_t}$ or $D_{t+1(i)} = \frac{D_t(i) \exp\left(-\alpha_t y_j h_t(x_i)\left(p_{t(i)}\right)^\lambda\right)}{Z_t}$</td>
</tr>
</tbody>
</table>

Output the final hypothesis:

$H(x) = \text{sign} \left(\sum_{t=1}^{T} \alpha_t h_t(x)\right)$

4 Future Studies – Class Imbalance Learning with Multiple Classes

What we have done in section 2 analyzed imbalanced data sets with binary classes (one minority class and one majority class). If there are more than two classes, more possible situations need to be considered. Intuitively, at least three cases exist pervasively in real-world applications: 1) multiple minority classes and one majority class (e.g. two rare diseases vs. normal cases); 2) one minority class and multiple majority classes (e.g. spam vs. multiple kinds of non-spam in e-mail filtering); 3) multiple minority classes and multiple majority classes. So far, however, little work has been done in class imbalance learning area with multi-class. Most current solutions are cost-sensitive approaches [8] [23] [22] [12], or alternatively multi-class is decomposed into several binary problems [10] [21].

Because imbalanced data set generally suffers unequal error costs (data from minority classes are harder to obtain or more costly), cost-sensitive learning has been used as good solutions [7] [18]. However, cost-sensitive methods are not limited to address class imbalance problem. They can be applied to any problem aiming at minimizing the misclassification cost. This generality of cost-sensitive learning makes us worried about some features that only imbalanced data sets with multiple classes might have. Therefore, we are more interested to know the characteristics multi-class imbalanced data bring. It can help us understand the multi-class problem better and propose corresponding solutions. Incorporating our previous work about the diversity analysis of class imbalance learning with binary classes, we have following subtopics in multi-class imbalance learning for future studies,

- Can existing ensemble solutions handle multi-class imbalanced data sets as
well as binary-class ones?

• What kinds of multi-class imbalanced data make classification more difficult?

• What is the impact of diversity on single-class measures when there are more than one minority or majority class?

• Are there any good solutions to deal with the problems caused by multi-class?

5 Updated Timetable

Our research basically followed the timetable in progress report 4, except some small adjustments. Instead of comparing more state-of-art ensemble solutions for imbalanced data sets and proposing new algorithm by using active learning method, we concentrated on the theoretical understandings of the ensemble methods used in class imbalance learning and did diversity analysis on single-class measures. In addition, we submitted a paper to ICDM 2009, and it will be published in the workshop proceedings. A journal paper about “Relationships Between Diversity of Classification Ensembles and Single-Class Measures” is under way. In summary, we made the following progress from May to October 2009,

• Analyze diversity of classification ensembles on imbalanced data sets with two classes. Study the relationships between Q-statistic and single-class measures (recall, precision, F-measure) theoretically and experimentally.

• Conclude the above work and write a paper for ICDM2009. (Accepted as workshop paper)

• Enrich the above work and write a journal paper.

• Propose a new algorithm to improve the classical NCL algorithm.

Besides, we did more literature reading related to our work. Main literature review was briefly introduced in section 2 and 3. For the future research, we still stick to the plan in report 4 starting from October 2009 with small modifications. The updated timetable in this section includes our current research about the improved algorithm of NCL, and further multi-class studies of class imbalance learning.

From October to December 2009

• Finish journal paper writing (“Relationships Between Diversity of Classification Ensembles and Single-Class Measures”); give AUC/G-mean comparisons.

• Work on the proposed NCL algorithm and compare with related state-of-art algorithms (e.g. SMOTEBoost, under-sampling methods), including performance and execution time.
• Read related research about multi-class classification and its popular solutions.
• Give a presentation on ICDM09 workshop.

From January to April 2010

• More literature review of multi-class classification.
• Formulate and define multi-class classification in class imbalance learning.
• Explore if the existing ensemble methods are still effective to solve imbalanced data sets with multiple minority or majority classes. Which kind of data distribution is likely to cause problems? Which ensemble methods are sensitive to multi-class?
• Write progress report 6.

From May to July 2010

• Propose an ensemble solution for multi-class imbalanced data sets
• Start writing thesis.

From August to October 2010

• Finish thesis writing.
• Prepare for progress report 7 and viva.

6 Appendix: Published Paper in Workshop of ICDM2009

Title: Theoretical Study of the Relationship Between Diversity and Single-Class Measures for Class Imbalance Learning

Author: Shuo Wang and Xin Yao (Page: 17)

References


Theoretical Study of the Relationship Between Diversity and Single-Class Measures for Class Imbalance Learning

Shuo Wang and Xin Yao
The Centre of Excellence for Research in Computational Intelligence and Applications (CERCIA)
School of Computer Science, University of Birmingham
Edgbaston, Birmingham B15 2TT, UK
Email: {s.wang, x.yao}@cs.bham.ac.uk

Abstract—This paper presents the theoretical research about the relationship between diversity of classification ensembles and single-class measures that are commonly used in class imbalance learning. Although there have been studies on diversity and its links to overall ensemble accuracy, little work has been done on the impact of diversity on single-class performance measures in class imbalance learning. The study of class imbalance learning is important, because many real-world problems, such as those in medical diagnosis, fraud detection, condition monitoring, etc., have imbalanced classes, where a minority class is usually more important and interesting than the majority class. In order to gain a deeper understanding of ensemble learning for imbalanced classes, this paper studies the impact of diversity on single-class performance measures theoretically and empirically. One of the main objectives of this paper is to find out if and when ensemble diversity can improve the classification performance on the important (minority) class.

Keywords—diversity; ensemble learning; imbalanced data; single-class performance measure

I. INTRODUCTION

Class imbalance learning is a typical classification problem, where the amount or cost of data from each class within a data set is skewed. The minority class usually has higher cost and are more important. Ensemble is a big group of solutions to imbalanced data problems. They either rebalance the original training data from data level or emphasize examples of the minority class from algorithm level, aiming at improving accuracy over the high-cost class and not sacrificing accuracy over the entire data set, such as SMOTEBoost [1], DataBoost [2], and AdaCost [3].

It is commonly agreed that the success of ensemble is attributed to diversity – the degree of disagreement within ensemble. It has been proved empirically and theoretically in regression context by decomposing generalisation error [4]. A similar role of diversity has been found empirically in classification context. The theoretical research of the relationship between diversity and classification performance has been reformulated mainly in two ways so far. Tumer and Ghosh [5] derived an expression for added classification error of the ensemble by using posterior probabilities, but it is limited to classifiers that only output real-valued numbers, such as neural networks. Kuncheva et al. [6] discussed the link between Q-statistic (a widely-used diversity measure) and overall accuracy based on majority voting and binomial theorem under a set of assumptions. Both of them put effort on the relationship between diversity and “multi-class” (overall) accuracy without considering which class the input data point belongs to.

In class imbalance learning, however, multi-class accuracy is not a proper performance criterion any more [7] [8]. Some single-class measures are defined to evaluate the classifier performance on a specific class, such as recall, precision, and F-measure [9]. Hence, several questions are raised: What is the impact of diversity on single-class measures? Is it different from the impact on multi-class measures? Our short answer here is yes, and diversity has positive impact on the minority class. In addition, our previous work about diversity analysis on imbalanced data sets [10] [11] has shown that the minority class and the majority class have different behaviors as diversity is changing. So, the questions are meaningful and interesting. This paper focuses on the impact of diversity on single-class evaluation measures of balanced data and unbalanced data, and explores the role of diversity on the minority class. In order to find out if and when diversity is able to improve the performance in class imbalance learning area, it is important to build the link between diversity and single-class measures that are commonly used to evaluate the classifier performance over one class. Our work extends Kuncheva et al.’s analysis [6] about the link between Q-statistic and majority vote accuracy to the relationship between Q-statistic and single-class measures (recall, precision, F-measure). For simplicity, we only discuss two-class classification problems.

II. DEFINITIONS AND NOTATIONS

In a two-class data set with labels \{+1, -1\}, examples having positive labels belong to the positive class, and others having negative labels belong to the negative class. Let \( Z = \{z_1, \ldots, z_N\} \) be the training data set. For each classifier \( D_i \ (i = 1, \ldots, L) \), we define \( y_{ji} = 1 \) if \( D_i \) recognizes \( z_j \) correctly, and 0 otherwise. Similarly for ensemble, we define \( y_{j,ens} = 1 \) if majority voting result is correct for \( z_j \), and 0 if
it is incorrect. For the convenience of calculation later, $L$ is restricted to an odd number. Some notations and definitions used in diversity analysis are given in this section.

1) $Q$-statistic [12]: a pairwise measure of diversity. For two classifiers $D_i$ and $D_k$, the Q statistics is expressed 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 examples $z_j$ of $Z$ for which $y_{j,i} = a$ and $y_{j,k} = b$. It is widely used because of its easy-understanding form. In addition, the possible range of $Q$ is not affected by the individual accuracies easily [13] [14].

Another reason that Q is chosen in the Kuncheva’s study is because it is designed for synthetic examples with fixed settings, and illustrated that there is no clear pattern of the relationship between $P_{maj}$ and Q’s. However, the Q’s with high $P_{maj}$ tend to achieve negative values. Motivated by the first part of study, they 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”. The two patterns are assumed to be the “best case” and the “worst case”. Under the assumptions of each pattern, the link between Q-statistic and $P_{maj}$ was found.

3) $p$: the accuracy of the classifier individuals. It is assumed that all $L$ classifiers have the same individual accuracy in Kuncheva’s work.

To measure the classification performance on the minority class, recall, precision, and F-measure are normally used for performance comparison. They are defined based on the $2 \times 2$ confusion matrix in Table I.

4) Recall of positive class (denoted as $Rec_p$)

$$Rec_p = TP / \left( TP + FN \right) \quad (2)$$

It is the accuracy within positive class (TP rate). Similarly, the recall of negative class (denoted as $Rec_n$) is defined as,

$$Rec_n = TN / \left( TN + FP \right) \quad (3)$$

5) Precision of positive class (denoted as $Acc_p$)

$$Acc_p = TP / \left( TP + FP \right) \quad (4)$$

It is the proportion of real positive examples within all examples predicted as positive, representing how accurate the learning model is to predict this class. Similarly, the precision of negative class (deoted as $Acc_n$) is defined as,

$$Acc_n = TN / \left( TN + FN \right) \quad (5)$$

6) F-measure (denoted as $F$) [9]

$$F = \frac{\left(1 + \beta^2\right) \cdot recall \cdot precision}{\beta^2 \cdot recall + precision} \quad (6)$$

where $\beta$ corresponds to relative importance of precision and recall, and it is usually set to 1. $F$ value incorporates both precision and recall on a specific class ($F_p$ – positive, $F_n$ – negative), in order to measure the “goodness” of a learning algorithm for the class.

III. THE RELATIONSHIP BETWEEN Q-STATISTIC AND MAJORITY VOTE ACCURACY IN KUNCHEVA’S STUDY

Kuncheva et al.’s study [6] builds a connection between Q-statistic and overall accuracy produced by majority vote. It has two main contributions. Firstly, they gave several synthetic examples with fixed settings, and illustrated that there is no clear pattern of the relationship between $P_{maj}$ and Q’s. However, the Q’s with high $P_{maj}$ tend to achieve negative values. Motivated by the first part of study, they 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”. The two patterns are assumed to be the “best case” and the “worst case”. Under the assumptions of each pattern, the link between Q-statistic and $P_{maj}$ was found.

A. Pattern of Success (Best)

Under the pattern of success, no correct votes are wasted. The expression of $P_{maj}$ with respect to Q is,

$$P_{maj} = \frac{L}{l+1} \frac{1 - Q}{2 - Q} \quad (7)$$

where $l = \lfloor L/2 \rfloor$ for the convenience. $P_{maj}$ 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 (Worst)

Under the pattern of failure, correct votes are wasted to the greatest extent. The expression of $P_{maj}$ with respect to Q is,

$$P_{maj} = \frac{L}{(l-1)(2-Q)} - \frac{l}{(L-l)} \quad (8)$$

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

Detailed discussion can be found in her work [6]. Based on the two patterns, it shows diversity is not always beneficial to ensemble performance. There is some “bad” diversity. When diversity is “good”, such as the pattern of success, diversity brings performance improvement. When it is “bad”, diversity causes performance deterioration. The pattern of failure is one of such cases. Most of the cases in real-life problems will fall between the two extreme patterns.
IV. THE RELATIONSHIP BETWEEN Q-STATISTIC AND SINGLE-CLASS MEASURES

The relationship in single-class case is not as explicit as the one in multi-class explained in Section III [10] [11]. Considering those single-class measures and the overall accuracy are correlated, it is possible to utilize the above link to help our study. In this section, we discuss how Q-statistic impacts the tendencies of single-class measures theoretically, which include recall, precision, and F-measure.

Suppose the data set has \( N \) data points, where \( N_p \) points belong to the positive data set \( Z_p \), and \( N_n \) points belong to the negative data set \( Z_n \). The imbalance rate is denoted as \( \theta \) and calculated by \( \theta = N_p/N \). If the class distribution is balanced, \( \theta \) is nearly equal to 0.5. Otherwise, we assume that the positive class is minority, the negative class is majority, and \( \theta < 0.5 \). Two probabilities can be approximated, \( p\{z_j \in Z_p\} = \theta \) and \( p\{z_j \in Z_n\} = 1 - \theta \). Furthermore, we denote three probabilities for later use:

\[
p_1 = p\{y_j,ens = 1 \cap z_j \in Z_p\} \quad (9)
\]
\[
p_2 = p\{y_j,ens = 0 \cap z_j \in Z_n\} \quad (10)
\]
\[
\eta = p\{z_j \text{ is predicted as positive}\} = p_1 + p_2 \quad (11)
\]

A. Recall Vs. Single-class Q

Let us recall the “pattern of success” and “pattern of failure” in Kuncheva et al.’s work. Their discussion focused on the classification accuracy over two classes. If we change the whole context into single-class, their analysis still holds true. However, all the two-class measures should be adjusted to the single-class measures accordingly. \( P_{maj} \) should be majority vote accuracy on the corresponding class – recall, and diversity measure \( Q \) is only calculated within the data with the same class label. Therefore, we obtain a direct relationship between recall and single-class Q-statistic (\( Q_p \) or \( Q_n \)).

In the context of single class, recall is monotone decreasing with single-class \( Q \) under the pattern of success; recall is monotone increasing with single-class \( Q \) under the pattern of failure. Generally speaking, diversity has a direct impact on recall. When diversity is “beneficial” to the ensemble system, increasing diversity helps identify more data belonging to the class. Otherwise, less data will be recognized. What about the other measures?

B. Recall, Precision, F-measure Vs. Q

The single-class measures emphasize different performance aspects of a learner. They have different definitions, but are correlated to each other. There is a trade-off between recall and precision. F-measure is decided by both recall and precision. To simplify the relationship problem, we firstly derive the expressions of recall (Rec), precision (Acc), and F-measure (F) with respect to diversity value \( Q \) under some assumption. In this section, it is assumed that an example \( z_j \) is classified correctly by the ensemble is independent of its real class label. The assumption is not appropriate for a very imbalanced data set, within which the data from minority class are harder to be recognized than the data from majority class. However, as data get less imbalanced, the dependency should get weaker. The analysis here can be regarded as the extreme case for balanced data sets. It is therefore verified empirically on a balanced data set. We just use “positive” symbol to stand for a single class without discriminating positive or negative, since the derivation for the negative class will be the same. The relationship without the assumption will be discussed in the next section.

Under the assumption, we further get the probabilities from Eq.(9) and Eq.(10).

\[
p_1 = p\{z_j \in Z_p\} \cdot p\{y_j,ens = 1\} = \theta P_{maj} \quad (12)
\]
\[
p_2 = p\{z_j \in Z_n\} \cdot p\{y_j,ens = 0\} = (1 - \theta) (1 - P_{maj}) \quad (13)
\]

According to the definition of recall and precision introduced in section II, we obtain

\[
Rec_p = \frac{p_1}{p_1 + p_2} = \frac{\theta P_{maj}}{\theta P_{maj} + (1 - \theta) (1 - P_{maj})} \quad (14)
\]
\[
Acc_p = \frac{P_{maj}}{P_{maj} + (1 - P_{maj})} \quad (15)
\]

1) Pattern of success: In the best case with no wasted vote, \( Rec_p \) holds the same expression of Eq.(7) from Eq.(14). Eliminating \( P_{maj} \) from Eq.(15) and Eq.(7),

\[
Acc_p = \frac{L \theta (1 - Q)}{I(1 - \theta)(2 - Q) + L(2 \theta - 1)(1 - Q)} \quad (16)
\]

\( Acc_p \) is a monotone decreasing function of \( Q \) as the derivative of \( Acc_p \) is always negative with respect to \( Q \). This result supports when \( Q \) gets larger, both recall and precision values will get smaller, and therefore F-measure will get smaller. In summary, under the condition with “good” pattern, increasing diversity can improve performance on a single class, which includes recall, precision and F-measure.

2) Pattern of failure: In the worst case where a lot of votes are wasted, \( Rec_p \) holds the same expression of Eq.(8). Eliminating \( P_{maj} \) from Eq.(15) and Eq.(8),

\[
Acc_p = \frac{L \theta (1 - Q)}{L(1 - 2\theta)(1 - Q) + L(2\theta - 1)\theta (2 - Q)} \quad (17)
\]

\( Rec_p \) and \( Acc_p \) are monotone increasing functions of \( Q \). When \( Q \) gets larger, both recall and precision values will get larger, and therefore F-measure will get better. Diversity will lead to deterioration of the performance on single class under this pattern. In summary, under the condition with “bad” pattern, increasing diversity can degrade the performance on a single class.

With the independence assumption, recall, precision, and F-measure have improvement or deterioration together depending on “good” or “bad” \( Q \). It is worth noting that the \( Q \) here is a two-class value.
To verify the arguments, we choose a simple balanced data set “Monks1” (“0” – assigned as positive class, “1” – assigned as negative class). It is separated into 124 training examples and 432 testing examples. Both training and testing sets are strictly balanced, where $\theta = 0.5$. In the ensemble, 9 decision trees are built independently by using Bagging strategy. To observe the changing tendencies of single-class measures relative to $Q$, we adjust $Q$ value by setting different sampling rates (10%, 50%, and 100%) for forming every training subset. Table II presents the average results of the measures after 10 runs under each rate.

As we expect, the changing tendencies of single-class measures have no difference between the two classes. Diversity is “harmful” to both of the classes for this data set. When $Q$ value is increasing from 26% to 59% (diversity degree gets lower), all the other 3 measures (recall, precision, F-measure) from both classes have significant improvement. It shows that the ensemble is in “bad” pattern. If data domain is imbalanced, the results will be different.

**C. Recall, Precision, F-measure Vs. $Q$ without Independence Assumption**

The independence assumption removes the trade-off between recall and precision. Without the assumption, their correlation makes them behave in different ways. Several situations need to be considered. Imbalanced data set is more suitable to this case. In this section, $P_{maj}$ is utilized to relate $Q$-statistic to single-class measures. It is reformulated in five different ways.

Starting from the definition of majority vote, the easiest way to reformulate $P_{maj}$ is,

$$ P_{maj} = (N_p \cdot Rec_p + N_n \cdot Rec_n) / N \tag{18} $$

Use Eq.(2) divided by Eq.(4),

$$ \frac{Rec_p}{Acc_p} = \frac{Rec_p + N_n (1 - Rec_n)}{N} \tag{19} $$

Eq.(19) presents the functional link among $Rec_p$, $Acc_p$, and $Rec_n$. Eliminating $Rec_n$ from Eq.(18) and Eq.(19), we obtain,

$$ P_{maj} = (1 - \theta) + \theta (2 - 1/Acc_p) Rec_p \tag{20} $$

Eq.(20) is the second form of $P_{maj}$, expressed by recall, precision, and imbalance rate. If imbalance rate is regarded as a fixed value, the equation is decided by two variables. It is not possible to remove either recall or precision from the equation and keep the other alone.

Similarly, the third form of $P_{maj}$ is expressed by precision and F-measure,

$$ P_{maj} = (1 - \theta) + \theta (2Acc_p - 1) / \left(\frac{2Acc_p}{F_p} - 1\right) \tag{21} $$

From the view of statistics, another link between recall and precision is found,

$$ \theta \cdot Rec_p = \eta \cdot Acc_p \tag{22} $$

Recall and precision are correlated by the probability of a learner labeling an input example as positive class. Eliminating $Acc_p$ from Eq.(20) and Eq.(22),

$$ P_{maj} = (1 - \theta) + \theta (2Rec_p - \eta/\theta) \tag{23} $$

Eliminating $Acc_p$ from Eq.(21) and Eq.(22),

$$ P_{maj} = 1 - (1 - F_p) (\theta + \eta) \tag{24} $$

Fig. 1 presents two special cases that we can observe from the above expressions of $P_{maj}$, when $Acc_p = 0.5$ and $Rec_p = 0$. Suppose the shadowed area on the left side represents minority data set in each graph, and the right side area is majority data set. As the data set is imbalanced, the shadowed area is always smaller than the other. The circle corresponds to the probability $\eta$ defined in Eq.(11). When $Acc_p = 0.5$, the total area that is correctly classified is the left half circle, plus the majority area apart from the right half circle. Therefore, $P_{maj}$ will be $1 - \theta$. When $Rec_p = 0$, the circle belongs to the majority area completely, but $Acc_p = 0$ at the same time. In this situation, the right graph shows that $P_{maj}$ will be the majority area excluding the circle, $1 - \theta - \eta$. It is consistent with Eq.(23).

Given the five different forms of $P_{maj}$ with two variables (Eq.(18) (20) (21) (23) (24)), the functional link of recall, precision and F-measure with respect to single-class $Q$ is achieved without the independence assumption. They do not
simply have the same performance behavior. Every two of them are correlated. The next problem is how to make use of this relationship in class imbalance learning area.

D. Diversity Impact on Imbalanced Data Sets

In this section, we generalize the two patterns proposed by Kuncheva to class imbalance learning with two hypotheses, and explain the impact of diversity on minority class based on section IV.C. Intuitively, two situations can be imagined for the minority and majority class respectively.

Hypothesis 1: Individual classifiers tend to misclassify or overfit the examples from the minority class, and very few of them are identified by over half of the individuals. Therefore, this class of data makes ensemble perform closer to the “pattern of success”, and the ensemble has low diversity degree. Increasing diversity on the minority class should have beneficial influence, particularly on the measure recall.

Hypothesis 2: Data from the majority class can provide sufficient information to learners. The individuals within the ensemble tend to make the same correct decision. A lot of correct votes are wasted. The majority class is more likely to belong to the “pattern of failure”. Increasing diversity may produce negative effect to recall.

The hypotheses are verified empirically on an imbalanced UCI data set “insurance” with imbalance rate 6.3%. The training set contains 5822 data examples, and the testing set contains 4000. The same experimental settings are applied. The only difference is the training subset is rebalanced, in order to avoid most of the minority data being ignored by learners and influencing our observation. The experiment is conducted firstly to give an insight of the changing tendency of recall of each class along with single-class Q. Table III presents the average results of 10 runs under each rate. Both Q_p and Q_n have significant increase as bootstrap rate varies from 10% to 100%. However, Rec_p and Rec_n are changing towards two different directions as we supposed at the beginning. The results are consistent with our hypotheses and the analysis in section IV.A. The minority class belongs to the “pattern of success”, and the majority class belongs to the “pattern of failure”. To get the information of other measures, the change of P\text{maj} is crucial. As Q_p and Q_n get larger,

Better P\text{maj}: For minority class, if P\text{maj} is enhanced, the most obvious result is the reduction of \eta from Eq.(23). Acc_p tends to increase no matter when Acc_p > 0.5 or Acc_p < 0.5 according to Eq.(20). How F-measure changes is still unclear. In this case, diversity enhancement means increasing the probability of classifying an example into the minority class, but sacrificing the classification precision. Given the hypotheses and similar analysis, measures on majority class should behave in an opposite way.

Worse P\text{maj}: For minority class, if P\text{maj} is reduced, probability \eta can either increase or decrease. If \eta increases, Acc_p will get smaller due to the reduction of recall from Eq.(22), which results in the decrease of F-measure. If \eta decreases, F-measure will get smaller from Eq.(24), but the change of Acc_p is unsure. In this case, diversity enhancement means the improvement of both recall and F-measure. Therefore, introducing diversity could be a good choice to deal with imbalanced data.

We continue the experiment on “insurance” data set, and further output the other measures (P\text{maj}, \eta, precision, F-measure) for each class in Table III. This data set is in the case with “Better P\text{maj}”, since it increases from 77% to 89%. The behavior of each single-class measure is consistent with our discussion in “Better P\text{maj}” case. For this data set, diversity is beneficial to the minority class, as both recall and F-measure have significant improvement and precision is not sacrificed too much. Recall is improved by a larger amount than the reduction of precision. Compared to the experiment on the balanced data set, the behavior of single-class measures between minority and majority class presents discrepancy, where they are not simply improving or reducing in the same direction.

V. Conclusions

This paper explores the relationship between diversity and single-class performance measures theoretically and empirically by extending Kuncheva et al.’s study [6]. Two situations are discussed. If we assume that the class label voted by the majority classifier members is independent of the expected class, which is likely to be the case for balanced data sets, recall, precision, and F-measure will have the same general behavior as Q changes. This is verified experimentally on a balanced data set. For imbalanced data sets, however, the independence assumption is not appropriate any more. The
Table III

<table>
<thead>
<tr>
<th>Bootstrap Rate</th>
<th>10</th>
<th>50</th>
<th>100</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Q_p$</td>
<td>0.4450 ± 1.4674E−3</td>
<td>0.6631 ± 6.6052E−4</td>
<td>0.7491 ± 3.3401E−4</td>
</tr>
<tr>
<td>$Rec_p$</td>
<td>0.4929 ± 8.4131E−4</td>
<td>0.2624 ± 1.8615E−4</td>
<td>0.2000 ± 1.8124E−4</td>
</tr>
<tr>
<td>$Acc_p$</td>
<td>0.12656 ± 5.59927E−5</td>
<td>0.14687 ± 1.35332E−5</td>
<td>0.15323 ± 1.12634E−4</td>
</tr>
<tr>
<td>$F_p$</td>
<td>0.20131 ± 1.35898E−4</td>
<td>0.18872 ± 3.11887E−5</td>
<td>0.17340 ± 1.20316E−4</td>
</tr>
<tr>
<td>$Q_{maj}$</td>
<td>0.68602 ± 2.83117E−4</td>
<td>0.74116 ± 1.54031E−4</td>
<td>0.83936 ± 6.82472E−4</td>
</tr>
<tr>
<td>$Acc_{maj}$</td>
<td>0.78909 ± 4.85177E−5</td>
<td>0.90337 ± 2.06353E−5</td>
<td>0.92996 ± 1.88991E−5</td>
</tr>
<tr>
<td>$F_{maj}$</td>
<td>0.86604 ± 2.00129E−5</td>
<td>0.92629 ± 4.72858E−6</td>
<td>0.93908 ± 5.16146E−6</td>
</tr>
<tr>
<td>$\eta$</td>
<td>0.20365 ± 4.85443E−5</td>
<td>0.86487 ± 1.32115E−5</td>
<td>0.88853 ± 1.39090E−5</td>
</tr>
</tbody>
</table>

majority vote accuracy is reformulated to relate Q-statistic to single-class measures. Considering the characteristics of imbalance distribution, we assume that the minority class belongs to the “good” diversity pattern, and the majority class belongs to the “bad” diversity pattern. “Good (bad)” diversity means increasing diversity degree is beneficial (harmful) to classification performance of the ensemble on a set of data. We find that single-class Q-statistic has a direct impact on measure recall. The change of precision and F-measure depends on the majority vote accuracy. We have verified our theoretical analysis experimentally on a real imbalanced data set. For the minority class, diversity introduces the risk of precision reduction, but F-measure tends to be improved. For the majority class, opposite results are obtained. Generally speaking, diversity has positive impact on the minority class at the risk of degrading performance on the majority class.

The findings presented in this paper have revealed the role of diversity in class imbalance learning using ensemble approaches. The future directions of this work include the development of novel ensemble learning algorithms 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 learning with more than two classes.

ACKNOWLEDGMENT

This work was partially supported by an Overseas Research Student Award (ORSAS) and a PhD scholarship from the School of Computer Science, University of Birmingham, UK.

REFERENCES


