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: \textit{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 \(Z = \{z_1, \cdots, z_N\}\) be the training data set. For each classifier \(D_i (i = 1, \cdots, L)\), we define \(y_{j,i} = 1\) 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,

\[
Q_{i,k} = \frac{N_{10}^i N_{01}^k - N_{11}^i N_{00}^k}{N_{10}^i N_{00}^k + N_{01}^k N_{11}^k}
\]

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 understanding form. In addition, the possible range of Q is not affected by the individual accuracies easily [13] [14].

Under the pattern of success, no correct votes are wasted. Under the pattern of failure, correct votes are wasted 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 \( \text{P}_{maj} \) was found.

A. Pattern of Success (Best)

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

\[
\text{P}_{maj} = \frac{L}{l + 1} \frac{(1 - Q)}{(2 - Q)}
\]

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 \( \text{P}_{maj} \) and Q’s. However, the Q’s with high \( \text{P}_{maj} \) 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 \( \text{P}_{maj} \) was found.

B. Pattern of Failure (Worst)

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

\[
\text{P}_{maj} = \frac{L}{(l - 1)} \frac{1}{(2 - Q)} - \frac{l}{(L - l)}
\]

\( \text{P}_{maj} \) 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.

\[
p_1 = p\{y_{j,ens} = 1 \cap z_j \in Z_p\}
\]

\[
p_2 = p\{y_{j,ens} = 0 \cap z_j \in Z_n\}
\]

\[
\eta = p\{z_j \text{ is predicted as positive}\} = p_1 + p_2
\]

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),

\[
p_1 = p\{z_j \in Z_p\} \cdot p\{y_{j,ens} = 1\} = \theta P_{maj}
\]

\[
p_2 = p\{z_j \in Z_n\} \cdot p\{y_{j,ens} = 0\} = (1 - \theta)(1 - P_{maj})
\]

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

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

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

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

\(Rec_p\) and \(Acc_p\) are monotone increasing functions of \(Q\). 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.(7) from Eq.(14). Eliminating \(P_{maj}\) from Eq.(15) and Eq.(7),

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

\(Rec_p\) 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 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,

\[
P_{maj} = \frac{(N_p \cdot Rec_p + N_n \cdot Rec_n)}{N}
\]  

(18)

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

\[
Rec_p/Acc_p = Rec_p + N_n (1 - Rec_n)/N
\]

(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,

\[
P_{maj} = (1 - \theta) + \theta (2 - 1/Acc_p) Rec_p
\]

(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,

\[
P_{maj} = (1 - \theta) + \theta (2Acc_p - 1)/\left(\frac{2Acc_p}{F_p} - 1\right)
\]

(21)

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

\[
\theta \cdot Rec_p = \eta \cdot Acc_p
\]

(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),

\[
P_{maj} = (1 - \theta) + \theta (2Rec_p - \eta/\theta)
\]

(23)

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

\[
P_{maj} = 1 - (1 - F_p) (\theta + \eta)
\]

(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), Eq.(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 failure”, 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 success”. 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 contains 4000. The same experimental settings are applied. For...
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 verified. 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


