

Classification Ensemble That Maximizes the Area Under Receiver Operating Characteristic Curve (AUC)

Eunsik Park¹ and Y-c Ivan Chang²

¹Chonnam National University, Gwangju, Korea

²Academia Sinica, Taipei, Taiwan

Outline

- **Motivation** : It is well known that there is no single classification rule that is overwhelmingly better than others in all situations.
- **Ensemble** : Thus, the classification ensemble method, which integrates many classification methods together, can usually improve on the performance of individual classification rules.
- **Our Proposal** : we study the ensemble method that integrates non-homogeneous classifiers constructed by different methods, and target at maximizing the area under receiver operating characteristic curve (**AUC**).
- **Evaluation** : AUC is used because of its threshold independent character and computational convenience that can help to resolve the difficulty due to non-homogeneity among base classifiers.
- **Numerical Study** : Ensemble is applied to some real data sets. The empirical results clearly show that our method outperforms individual classifiers.

Review - Ensemble Algorithms

- Some methods that can also be viewed as ensemble algorithms have been already proposed such as voting, bagging, and even more boosting-like algorithms.
- However, most of them are aggregations of results from homogeneous classification rules, which may somewhat improve the overall performance, but on the other hand, they usually share the same shortcomings as those of their base classifiers.
- Among them, the bagging algorithm that relies on the idea of bootstrapping is an typical example since it only reduces the variation of the final classifier, but not its bias (Bauer & Kohavi, 1999).

Review - Building a Classification Rule

- There are many factors usually considered in building a classification rule such as loss/objective function, feature selection, threshold determination, subject-weighting, and these factors are treated differently in different classification methods.
- Moreover, there are many measures of classification performance. Depending on the criterion chosen, the final ensemble will also perform differently. As mentioned, individual algorithms are usually designed for some specific demands depending on classification problems.
- These heterogeneities usually increase the difficulties of constructing ensemble of non-homogeneous classifiers.

Our Proposal - Ensemble

- In order to incorporate non-homogeneous classifiers and take the advantage of the specific natures of individual classification methods, we take their function-value outputs, instead of their predicted labels, as new features to construct the new ensemble such that the final classifier is more robust than individual classifiers.
- Moreover, in order to prevent the ambiguity of voting due to threshold selection, we would like to adopt some threshold-independent measure as our targeted performance measure.
- Therefore, we use AUC because AUC shares the threshold-independent advantage of ROC curve, while provides us with an easy operation nature (Pepe, 2003 & Fawcett, 2006).

Ensemble Based on AUC - I

- We study an ensemble method, targeting at maximizing the area under ROC curve, with non-homogeneous classifiers as its ingredients. Since all classifiers are applied to the same data set, their outputs should be correlated.
- It is, however, difficult to have information about the correlation among outputs from different classifiers, which makes the ensemble method dependant on such an information less useful here.
- Hence, the PTIFS method of Wang et al. (2007) is adopted in our paper as the integration method due to its nonparametric character.

Ensemble Based on AUC - II (PTIFS)

- A parsimonious threshold-independent protein feature selection (**PTIFS**) method through the area under receiver operating characteristic (ROC) curve. Bioinformatics, 2007, Vol. 23, 2788-2794, Zhanfeng Wang, Yuan-chin I. Chang, Zhiliang Ying, Liang Zhu, Yanning Yang.
- Starting from an anchor feature, the PTIFS method selects a feature subset through an iterative updating algorithm. Highly correlated features that have similar discriminating power are precluded from being selected simultaneously.

Ensemble Based on AUC - III

- Each base-classifier will be optimally trained if it has such an option available, and the features selected can be different if the classifier itself has an internal feature selection function. In other words, our method allows each classifier to do its best in all possible senses.
- Then we take their classification function output values as new features to conduct final ensemble while maximizing AUC as the final objective.
- That is, our method can integrate nonhomogeneous base-classifiers and each classifier is well-trained before being included into the final ensemble.

Setup - I

- The gene selection is based on the logistic regression analysis assuming significance level, α , is 0.01.
- 50% of total samples are randomly selected as the training set, and the rest samples are assigned to the testing set.
- Real Datasets

Table: Number of samples/genes in two data sets

data sets	sample size	normal samples	cancer samples	genes
hepatocellular carcinoma	60	20	40	7,129
breast cancer	102	62	40	1,368

Setup - II

- Ensemble Method
 - PTIFS (Wang et al, 2007) : Non-parametric algorithm maximizing AUC, LARS type, deal with high-dimensional data.
 - Su and Liu (Su & Liu, 1993) : Maximizing AUC under normal assumption, Based on LDA.
 - LogitBoost (Friedman, Hastie, Tibshirani, 2000)
 - AdaBoost (Freund & Schapire, 1996)
 - AdaBag (Breiman, 1996) : Bootstrapping
- Individual Classifier
 - SVM (Support Vector Machine)
 - KDA (Kernel Fisher Discriminant Analysis)
 - LDA (Linear Fisher Discriminant Analysis)
 - DDA (Shrinkage Discriminant Analysis - Diagonal) : Schafer and Strimmer, 2005
 - QDA (Quadratic Fisher Discriminant Analysis)

Results I - Ensemble Comparison : hepatocellular carcinoma data

Table: Misclassification rate, AUC, sensitivity and specificity(iteration=1,000)




Ensemble	Classifiers*	Misclassification rate		AUC		Sensitivity		Specificity	
		Train	Test	Train	Test	Train	Test	Train	Test
	SVM	0.00(0.00)	0.18(0.06)	1.00(0.00)	0.90(0.04)	1.00(0.00)	0.74(0.15)	1.00(0.00)	0.88(0.06)
	KDA	0.48(0.09)	0.34(0.07)	0.56(0.05)	0.50(0.00)	0.36(0.10)	0.15(0.02)	0.68(0.10)	0.66(0.06)
	LDA	0.03(0.03)	0.13(0.05)	1.00(0.00)	0.94(0.04)	0.99(0.04)	0.82(0.12)	0.97(0.03)	0.90(0.06)
	DDA	0.07(0.03)	0.11(0.06)	0.96(0.03)	0.95(0.04)	0.88(0.07)	0.83(0.13)	0.96(0.03)	0.93(0.06)
	QDA	0.08(0.04)	0.16(0.07)	0.60(0.06)	0.63(0.10)	0.92(0.08)	0.85(0.12)	0.92(0.03)	0.85(0.09)
PTIFS	All	0.00(0.00)	0.16(0.06)	1.00(0.00)	0.91(0.04)	1.00(0.00)	0.81(0.13)	1.00(0.00)	0.86(0.10)
Su & Liu	All	0.00(0.01)	0.00(0.01)	1.00(0.00)	1.00(0.05)	0.99(0.03)	0.99(0.03)	1.00(0.01)	1.00(0.01)
LogitBoost	All	0.00(0.00)	0.17(0.08)	1.00(0.00)	0.88(0.07)	1.00(0.00)	0.61(0.22)	1.00(0.00)	0.94(0.11)
AdaBoost	All	0.00(0.00)	0.18(0.06)	1.00(0.00)	0.81(0.06)	1.00(0.00)	0.77(0.13)	1.00(0.00)	0.85(0.10)
AdaBag	All	0.00(0.00)	0.18(0.06)	1.00(0.00)	0.81(0.06)	1.00(0.00)	0.77(0.13)	1.00(0.00)	0.85(0.10)

Results II - Ensemble Comparison : breast cancer data




Table: Misclassification rate, AUC, sensitivity and specificity (iteration=1,000)

Ensemble	Classifiers*	Misclassification rate		AUC		Sensitivity		Specificity	
		Train	Test	Train	Test	Train	Test	Train	Test
	SVM	0.00(0.00)	0.18(0.04)	1.00(0.00)	0.90(0.03)	1.00(0.00)	0.80(0.09)	1.00(0.00)	0.85(0.06)
	KDA	0.00(0.00)	0.26(0.06)	0.84(0.18)	0.72(0.08)	1.00(0.01)	0.66(0.11)	1.00(0.00)	0.80(0.07)
	LDA	0.00(0.01)	0.16(0.04)	1.00(0.00)	0.91(0.03)	0.99(0.02)	0.81(0.09)	1.00(0.01)	0.86(0.05)
	DDA	0.10(0.03)	0.13(0.04)	0.95(0.02)	0.93(0.03)	0.84(0.06)	0.80(0.08)	0.94(0.03)	0.92(0.05)
	QDA	0.04(0.02)	0.16(0.05)	0.53(0.03)	0.59(0.07)	0.97(0.04)	0.86(0.10)	0.96(0.02)	0.84(0.07)
PTIFS	All	0.00(0.00)	0.17(0.04)	1.00(0.00)	0.90(0.03)	1.00(0.00)	0.76(0.10)	1.00(0.00)	0.88(0.07)
Su & Liu	All	0.00(0.01)	0.00(0.01)	1.00(0.00)	1.00(0.00)	1.00(0.01)	1.00(0.01)	1.00(0.01)	1.00(0.01)
LogitBoost	All	NA(NA)	NA(NA)	NA(NA)	NA(NA)	NA(NA)	NA(NA)	NA(NA)	NA(NA)
AdaBoost	All	0.00(0.00)	0.19(0.06)	1.00(0.00)	0.80(0.07)	1.00(0.00)	0.70(0.16)	1.00(0.00)	0.88(0.08)
AdaBag	All	0.00(0.00)	0.20(0.06)	1.00(0.00)	0.80(0.07)	1.00(0.01)	0.69(0.16)	1.00(0.00)	0.88(0.07)





References I

-  Bauer, E. and R. Kohavi (1999).
An Empirical Comparison of Voting Classification Algorithms:
Bagging, Boosting, and Variants
Machine Learning, 30, 105 – 139.
-  Breiman, (1996).
Bagging redictors
Machine Learning, 24, 123–140.
-  Buhlmann, P. and B. Yu (2003).
Boosting With the L2 Loss: Regression and Classification
Journal of the American Statistical Association., 98, 324 – 339.

References II

-  Dietterich, T. (2008).
Ensemble Methods in Machine Learning
Multiple Classifier Systems 2000, Eds. J. Kittler and F. Roil, LNCS 1857, 1 – 15, Springer-Verlag, Berlin Heidelberg.
-  Fawcett, T. (2006).
An introduction to ROC analysis
Pattern Recognition Letters, 27, 861–874.
-  Freund, Y. and R. Schapire (1996).
Experiments with a new boosting algorithm
Machine Learning: Proceedings of the Thirteenth International Conference, 148 – 156, 1996.

References III

-  Friedman J., T. Hastie and R. Tibshirani (2000).
Additive logistic regression: a statistical view of boosting
Annals of Statistics, 28(2), 337 – 407.
-  Lim, T. and Y. Shih (2000).
A comparison of Prediction Accuracy, Complexity, and Training
Time of Thirty-three Old and New Classification Algorithms
Machine Learning, 40, 203 – 229.
-  Pepe, M. S. (2003).
*The statistical evaluation of medical tests for classification and
prediction*, New York: Oxford.
-  Su, J. Q. and Liu, J. S. (1993).
Linear combinations of multiple diagnostic markers.
Journal of the American Statistical Association 88, 1350-1355.

References IV



Wang, Z., Y. Chang, Z. Ying, L. Zhu, and Y. Yang (2007).

A parsimonious threshold-independent protein feature selection method through the area under receiver operating characteristic curve.

Bioinformatics 23, 2788 – 2794.



Su, J. Q. and Liu, J. S. (1993).

Linear combinations of multiple diagnostic markers.

Journal of the American Statistical Association 88, 1350-1355.



Schafer, J. and Strimmer, K. (2005).

A Shrinkage Approach to Large-Scale Covariance Matrix Estimation and Implications for Functional Genomics.

Statistical Applications in Genetics and Molecular Biology 4, article 32.

References V



Sharma, P., Sahni, N. S., Tibshirani, R., Skaane, P., Urdal, P., Berghagen, H., Jensen, M., Kristiansen, L., Moen, C., Sharma, P., Zaka, A., Arnes, J., Børresen-Dale, T. and Lonneborg, A. (2005) Early detection of breast cancer based on gene-expression patterns in peripheral blood cells.
Breast Cancer Res., 7, R634 – R644.