Bias-Variance Analysis of ECOC and Bagging Ensembles Using Neural Nets

Cemre Zor[1], Terry Windeatt[1] and Berrin Yanikoglu[2]

[1]Center for Vision, Speech and Signal Processing, University of Surrey, UK, GU2 7XH
(c.zor, t.windeatt)@surrey.ac.uk
[2]Sabanci University, Tuzla, Istanbul, Turkey, 34956
berrin@sabanciuniv.edu

Abstract. to be filled in.

1 Introduction

Within the machine learning research, many techniques have been proposed in order to understand and analyse the success of ensemble classification methods over single-classifier classifications. One of the main approaches considers tightening the generalization error bounds by using the margin concept [ ]. Though being theoretically interesting, bounds found are not usually tight enough to be used in practical design issues. [ ]. Bias and variance analysis is another method used to show why ensembles work well. In this paper, we try to analyse the success of bagging [ ] and Error Correcting Output Coding (ECOC) [ ] as ensemble classification techniques, by using Neural Networks (NNs) as the base classifiers within the bias and variance framework of James [ ]. As the characteristics of the ensemble depend on the specifications of the base classifiers it is composed of, having a detailed look at the parameters of the base classifiers within the bias-variance analysis is of importance. Similiar work of Valentini and Dietterich on Support Vector Machines (SVMs) can be found in [ ].

ECOC is an ensemble technique [3], where multiple base classifiers are trained according to a preset binary *code matrix*. Consider an ECOC matrix C, where a particular element $C_{ij}\ \epsilon\ (+1, -1)$ indicates the desired label for class i, to be used in training the base classifier j. The base classifiers are the dichotomizers which carry out the two-class classification tasks for each column of the matrix, according to the input labelling. Each row, called a *codeword*, indicates the desired output for the whole set of base classifiers for the class it is indicating. During decoding, a given test sample is classified by computing the similarity between the output (hard or soft decisions) of each base classifier and the codeword for each class by using a distance metric, such as the Hamming (L1 Norm) or the Euclidean (L2 norm) distance. The class with the minimum distance is then chosen as the estimated class label. The method can handle incorrect base classification results up to a certain degree. Specifically, if the minimum Hamming distance (HD) between any pair of codewords is d, then up to $\lfloor (d-1)/2 \rfloor$ single bit errors can be corrected.

As for bias and variance analysis, after the initial work of Geman [ ] on the regression setting using squared-error loss, others like Breiman [ ], Kohavi and Wolpert [ ], Dietterich and Kong [ ], Friedman [ ], Wolpert [ ], Heskes [ ], Tibshirani [ ], Domingos and James [ ] have tried to extend the analysis for the classification setting. One of the problems with the above definitions of bias and variance is that most of them are given for specific loss functions such as the zero-one loss, and it is hard to generalize them for all the other loss functions. Usually, new definitions are driven for each loss function. Even if the definitions are proposed to be general, they may fail to satisfy the additive decomposition of the prediction error defined in [?]. The definition of James [?] has got advantages over the others as it proposes to construct a scheme which is generalizable to any symmetric loss function. Furthermore, it constructs two more concepts called "systematic effect" and "variance effect" which help assure the additive prediction error decomposition for general loss functions and realize the effects of bias and variance on the prediction error.

Some characteristics of the other definitions which make James' more preferable for us are as follows: -Dietterich allows a negative variance and it is possible for the Bayes classifier to have positive bias. -Experimentally, the trends of Breiman's bias and variance graphs closely follow James' systematic effect and variance effect ones respectively. However, for each test input pattern, Breiman separates base classifiers into two sets, as biased and unbiased; and considers each test pattern only to have either bias or variance accordingly. -Kohavi and Wolpert also assign a nonzero bias to the Bayes classifier and the Bayes error is absorbed within the bias term. Although it helps avoid the need to calculate the Bayes error in real datasets through making insufficient assumptions, it is not preferable as the bias term comes out to be too high. -The definitions of Tibshirani, Heskes and Breiman are hard to be generalized and extended for the loss functions other than the ones they were defined in. -Friedman proposes that bias and variance do not always need to be additive.

After all these differences, it should also be noted that the characteristics of bias and variance terms Domingos has defined are actually close to James', although the decomposition can be considered as being multiplicative [ ].

In the literature, attempts have also been made to explore the bias-variance characteristics of ECOC and bagging ensembles. Examples can be found in [james] [kongdie] [breiman][terry][ydoes]. In this paper a detailed analysis on ECOC and bagging ensembles using NNs as the base classifiers through changing the parameters, namely nodes and epochs, has been given.

2 Bias and Variance Analysis of James

I need around half a page here.

3 Experiments

3.1 Experimental Setup

Experiments have been carried out on 5 artificial and 4 UCI MLR [5] datasets. 3 of the artificial datasets have been created according to Breiman's description in []. Detailed information about the sets can be found in table []. The optimization method used in NNs is the Levenberg-Marquart (LM) technique; the level of training (epochs) varies between 2 and 15; and the number of nodes between 2 and 16.

The ECOC matrices are created by randomly assigning binary values to each matrix cell. Hamming Distane is used as the metric in the decoding stage and the number of columns each ECOC has is set to 50. 50 is also the number of classifiers used in bagged ensembles and the total number of base classifiers that the bias-variance analysis has been carried out on. That is, for each of the three settings of single classifier, bagging and ECOC classifications, 50 base classifiers are created. Each base classifier is either a single classifier, or an ensemble consisting of 50 bagged classifiers or ECOC matrices of 50 columns.

Experiments have been performed 10 times for the artificial datasets by using different training & test data, and ECOC matrices in each run; and the results are averaged. The number of training patterns per base classifier is equal to 300; and the number of test patterns to be used in every run is 18000. For the UCI datasets having separate test sets, the analysis has been done just once for the single classifier and bagging settings, and 10 times with different matrices for the ECOC setting. Here, bootstrapping is applied while creating the base classifiers, as it is expected to be a close enough approximation to random & independent data generation from a known underlying distribution [ ]. As for the UCI datasets without separate test sets, the $ssCV$ cross-validation method of Webb and Conilione [ ], which allows the usage of the whole dataset both in training and test stages, has been implemented. In $ssCV$, the shortcomings of the hold-out approach like the usage of small training and test sets; and the lack of inter-training variability control between the successive training sets has been overcome [ ]. In our experiments, we set the inter-training variability constant δ to $1/2$.

The Bayes error is analytically calculated for the artificial datasets, as the underlying likelihood probability distributions are known. As for the real datasets, the motivation is to find the best optimal classifier parameters giving the lowest error rate possible, through cross-fold validation (CV); and then to use these parameters to construct a classifier which is expected to be close enough to the Bayes classifier. This classifier is then used to calculate the output probabilities per pattern in the dataset. For this, we first find an optimal set of parameters for RBF SVMs by applying 10 fold CV; and then, obtain the underlying probabilities by utilizing the leave-one-out approach. Using the leave-one-out approach instead of training and testing the whole dataset with the found CV parameters helps us avoid overfitting. It is assumed that the underlying distribution stays almost constant per each fold of the leave-one-out procedure.

Table 1. Summary of the 5 UCI MLR datasets

	Type	# Training Samples	# Test Samples	# Attributes	# Classes	Bayes Error
TwoNorm [ ]	Artificial	300*	18000*	20	2	
ThreeNorm [ ]	Artificial	300 *	18000*	20	2	
RingNorm [ ]	Artificial	300 *	18000*	20	2	
ArtificalMultiClass1	Artificial	300*	18000*	2	5	
ArtificalMultiClass2	Artificial	300 *	18000*	3	9	
Glass Identification	UCI	214	-	10	6	
Dermatology	UCI	358	-	33	6	
Segmentation	UCI	210	2100	19	7	
Yeast	UCI	1484	-	8	10	

*: The training and test samples for the artificial datasets change per each base classifier and per each run respectively.

3.2 Results

When the graphs driven from the experiments are explored, some clear trends are observed. The observations are stated below together with some graphs that are representatives of results.

- Prediction errors obtained by using bagging and ECOC ensembles are always lower than the ones of the single classifier; and the reduction in the error is almost always coming out as a result of reductions both in variance effect and in systematic effect. Among these two, the reductions in the variance effect have greater magnitude. This observation means that the contributions of bias and variance to the prediction error are smaller when ensembles are used. In [ ] and [ ], bagging and ECOC are also stated to have low variance in the additive error decomposition.
- When the single classifier case is taken into account; we see that variance effect, which is the contribution of variance to the prediction error, does not necessarily follow the trend of variance. It happens especially when the number of nodes and epochs is small, that is when the network is relatively weak. In this scenario, the variance goes lower and the variance effect goes higher. This is actually an expected observation as one would expect having high variance to help hitting the right target class, when the network is relatively less decisive. Ensemble methods do not show this property usually. This is due to the fact that they already make use of variance in an amount which is enough to account for the above mentioned situation for weak networks, as they themselves are composed of many base classifiers.
- In the above mentioned scenario of variance effect showing opposite trend to variance, the bias-variance trade-off can be observed. At the points where the variance effect increases, bias effect decreases in an amount enough to reveal an overall decrease in the prediction error. However, these points are not necessarily the optimal points when the prediction error is considered.
- In sense of the prediction error, the convergence points of single classifiers to the optimal are usually at higher epochs than those of bagged ensembles, per

Fig. 1.

node. The points where ECOC ensembles are converging are mostly at even lower points than the ones of bagging. Meanwhile, the prediction errors also come out in the same descending order: single classifier, bagging and ecoc. The only exceptions to these happen during the 2 class problems, where bagging and ECOC are showing quite similar trends, bagging sometimes doing better. This is because, under this circumstance ECOC with HD decoding can be considered as bagging: Although bootstrapping has not been used, random initial weights given by LM algorithm are expected to give a similar effect [?].

- It is also almost always the case that the prediction error of ECOC converges to a minima in 2 nodes, using n epochs whereas the prediction error of a single classifier converges to its optima, which has got a value higher than or equal to the one of ECOC, using 4, 8 or 16 nodes and greater than or equal to n epochs. Therefore, instead of a single classifier trained with high number of epochs and nodes, an ECOC trained with parameters of quite low values gives out better results. The trend is similar when bagging is considered. It is usually in between the single classifier and ECOC, based on both accuracy and convergence.
- Put information about the speed of bagging vs ECOC.
- Tell about overfitting in discussion, didnt happen in out experiments but might happen with higher nodes or with svms?

4 Discussion

References

1. Tumer K., Ghosh, J.: Error Correlation and Error Reduction in Ensemble Classifiers. Connection Science, Special Issue on Combining Artificial Neural Networks: Ensemble Approaches, 8(3) 385–404 (1996)
2. Escalera, S., Tax, D. M. J., Pujol, O., Radeva, P., Duin, R. P. W.: Subclass Problem-Dependent Design for Error-Correcting Output Codes. In: IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 30, no. 6, pp. 1041-1054 (2008)
3. Dietterich, T.G., Bakiri, G.: Solving Multi-class Learning Problems via Error-Correcting Output Codes. J. Artificial Intelligence Research 2. 263–286 (1995)
4. Allwein, E., Schapire, R., Singer, Y.: Reducing Multiclass to Binary: A Unifying Approach for Margin Classifiers. JMLR 1. 113–141 (2002)
5. Asuncion, A., Newman, D.J.: UCI Machine Learning Repository, http://www.ics.uci.edu/~mlearn/MLRepository.html. School of Information and Computer Science, University of California, Irvine, CA (2007)
6. Escalera, S., Pujol, O., Radeva, P.: On the Decoding Process in Ternary Error-Correcting Output Codes. In: CIARP, vol. 4225, pp. 753–763 (2006)
7. Escalera, S., Pujol, O., Radeva, P.: Recoding Error-Correcting Output Codes. Proceedings of the 8th International Workshop on MCS, vol. 5519, pp. 11–21 (2009)

8. Alpaydin, E., Mayoraz, E.: Learning error-correcting output codes from data. In: Proc. Int. Conf. Neural Networks (ICANN) (1999)
9. James, G. M.: Majority Vote Classifiers: Theory and Applications. PhD Thesis, Department of Statistics, University of Standford (1998)
10. James, G. M., Hastie, T.: The Error Coding Method and PICT's, Computational and Graphical Statistics, vol. 7, no. 3, pp. 377-387 (1998)
11. Windeatt, T., Ghaderi R.: Coding and Decoding Strategies for Multi-class Learning Problems. Information Fusion, 4(1), pp. 11-21 (2003)
12. Dietterich, T.G., Bakiri, G.: Solving Multi-class Learning Problems via Error-Correcting Output Codes. J. Artificial Intelligence Research 2. 263–286 (1995)
13. Allwein, E., Schapire, R., Singer, Y.: Reducing Multiclass to Binary: A Unifying Approach for Margin Classifiers. JMLR 1. 113–141 (2002)