Available online at www.sciencedirect.com
SCIENCE DIRECT®
ELSEVIER
Information Fusion 6 (2005) 3–4
INFORMATION FUSION
www.elsevier.com/locate/inffus

Guest editorial

# Diversity in multiple classifier systems

Fifteen years ago, the reader would have questioned a statement that an ensemble of classifiers is generally better than a single classifier. Now this is the prevailing opinion based on a substantial amount of theoretical and empirical evidence, and on the availability of smart training methods for classifier ensembles.

It is intuitively clear that an ensemble of identical classifiers will be no better than a single member thereof. If we have "the perfect classifier", then no ensemble is needed. If the ensemble members are imperfect, they should be different so that at least some of them are correct where the others are wrong. We call this loosely specified property *diversity*, and set off to explore why and how it works for the success of the ensemble, if at all.

Diversity does work! Classifier ensembles that enforce diversity fare better than ones that do not. The classical example is boosting versus bagging, the two currently most successful ensemble strategies. Both approaches build the ensembles by training each classifier on a bespoke data set. Boosting promotes diversity actively whereas bagging relies on independent re-sampling from the training set. Boosting has been crowned as the "best off-the-shelf classifier" by Leo Breiman himself, the creator of bagging. Numerous theoretical studies explain the success of Boosting by proving bounds and margins on its error. The secret lies with the ingenious construction of the subsequent training data sets so that classifiers trained on them form a diverse ensemble. Can we not measure and use diversity explicitly to create better ensembles?

Our previous studies led us to the somewhat surprising and discouraging conjecture that diversity is not unequivocally related to the ensemble accuracy. Is this a fault of defining and measuring diversity? Should diversity be always related to accuracy? Should diversity be perceived as a property of the set of classifiers or should it be related to the combination method too? This special issue, consisting of seven original contributions, looks into diversity through a magnifying glass. The efforts of leading researchers and teams are being presented together in search of answers to some of the above questions.

The first paper is a survey on diversity in classification and regression. Brown et al. have taken up the difficult task to tidy the diversity drawer of multiple classifier systems cabinet. They start with diversity in regression ensembles which allows for a much more rigorous treatment than diversity in classification ensembles. Their systematic approach leads them to propose a taxonomy of methods for creating diversity in classifier ensembles. As a result of this in-depth look into the core concept, the authors are able to offer useful tips for measuring diversity as well as provide a more formal analysis of diversity.

Windeatt studies measures of diversity in relation to the complexity of the base classifiers in the ensemble (of neural networks). He proposes a new measure that is better related to the classification accuracy than some of the most commonly used measures when the complexity of the base classifiers is varied. The experimental results suggest that using diversity might be a way towards developing reliable methods for tuning the complexity of the base classifiers.

Gal-Or et al. explore the effectiveness of diversity measures in classifying television viewers for the purposes of targeted advertising. They investigate the case of two classes with unequal misclassification costs and identify diversity measures that are good predictors of the classification accuracy. The authors expand the existing research by drawing a parallel between the behaviours of the diversity measures for oracle representation (0 = incorrect label/1 = correct label) and direct representation where the two classes are coded as 0 and 1.

Banfield et al. propose an interesting performance-based diversity measure with a direct application to pruning the ensemble, called "thinning". A rich experiment has been carried out using 22 publicly available data sets. The ensemble size was chosen to be 1000 (classifiers generated through variants of bagging), "thinned" down to 100. The results support the authors' thesis that thinning reduces computational complexity of the ensemble without a significant adverse effect on the accuracy.

Ruta and Gabrys summarize methods for selecting classifiers to form an ensemble from a set of trained

1566-2535/$ - see front matter 
doi:10.1016/j.inffus.2004.04.009

classifiers. Contrary to the findings of Banfield et al., here the authors do not advocate using diversity measures to gauge the ensemble performance and propose instead to base the choice directly on the majority vote accuracy. The experiments, using 27 publicly available data sets, are rich and thorough as well. The contradiction between the results of the two studies is only superficial. Banfield et al. consider large ensembles (of 1000 classifiers) and their reduction down to a relatively large figure of 100 classifiers. Ruta and Gabrys consider ensembles of 15 classifiers from which to select, where exhaustive search is also a possible option. The diversity measure in the Banfield's study is incorporated in the "thinning" procedure and is not used as an overall criterion that is supposed to replace the evaluation of the ensemble accuracy. The two studies make an interesting compound suggesting that there is no point in substituting a diversity measure as a selection criterion but there are other ways in which diversity may be useful in the selection process.

Tsymbal et al. propose a feature selection framework for classifier ensembles. They calculate a "fitness function" for each classifier composed of an accuracy term and a diversity term. The diversity term reflects the contribution of the classifier to the ensemble diversity. Various measures of diversity, ensemble combination methods and feature selection algorithms are investigated through an experiment with 21 data sets.

Melville and Mooney suggest that diversity should be measured with respect to the ensemble prediction. They proceed to design a simple and appealing ensemble training algorithm, called DECORATE, which adds one classifier at a time for creating the training set of the new classifier using diversity explicitly. The original training set is augmented by a set of new data points, called "diversity data", whose labels are decided so as to be most diverse from the ensemble prediction. A large experiment involving 33 data sets has been carried out to demonstrate that DECORATE compares favourably with the best available ensemble methods such as bagging and boosting.

We had the luxury of great many submissions and thus were faced with pleasant but yet challenging task of having to select among them the very best for this special issue. I wish to thank the authors of all the submitted papers for considering this special issue as a possible forum for presenting their work. I acknowledge with sincere thanks the invaluable help of all the reviewers.

At the conception of this special issue, the main question for me was "Is the quest for diversity leading us to a dead end?" The sheer amount of interesting research that was submitted as a response to the call for papers is a clear answer "no". The variety of inspiring ideas within the submissions is a clear answer "no". The strong positive statements by most of the studies in this issue show that there is a way forward. The one negative statement is a warning that this way may still be bumpy. The abundance of profound expertise on the subject is another clear sign that diversity is presently an active pursuit. And the ambivalence of the opinions makes it a bigger challenge and more fun.

In closing, this special issue is being offered with the sincere belief that it will indeed turn out to be a significant milestone in the path towards a better understanding of the diversity concept and how it can be exploited in improving performance robustness in real world applications.

Ludmila I. Kuncheva
*School of Informatics*
*University of Wales, Bangor, Dean Street*
*Bangor Gwynedd LL57 1UT*
*United Kingdom*
*Tel.: +44-1248-383661; fax: +44-1248-361429*
*E-mail address:* l.i.kuncheva@bangor.ac.uk

Available online 19 May 2004