Diversity Analysis on Imbalanced Data Sets by Using Ensemble Models

Shuo Wang and Xin Yao

Abstract— **Many real-world applications have problems when learning from imbalanced data sets, such as medical diagnosis, fraud detection, and text classification. Very few minority class instances cannot provide sufficient information and result in performance degrading greatly. As a good way to improve the classification performance of weak learner, some ensemble-based algorithms have been proposed to solve class imbalance problem. However, it is still not clear that how diversity affects classification performance especially on minority classes, since diversity is one influential factor of ensemble. This paper explores the impact of diversity on each class and overall performance. As the other influential factor, accuracy is also discussed because of the trade-off between diversity and accuracy. Firstly, three popular re-sampling methods are combined into our ensemble model and evaluated for diversity analysis, which includes under-sampling, over-sampling, and SMOTE [1] – a data generation algorithm. Secondly, we experiment not only on two-class tasks, but also those with multiple classes. Thirdly, we improve SMOTE in a novel way for solving multi-class data sets in ensemble model – SMOTEBagging.**

I. Introduction

IMBALANCED data sets (IDS) correspond to domains where there are many more instances of some classes than others. Classification on IDS always causes problems because standard machine learning algorithms tend to be overwhelmed by the large classes and ignore the small ones. Most classifiers operate on data drawn from the same distribution as the training data, and assume that maximizing accuracy is the principle goal [2], [3]. Many real-world applications encounter the problem of imbalanced data, such as medical diagnosis, fraud detection, text classification, and oil spills detection [4].

Some solutions to the class imbalance problem have been proposed at both data level and algorithm level. At the data level, various re-sampling techniques are applied to balance class distribution, including over-sampling minority class instances and under-sampling majority class instances [5], [6], [7], [8] . Particularly, SMOTE (Synthetic Minority Over-sampling Technique) [1] is a popular approach designed for generating new minority class data, which could expand decision boundary towards majority class. At the algorithm level, solutions are proposed by adjusting algorithm itself, including adjusting the costs of various classes to counter the class imbalance, adjusting the decision threshold, and recognition-based (i.e., learning from one class) rather than discrimination-based (two class) learning. When working with decision trees, we could also adjust the probabilistic estimate at the tree leaf [2] . Cost-sensitive learning and semi-supervised learning are related research on class imbalance learning.

Shuo Wang (email: s.wang@cs.bham.ac.uk) and Prof. Xin Yao (email: x.yao@cs.bham.ac.uk) are with the School of Computer Science, University of Birmingham, Birmingham, UK.

As one of the solutions, ensemble systems have been drawn more and more attention because of their flexible characteristics. Firstly, for ensemble itself, multiple classifiers could have better answer than single one. A lot of study has been working on ensemble models and proved that it can average prediction errors and reduce bias and variance of errors. Secondly, most current ensemble models have the same learning procedure – re-sampling, base learning algorithm, voting, but different strategies in each phase. Each phase provides a chance to make the model better for classifying minority class. For example, Bagging [9] and Boosting [10] are two of the most popular techniques. These methods operate by taking a base learning algorithm and invoking it many times with different training sets. Therefore, some algorithms are proposed based on these two ensemble models by changing their re-sampling methods, such as BEV (Bagging Ensemble Variation) [11], SMOTEBoost [1], and DataBoost [12]. More details will be introduced in the Section 2. In the second phase of constructing base learners, algorithm-level methods can be applied. There are also some voting strategies beneficial to minority class instead of standard majority voting, such as adjusting weights of each classifier according to different cost, distance of instances, and F-measure value [13], [14].

Performance of ensemble models is decided by two factors: accuracy of individual classifier and diversity among all classifiers. Diversity is the degree to which classifiers make different decisions on one problem. Diversity allows voted accuracy to be greater than that of single classifier. Among above ensemble solutions for imbalanced data sets, however, it is still not clear that how diversity affects classification performance especially on minority classes. Understanding of diversity on minority class can help us improve ensemble solutions better. In this paper, therefore, the goal is to discover the impact of diversity on imbalanced data sets. Inevitably accuracy analysis is involved. Particularly, firstly, we combine three popular re-sampling methods into our ensemble model based on Bagging for diversity analysis, which includes under-sampling, over-sampling, and SMOTE. Secondly, we experiment not only on two-class tasks but also those with multiple classes to make our analysis sound. Thirdly, we extend SMOTE in a novel way for solving multi-class data sets in ensemble model – SMOTEBagging.

Around our research problem, we consider the following questions in our analysis, which are also the contributions of

978-1-4244-2765-9/09/$25.00 ©2009 IEEE

this paper:

- What is the performance tendency under different "diverse" degree by using different re-sampling techniques in ensemble? Three basic re-sampling methods are included: under-sampling of majority, over-sampling of minority, SMOTE, which generates synthetic minority class instances.
- What is the difference or similarity of diversity between two-class cases and multi-class cases?
- Can SMOTE bring diversity into ensemble?

The paper is organized as follows: Section 2 discusses related work of ensemble in class imbalance learning. Section 3 describes our experimental design including three improved ensemble models – OverBagging, UnderBagging, and SMOTEBagging. Section 4 gives observations from experiments and analyzes experimental results. Finally, section 5 presents the conclusions.

II. Related Work

In this field, ensembles have been used to combine several classifiers, each constructed after over-sampling or under-sampling training data, in order to balance the class distribution [15]. Among different re-sampling techniques, random over-sampling and random under-sampling are the simplest ones to be applied by duplicating or eliminating instances randomly. To avoid overfitting of random over-sampling, SMOTE is proposed by Chawla [1], which is a popular method of over-sampling by generating synthetic instances. Generally, SMOTE generates synthetic instances in the following way:

SMOTE generates new synthetic minority examples by interpolating between minority examples that lie together. It makes the decision regions larger towards majority class and less specific. Synthetic examples are introduced along the line segment between each minority class example and one of its k minority class nearest neighbors. Its generation procedure for each minority class example can be explained as: firstly, choose one of its k minority class nearest neighbors. Then, take the difference between the two vectors. Finally, multiply the difference by a random number between 0 and 1, and add it to this example. One of its problems is that SMOTE can only solve two-class problems by adjusting generating rate (i.e., from 100 to 500) to rebalance class distribution. This would cause confusion if more than one minority class exist. In addition, SMOTE is sensible to data complexity of data sets.

Current ensemble solutions are mostly based on various re-sampling methods, such as SMOTEBoost [1], DataBoost [12], and BEV [11]. The first two improve Boosting by combining data generating methods. Instead of changing the distribution of training data by updating the weights associated with each example in standard Boosting, SMOTEBoost alters the distribution by adding new minority-class examples using the SMOTE algorithm. Experimental results indicate that this approach allows SMOTEBoost to achieve higher F-values than standard Boosting and SMOTE algorithm with a single classifier. DataBoost has a different goal – improve performance of minority class without sacrificing the performance of majority class. Therefore, hard instances from both majority class and minority class are identified. BEV use Bagging by under-sampling majority class.

A number of researchers have been working on this topic, however, very few discuss the diversity and give us a clear idea that "why the ensemble model can improve performance of minority". Therefore, in order to achieve our goal, we choose three re-sampling methods in our experiments based on Bagging ensemble model – random over-sampling, random under-sampling, SMOTE. The limitation of the above solutions is that they are designed and tested on two-class applications. So, we extend the three Bagging models to multi-class cases where multiple minority classes and multiple majority classes exist.

Class imbalance has its own evaluation criteria on minority class and whole data set. For evaluating performance of one class, recall, precision, and F-measure are commonly used. Recall values tell us how many minority class instances are identified in the end, but may sacrifice system precision by misclassifying majority class instances. For a two-class problem, if we assume positive class is the minority, then recall value is formulated as "$TP/\left(TP+FN\right)$", where TP denotes the number of "true positive" instances and FN denotes the number of "false negative" instances. Value of F-measure (or F-value) incorporates both precision and recall, in order to measure the "goodness" of a learning algorithm for the class. It is formulated as,

$$F-value=\frac{\left(1+\beta^2\right)\cdot recall\cdot precision}{\beta^2\cdot recall+precision} \tag{1}$$

where β corresponds to relative importance of precision ($TP/\left(TP+FP\right)$, FP is "false positive") and recall, and it is usually set to 1. For evaluating overall performance, geometric mean (G-mean) and ROC analysis are better choices. G-mean is geometric average of recall values of each class. In this work, we choose recall, F-measure and G-mean value to describe performance tendency at different diversity degrees. Q-statistics is selected as our diversity measurement because of its easily understood form [16]. For two classifiers L_i and L_k, Q-statistic value is,

$$Q_{i,k}=\frac{N^{11}N^{00}-N^{01}N^{10}}{N^{11}N^{00}+N^{01}N^{10}} \tag{2}$$

where N^{ab} is the number of training instances for which L_i gives result 'a' and L_k gives result 'b' (It is supposed that the result here is equal to 1 if an instance is classified correctly and 0 if it is misclassified). Then for an ensemble system with a group of classifiers, the averaged Q-statistics is calculated to express the diversity over all pairs of classifiers,

$$Q_{av}=\frac{2}{M\left(M-1\right)}\sum_{i=1}^{M-1}\sum_{k=i+1}^{M}Q_{i,k} \tag{3}$$

For statistically independent classifiers, the expectation of Q-value is 0. Q-value varies between -1 and 1. It will be

positive if classifiers tend to recognize the same instances correctly, and will perform negative if they commit errors on different instances [17]. The larger the value is, the less diverse classifiers are.

III. Experimental Design

This section presents our experimental design for diversity analysis on both two-class and multi-class data sets. We implemented three ensemble models, each using Bagging to integrate every individual classifier, but different re-sampling methods. They are referred to UnderBagging, OverBagging and SMOTEBagging respectively. Firstly, the description and definition of these models are given. Then, experimental configuration is presented. It is worth to note that the following experiments and corresponding analysis emphasize performance on minority more than majority class. The reason is that information provided by minority class is commonly more meaningful in real-world problems, although performance is influenced by the relative proportion of both minority class and majority class.

A. Notations and Three Bagging Models in Our Work

Suppose there are C classes. The i-th class has N_i number of training instances. Those classes are sorted by N_i such that for the i-th class and the j-th class, if $i < j$ then $N_i \leq N_j$. Therefore, N_C is the number of the class having the most instances. Moreover, suppose there are H minority classes and $(C - H)$ majority classes, which is defined manually. Now we construct each classifier in ensemble iteratively using subset S_k of training set S. M classifiers are built, $k = 1, 2, \ldots, M$.

1) UnderBagging and OverBagging: In UnderBagging, each subset S_k is created by under-sampling majority classes randomly to construct the k-th classifiers. In the similar way, OverBagging forms each subset simply by over-sampling minority classes randomly. After construction, majority vote is performed when a new instance comes. Each classifier gives its judgment. Final classification decision follows the most voted class. If a tie appears, then the class with minor instances is returned. The whole procedure could be described as 3 steps – "re-sampling, constructing ensemble, voting" from training phase to testing phase. Because there may be multiple minority and majority classes, it brings more difficulty to decide which re-sampling rate we should use. *How to decide re-sampling rate in multi-class cases?* In order to keep every subset having same number of instances from each class, we use a "uniform" way of controlling re-sampling rate $a\%$. It refers to sampling rate of class C, containing the most instances. Other $(C - 1)$ classes has re-sampling rate $(N_C/N_i) \cdot a\%$. 'a' ranges from 10 to 100. For example, when 'a' equals to 100, N_C instances are bootstrapped from class C which has the most instances firstly. For other classes from class 1 to class $(C - 1)$, each has sampling rate $(N_C/N_i) \cdot 100\%$. When 'a' equals to 10, $10\% \cdot N_C$ instances are bootstrapped from class C, and other classes have sampling rate $(N_C/N_i) \cdot 10\%$. This method builds subset with same number of each class. In the former case, all classes are over-sampled. In the second case, minority classes are more likely to be over-sampled or keep the same number, and majority classes are under-sampled. Therefore, as 'a' increasing, it is a procedure of changing ensemble from UnderBagging to OverBagging. We handle these two strategies in the same way. The algorithm detail is shown in Table I.

TABLE I
From UnderBagging to OverBagging

Training:
1. Let S be the original training set.
2. Construct subset S_k containing instances from all classes with same number by executing the following:
   2a. Set re-sampling rate at a%.
   2b. For each class i, re-sample instances with replacement at the rate of $(N_C/N_i) \cdot a\%$.
3. Train a classifier from S_k.
4. Repeat step 2 and 3 until k equals M.

Testing on a new instance:
1. Generate outputs from each classifier.
2. Return the class which gets the most votes.

Another advantage of this method is its convenience to analyze diversity and performance tendency by controlling the value of 'a'. In our experiments, 'a' is set at multiples of 10. In this way we can get 10 ensembles for one data set. We expect that smaller 'a' results in more diverse ensemble system. And actually that is the fact, which will be discussed in the following experiments. It is worth to note that the statement is not always true. The change of diversity may also depend on other factors, such as learning algorithm, size of data set and data complexity. Diversity degree is more easily influenced by nonlinear learning methods when re-sampling rate varies, such as decision tree and neural networks, but SVM is less sensitive to the number of training instances. However, the former type of learning algorithms is more often used in ensemble learning. Similarly, some data set properties may also slow down the changing of diversity, but general tendency is not influenced. It can be explained by equation (2). If decision tree or ANN is selected as base learner, increasing re-sampling rate makes classification boundary more and more specific. Then the value of "$N01 \cdot N10$" gets smaller, and causes Q-value becomes larger, which means the decrease of diversity.

2) SMOTEBagging: Different from UnderBagging and OverBagging, SMOTEBagging involves generation step of synthetic instances during subset construction. According to SMOTE, two parameters need to be decided: k nearest neighbors and the amount of over-sampling from minority class – N. In Chawla's paper, their implementation uses five nearest neighbors and set N at 100, 200, 300, 400 and 500. We cannot use this in our experiments directly because there may exist multiple minority classes. We must consider the relative class distribution among all minority classes after re-sampling instead of over-sampling each class independently

by using different N values. For example, minority class A has 10 instances and minority class B has 50 instances. We use the same N to over-sample both A and B. After that, the two classes are still "inner-imbalanced". To avoid it, we use a percentage value $b\%$ to control the number of new generated instances in each class. Every classifier has different 'b' values, which range from 10 to 100. Each possible value is the multiple of 10. The algorithm detail is shown in Table II.

TABLE II
SMOTEBAGGING

Training:
1. Let S be the original training set.
2. Construct subset S_k containing instances from all classes with same number by executing the following:
   2a. Re-sample class C with replacement at percentage 100%.
   2b. For each class $i\,(1, \ldots, C-1)$:
      Re-sample from original instances with replacement at the rate of $(N_C/N_i) \cdot b\%$.
      Set $N = (N_C/N_i) \cdot (1 - b\%) \cdot 100$.
      Generate new instances by using SMOTE (k, N).
3. Train a classifier from S_k.
4. Change percentage $b\%$.
5. Repeat step 2 and 3 until k equals M.

Testing on a new instance:
1. Generate outputs from each classifier.
2. Return the class which gets the most votes.

Note that after constructing a subset S_k, every class has the same number of instances N_C, and every minority class has the same percentage of new instances and original instances. To make our system more diverse, we use different percentage value when building each classifier. So, if we build 20 classifiers as ensemble members, every 10 classifiers have different $b\%$ from 10% to 100%.

B. Data Sets and Configuration

Our experiments test on 8 UCI data sets including 6 two-class data sets and 2 multi-class data sets. They are well chosen with various imbalance rate and data set size and concluded in Table III. Particularly, we treat the first four classes in Glass as minority classes, and the first eight classes in Yeast as minority classes. Therefore, Glass has four minority classes and two majority classes. Yeast has eight minority classes and two majority classes.

In the experimental study, C4.5 decision tree is used as base learner in all of ensemble strategies described in this section. 10-fold cross validation is performed on each data set by running 30 times. The test result is the average of 30 runs of 10 folds. Each ensemble model creates 20 classifier members.

C. Relationship Between Re-sampling and Diversity or Accuracy

Before our experiments, we need to clarify the relationship between re-sampling and diversity. Our diversity analysis is based on the adjustment of re-sampling rate in ensemble

TABLE III
EXPERIMENTAL DATA SETS

Data Set	Size	Attributes	Class	Class Distribution (from minority to majority)
Hepatitis	155	19	2	45:55
Heart	270	13	2	44:56
Liver	345	6	2	42:58
Pima	768	8	2	35:65
Ionosphere	351	34	2	35:65
Breast-w	699	9	2	34:66
Glass	214	9	6	4.2:6.0:8.0:13.6:32.7:35.5
Yeast	1484	8	10	0.3:1.3:2.0:2.5:3.0 3.4:11.0:16.4:28.9:31.2

models. However, we don't treat re-sampling rate and diversity as the same concept. When re-sampling rate changes, accuracy of each classifier and diversity are changing at the same time. It is obvious that accuracy varies with re-sampling because more instances are used for classification. Therefore, when we analyze the diversity in the next section, we don't ignore the influence of accuracy. To discriminate accuracy and diversity, we use the algorithm shown in Table I on single classifier firstly, and adjust re-sampling rate in the same way. The results show the relationship between re-sampling and accuracy before we do the diversity analysis. Figure 1 illustrates increasing tendency of output values (Recall and F-measure of minority, G-mean) by using one classifier in data set Breast-w. If we build only one classifier, classifier accuracy increases without diversity involved, caused by re-sampling rate. It results in the improvement of other metrics. Other data sets have similar results, which fluctuate in a much lower range than ensemble. More diversity analysis is given in section "Experimental Analysis".

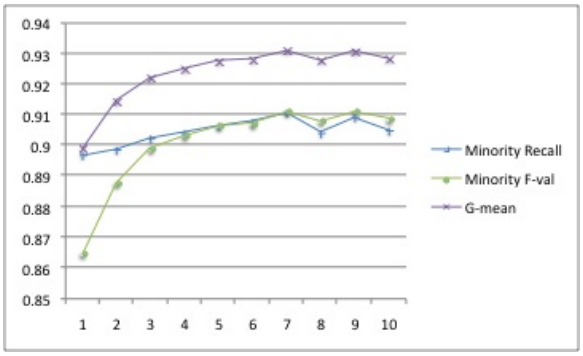

Fig. 1. Performance tendency of data set "Breast-w" by using single classifier. X-axis: the sampling rate from 10 to 100; Y-axis: the average values of final outputs. (Recall of Minority, F-value of Minority, G-mean)

IV. EXPERIMENTAL ANALYSIS

We firstly study the models UnderBagging and OverBagging on the eight data sets in Table III. In order to analyze diversity and performance tendency, percentage value 'a' is chosen from 10 to 100. When 'a' equals to 10, most classes from one data set will be under-sampled except the ones ten times smaller than the class with largest number of instances.

In this case, ensemble diversity should also be the largest. When 'a' equals to 100, all classes will be over-sampled to the largest number, in which case ensemble diversity should be the smallest, because a number of instances are duplicated. The fewer instances one class contains, the higher duplication degree is. In other words, overfitting is caused. We compare the results of recall values and F-values for each class, and G-mean as overall criterion. Different from other related studies, we calculate Q-statistics as diversity value not only on whole training data, but also on data in each class. This means every class has a diversity value, in order to make our experiments more accurate and convincing.

A. Two-class Data: From UnderBagging to OverBagging

In the two-class data sets, we give the curves to show the changes of each metric in Figure 2. X-axis presents the under-sampling percentage from 10 to 100, and Y-axis presents the average values of final outputs. However, for space considerations, we only put diversity results from data set "Pima" here in Table IV. Other five data sets perform similar on Q-statistic values. Q-statistic values of minority class and whole data set are both increasing as value 'a' becomes larger and larger, which means diversity is decreasing. In Figure 2, it is evident that recall value of minority class from five data sets out of six keeps decreasing when diversity becomes smaller and smaller. There is no phase of going up. Recall value of majority class performs in the opposite way, which keeps increasing. Data set "Ionosphere" is an exception. Recall value, however, can only tell us how many minority instances could be found (hit rate). F-value is more meaningful for most real world problems. F-value of minority class is the curve with circle marker '$\circ$' in the figure 2. As we can observe, none of F-values from six data sets decrease when diversity gets smaller during the first several steps. They all have a significant improvement at the first few points of x-axis. Then three of them start to decrease, and others stay at the same level. G-mean values presenting overall performance have similar tendency with F-values.

TABLE IV
Q-STATISTICS OF PIMA

Re-sample Percentage	Minority Q-statistic	Overall Q-statistic
10%	0.449	0.496
20%	0.513	0.570
30%	0.530	0.598
40%	0.543	0.618
50%	0.547	0.625
60%	0.549	0.628
70%	0.546	0.632
80%	0.552	0.634
90%	0.550	0.637
100%	0.552	0.638

The behavior of recall value is easy to understand. Higher diversity gives more chance to find out minority instances, and vice versa. At first, the re-sampling rate for majority class is low. One instance has lower probability to be classified as majority. In other words, system has a low accuracy on majority. Compared with single classifier in Figure 1, diversity exerts more significant influence on minority class than majority class. An instance is more likely to be classified as minority when accuracy is low. Therefore, recall of minority is comparatively high. As accuracy on majority and minority becomes higher, diversity goes down. Accuracy on minority also means overfitting, which causes low diversity and low recall. In fact, it can also be explained from the recall formulation ($recall = TP/(TP + FN)$) in section II. Imagine that classification boundary is getting more and more specific. TP get smaller and FN gets larger correspondingly because the number of minority instances is fixed. Too much duplication lowers the probability of classifying an instance as minority.

When discussing about diversity, we cannot ignore accuracy, because there is a trade-off between accuracy of each classifier and ensemble diversity [18], [17]. Assume accuracy and diversity have low-medium-high three levels respectively. Then there are the following possible statuses:

- Low accuracy, low diversity: every classifier is more likely to misclassify instances and makes the same errors. This rarely happens if a proper learning algorithm is chosen.
- Low accuracy, high diversity: every classifier is more likely to misclassify instances but makes different errors.
- High accuracy, low diversity: every classifier is more likely to make the same correct decision on instances.
- Medium accuracy, medium diversity: intermediate status between status 2 and 3.

During the analysis of F-values of minority class, the tendency can be explained based on the above statuses. At first, the classification capacity of ensemble system is in status 2. As re-sampling rate going up, status changes into 4. F-value is the geometric average of recall and precision. Recall is decreasing and precision is increasing, but accuracy is more influential so that F-value has improvement. Normally when re-sampling rate varies from 40% to 100%, F-value stops increasing or even starts decreasing, because the status changes from 4 to 3. Diversity factor is playing a more important role in the ensemble system. From this stand of view, the point with re-sampling rate 40% is better than the point with rate 100% for minority class, because they have similar F-values but the former case gets better recall value. In class imbalance field, high recall value is more useful than precision some times. For example, if we need to detect fraud, overfitting may harm fraud prevention, but recall can help us to find more potential fraud cases even if some of them are not. Therefore, status 4 with medium accuracy and medium diversity could be a better choice.

G-mean is actually the geometric average of recall value from each class. In the six cases, the increasing of majority recall value is faster than the decreasing of minority recall value. So, G-mean goes up at the first phase like F-value. In the second phase, the increasing speed slows down. G-mean values stop increasing or even start decreasing slightly.

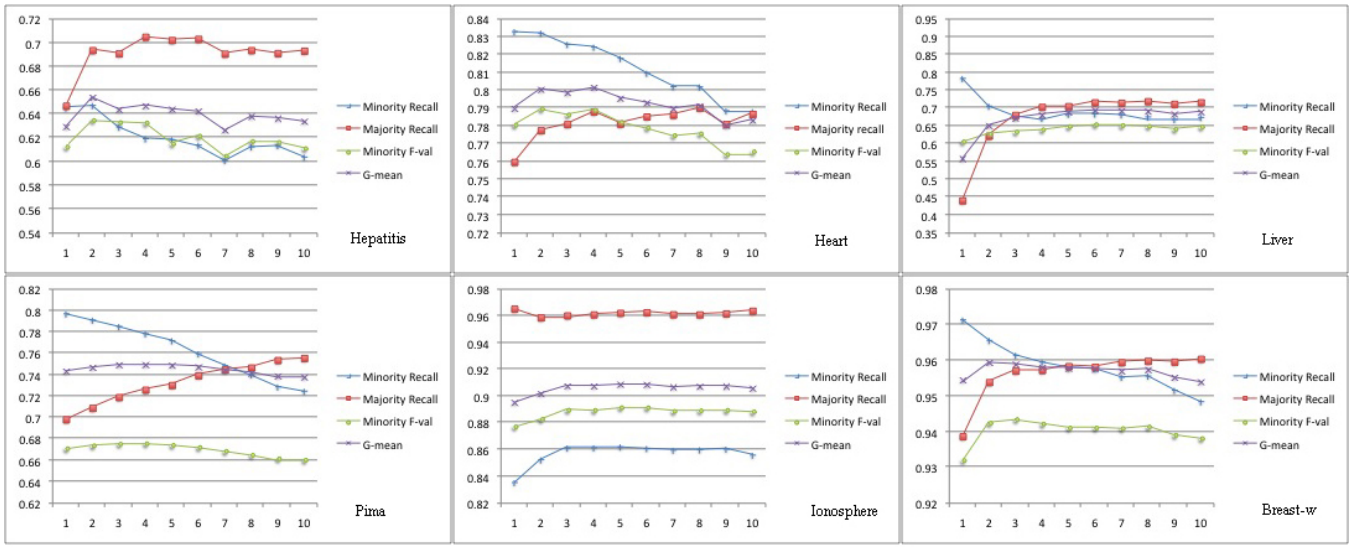

Fig. 2. Performance tendency of two-class data sets. X-axis: the sampling rate from 10 to 100; Y-axis: the average values of final outputs. (Recall of Minority and Majority, F-value of Minority, G-mean)

TABLE V

PERFORMANCE TENDENCY OF EACH CLASS IN MULTI-CLASS DATA SETS. FIRST COLUMN IS THE NUMBER OF CLASS SORTED BY IMBALANCE DEGREE FROM HIGHLY IMBLALANCED TO SLIGHTLY IMBALANCED. UP ARROW: SIGNIFICANT INCREASE; DOWN ARROW: SIGNIFICANT DECREASE.

Glass	Recall	F-val	Q-statistic
1	↓	↑↓	↑
2	↓	↑↓	↑
3	-	↑↓	↑
4	-	-	↑
5	↑	↑	↑
6	↑	↑	↑
Yeast	Recall	F-val	Q-statistic
1	↓	↑↓	↑
2	↓	↑↓	↑
3	↓	↓	↑
4	↓	↓↑	↑
5	↓	↓	↑
6	↓	-	↑
7	-	-	↑
8	-	↓	↑
9	↑	↑	↑
10	↑	↑	↑

TABLE VI

F-VALUE AVERAGES AND STANDARD DEVIATIONS OF 30 RUNS OF 10-FOLD CROSS-VALIDATION T TESTS OF THE CASE WITH THE BEST F-VALUE AND THE CASE WITH THE RE-SAMPLING RATE 100% FOR EACH MINORITY CLASS OF DATA SET "GLASS" AND "YEAST". SYMBOL "*" DENOTES STATISTICAL SIGNIFICANT DIFFERENCE WITH 90% OF CONFIDENCE. THE FIRST COLUMN LISTS THE NUMBERS OF MINORITY CLASSES.

Glass	Best F-val	F-val with 100% re-sampling rate	T
1	0.91067 ± 0.03586	0.85067 ± 0.07555	*
2	0.693 ± 0.09087	0.56500 ± 0.09475	*
3	0.39150 ± 0.07306	0.32367 ± 0.10052	*
4	0.87500 ± 0.04226	0.85467 ± 0.04261	
Yeast	Best F-val	F-val with 100% re-sampling rate	T
1	0.92467 ± 0.02630	0.86833 ± 0.06405	*
2	0.41706 ± 0.03578	0.31730 ± 0.02650	*
3	0.06608 ± 0.02364	0.03168 ± 0.01183	*
4	0.46140 ± 0.03892	0.42093 ± 0.04798	*
5	0.76751 ± 0.03747	0.72659 ± 0.01807	*
6	0.40013 ± 0.05017	0.37769 ± 0.05408	
7	0.76729 ± 0.01156	0.76696 ± 0.01350	
8	0.57919 ± 0.01252	0.56124 ± 0.01212	*

B. Multi-class Data: From UnderBagging to OverBagging

In the multi-class data sets, the performance tendency is more obvious, and similar with two-class data sets. Table V describes the changing by using up/down arrows. Mark "-" means there is not significant change. Double arrows show two changes happen sequentially. Recall and F-value are included.

In data set "Glass", the first four classes (No.1-4) are minority classes, sorted by imbalance rate. In the same way, the first eight classes (No.1-8) in "Yeast" are minority. Most recall values in minority classes are reducing. When the class is less imbalanced, the reducing speed slows down. We can also observe that most F-values in minority classes have a phase of decreasing, but not for the majority classes. T test with 90% of confidence between the case with the best F-value and the case with the highest re-sampling rate 100% is done in Table VI, in order to show that the best class performance does not appear in the case with high accuracy / low diversity. Proper diversity is necessary. Nine out of twelve minority classes have significant difference.

Between two-class and multi-class problems, diversity has similar impact on each class. The impact is weakened as the imbalance rate gets smaller for each class in the observations of multi-class. The imbalance rate here is a relative concept

within one data set, not an absolute value. In the first two two-class data sets, even if the data is not very imbalanced, the recall of minority still decreases significantly. If there exist multiple minority classes, less imbalanced minority class is more difficult to be influenced by diversity. Diversity is distracted on more comparatively imbalanced classes. There is an interactive influence among minority classes.

In summary, we have the following observations: recall values of minority classes keep decreasing while recall values of majority classes keep increasing as diversity is reducing. At the same time, F-values of minority classes and G-mean values perform two phases – increasing firstly and then have a reduction or stay at the same level. Finally, medium accuracy and medium diversity of an ensemble system could be a better choice in the field of class imbalance.

TABLE VII
EXPERIMENTAL RESULTS OF OVERALL PERFORMANCE ON MULTI-CLASS DATA SETS

Glass	G-mean	Overall Q-statistics
Over	0.927	0.664
SMOTE	0.960	0.621
Yeast	G-mean	Overall Q-statistics
Over	0.941	0.675
SMOTE	0.969	0.615

C. Multi-class Data: OverBagging and SMOTEBagging

In this section, we compare two models OverBagging and SMOTEBagging. We are interested in the questions that whether SMOTE brings diversity into ensemble model and whether the ensemble system has better performance. To find out the answer, we combine SMOTE algorithm into our Bagging model and extend it to solve multi-class data sets, which is described section 3. Because we do not analyze tendency in this part, all classes are over-sampled so that each has the same number of instances with the class having the most instances. OverBagging is same as the one in previous experiments whose re-sampling percentage is 100%. In SMOTEBagging, we use a percentage value $b\%$ to control the number of instances from each class that is used for generating new instances for one subset. This part of experiments is based on the multi-class data sets, so as to compare the outputs among different minority classes and keep results consistent. Minority classes from one data set have the same data properties.

Table VII presents overall performance of data set "Glass" and "Yeast". From Table VII, both data sets have a reduction on Q-statistics and an improvement on G-mean in SMOTEBagging. Generating synthetic instances generates more diverse ensemble systems. Table VIII and Table IX are the results of minority classes from each data set. In "Glass", three in four minority classes have lower Q-statistic values in model SMOTEBagging. All of the three classes have higher recall values. In "Yeast", seven in eight minority classes have lower Q-statistic values, and six in the seven achieve better recall except the last one. One interesting observation in this data set is that all classes get higher F-value in SMOTEBagging rows. For more imbalanced classes, F-values enhance more; for less imbalanced ones, F-values enhance less. However, we cannot get strong conclusion that there is a relationship between imbalance rate and changing degree of F-value. Generally speaking, SMOTE injects diversity into ensemble system in most cases and improve its overall performance.

V. CONCLUSIONS

In this paper, the effect of diversity is studied empirically on eight UCI data sets with three ensemble models. The results suggest that diversity influences recall value significantly. Basically, larger diversity causes better recall for minority but worse recall for majority classes. As diversity decreases, recall values tend to be smaller for minority classes. This is because diversity enhances the probability of classifying an instance as minority when accuracy is not high enough. Tendency of F-measure and G-mean are decided by classifier accuracy and diversity together. In our opinion, the best F-measure value and G-mean value don't appear at the status with high accuracy and low diversity, but the status with medium accuracy and medium diversity. Secondly, to make our research more convincing, we experiment on both two-class data sets and multi-class data sets. Three ensemble models are proposed to solve data with multiple classes. Multi-class is more flexible and beneficial to our diversity analysis. According to our results, diversity has similar impact on each class between two-class and multi-class, but the impact is weakened by the falloff of imbalance rate in the observations of multi-class, not for two-class. There is interaction among classes. If some classes have higher probability to be identified as, then other classes have lower probability. Finally, SMOTE does bring diversity into ensemble system in multi-class data sets. Both overall performance (G-mean) and diversity degree have improvement. Multi-class studied in this paper contains only two data sets. This is sufficient for exploring the diversity, but may need more to analyze the difference of performance between two-class and multi-class. It is an interesting topic in our future work. As part of future work, better evaluation criteria for multi-class also need to be explored.

ACKNOWLEDGMENT

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

REFERENCES

[1] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, pp. 341–378, 2002.

[2] N. V. Chawla, N. Japkowicz, and A. Kotcz, "Editorial: Special issue on learning from imbalanced data sets," *SIGKDD Explor. Newsl.*, vol. 6, no. 1, pp. 1–6, 2004.

[3] S. Visa and A. Ralescu, "Issues in mining imbalanced data sets- a review paper [c]," in *Proceedings of the Sixteen Midwest Artificial Intelligence*, 2005.

TABLE VIII

EXPERIMENTAL RESULTS OF MINORITY CLASSES ON GLASS

Class	Algorithm	Minority Recall	Minority F-val	Minority Q-statistic
1	Over	0.932	0.873	0.797
	SMOTE	0.917	0.861	0.820
2	Over	0.710	0.616	0.675
	SMOTE	0.750	0.663	0.636
3	Over	0.526	0.366	0.463
	SMOTE	0.573	0.333	0.292
4	Over	0.861	0.850	0.844
	SMOTE	0.887	0.871	0.832

TABLE IX

EXPERIMENTAL RESULTS OF MINORITY CLASSES ON YEAST

Class	Algorithm	Minority Recall	Minority F-val	Minority Q-statistic
1	Over	0.918	0.876	0.636
	SMOTE	1	0.947	0.678
2	Over	0.394	0.331	0.704
	SMOTE	0.512	0.428	0.609
3	Over	0.062	0.057	0.827
	SMOTE	0.078	0.069	0.743
4	Over	0.475	0.426	0.825
	SMOTE	0.497	0.434	0.779
5	Over	0.777	0.736	0.767
	SMOTE	0.784	0.748	0.718
6	Over	0.419	0.387	0.772
	SMOTE	0.467	0.402	0.713
7	Over	0.839	0.760	0.679
	SMOTE	0.865	0.761	0.626
8	Over	0.587	0.563	0.764
	SMOTE	0.520	0.569	0.687

[4] N. Japkowicz and S. Stephen, "The class imbalance problem: A systematic study," *Intelligent Data Analysis*, vol. 6, no. 5, pp. 429 – 449, 2002.

[5] G. E. A. P. A. Batista, R. C. Prati, and M. C. Monard, "A study of the behavior of several methods for balancing machine learning training data," *Special issue on learning from imbalanced datasets, Sigkdd Explorations*, vol. 6, no. 1, pp. 20–29, 2004.

[6] H. Han, W.-Y. Wang, and B.-H. Mao, "Borderline-smote: A new over-sampling method in imbalanced data sets learning," in *Advances in Intelligent Computing*, 2005, pp. 878–887.

[7] 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.

[8] I. Tomek, "Two modifications of cnn," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 6, no. 11, pp. 769–772, 1976.

[9] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, no. 2, pp. 123–140, 1996.

[10] 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.

[11] 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.

[12] 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.

[13] R. Valdovinos and J. 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.

[14] N. V. Chawla and J. Sylvester, "Exploiting diversity in ensembles: Improving the performance on unbalanced datasets," *Multiple Classifier Systems*, vol. 4472, pp. 397–406, 2007.

[15] V. Garcia, J. Sanchez, R. Mollineda, R. Alejo, and J. Sotoca, "The class imbalance problem in pattern classification and learning."

[16] G. U. Yule, "On the association of attributes in statistics," *Philosophical transactions of the Royal society of London*, vol. A194, pp. 257–319, 1900.

[17] 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.

[18] G. Brown, J. L. Wyatt, and P. Tino, "Managing diversity in regression ensembles," *The Journal of Machine Learning Research*, vol. 6, pp. 1621 – 1650, 2005.