A Study on Class-Specifically Discounted Belief for Ensemble Classifiers

Ronnie Johansson
School of Humanities and Informatics
University of Skövde
Skövde, Sweden
Email: ronnie.johansson@his.se

Henrik Boström
School of Humanities and Informatics
University of Skövde
Skövde, Sweden
Email: henrik.bostrom@his.se

Alexander Karlsson
School of Humanities and Informatics
University of Skövde
Skövde, Sweden
Email: alexander.karlsson@his.se

***Abstract*—Ensemble classifiers are known to generally perform better than their constituent classifiers. Whereas a lot of work has been focusing on the generation of classifiers for ensembles, much less attention has been given to the fusion of individual classifier outputs. One approach to fuse the outputs is to apply Shafer's theory of evidence, which provides a flexible framework for expressing and fusing beliefs. However, representing and fusing beliefs is non-trivial since it can be performed in a multitude of ways within the evidential framework. In a previous article, we compared different evidential combination rules for ensemble fusion. The study involved a single belief representation which involved discounting (i.e., weighting) the classifier outputs with classifier reliability. The classifier reliability was interpreted as the classifier's estimated accuracy, i.e., the percentage of correctly classified examples. However, classifiers may have different performance for different classes and in this work we assign the reliability of a classifier output depending on the class-specific reliability of the classifier. Using 27 UCI datasets, we compare the two different ways of expressing beliefs and some evidential combination rules. The result of the study indicates that there is indeed an advantage of utilizing class-specific reliability compared to accuracy in an evidential framework for combining classifiers in the ensemble design considered.**

Keywords: ensemble classifiers, random forests, evidence theory, Dempster-Shafer theory, combination rules

I. Introduction

Information fusion researchers have pointed out the potential benefits of learning predictive models to improve fusion-based state estimation [1]. Conversely, machine learning (or data mining) researchers have acknowledged the contribution of information fusion to the construction of predictive models [2]. A predictive model (or classifier) is constructed from examples with known class labels to suggest the most likely class for novel, i.e., previously unseen, examples. Many different ways of constructing predictive models have been proposed, and it is widely acknowledged that there is no single method that is optimal for all possible problems [3]. Instead, the fact that individual classifiers generated in different ways or from different sources are *diverse*, i.e., make different classification errors, can be exploited by combining (or fusing) their outputs to improve the classification performance [4], [5]. There has been a substantial amount of work in the field of machine learning on developing different methods to exploit the idea of learning such *ensembles* of classifiers, including varying the set of training examples given to the learning algorithm or randomizing the process for generating each classifier, see e.g. [6].

The main focus of previous research on ensembles of classifiers has been on the generation of the constituent classifiers, rather than on the way in which they are combined. Similarly to the learning methods, no single combination rule can be expected to be optimal for all situations, but instead each rule has its individual strengths and weaknesses. Still, it may be the case that some of the rules are better suited than others to combine the output of certain types of ensemble classifier. Most commonly, straightforward fusion approaches, such as voting, are employed (explained in e.g., [4], [7]–[10]). However, some authors have proposed using Shafer's evidence theory to combine the ensemble classifiers by expressing their outputs in terms of *mass functions* [10]–[14], representing the belief of each classifier. Originally, Dempster's rule was proposed as *the* means to combine mass functions [15]. Since then, many alternative combination rules have been proposed to counter seemingly deficient properties of Dempster's rule, such as Yager, Dubois-Prade, and the modified Dempster's rule [16].

In a previous article [17], we compared different evidential combination rules for a specific ensemble design. The results indicated that some of the combination rules seem to be more appropriate than others for ensemble classifiers. The combination rules we selected were Dempster's and the modified Dempster's rule.

In the previous work, we employed a simple class belief representation that discounted each mass function with the reliability of the classifier. The reliability was expressed as the estimated accuracy of each classifier. However, it should be noted that this reliability measure is independent of the predicted class. Hence, it does not allow for expressing that certain classifiers are more reliable when predicting certain classes. In this work, we add an alternative way to represent belief which is based on the classifier's classification accuracy for each class.

In the next section, we give a brief description of ensemble classifiers (random forests in particular) and discuss how the output of members of ensembles commonly are combined. In Section III, we give a brief introduction to evidential theory and present the combination rules that are compared in this study. In Section IV, we discuss previous approaches to evidence-based ensemble combination. In Section V, we describe the experimental setup of the study and present results from using the evidential combination rules and belief representations for random forests. Finally, in Section VI, we present the main conclusions from this study and point out some directions for future research.

II. Ensembles of Classifiers

A. Basic Terminology

A classifier e is a function that maps a vector of attribute values **x** (also called *example*) to classes $c \in C = \{c_1, \ldots, c_l\}$. An ensemble classifier consists of a set of classifiers, $E = \{e_1, \ldots, e_m\}$, whose output is dependent on the outputs of the constituent classifiers.

B. Random Forests

Classification trees have many attractive features, such as allowing for human interpretation and hence making it possible for a decision maker to gain insights into what factors are important for particular classifications. However, recent research has shown that significant improvements in predictive performance can be achieved by generating large sets of models, i.e., *ensembles*, which are used to form a collective decision on the value for the dependent variable [6]. It can be shown that as long as each single model performs better than random, and the models make independent errors, the resulting error can in theory be made arbitrarily small by increasing the size of the ensemble. However, in practice it is not possible to completely fulfill these conditions, but several methods have been proposed that try to approximate independence, and still maintain sufficient accuracy of each model, including the introduction of randomness in the process of selecting examples and attributes when building each individual model. One popular method of introducing randomness in the selection of training examples is bootstrap aggregation, or *bagging*, as introduced by Breiman [18]. It works by randomly selecting n examples with replacement from the initial set of n examples, leading to that some examples are duplicated while others are excluded. Typically a large number (at least 10) of such sets are sampled from which each individual model is generated. Yet another popular method of introducing randomness when generating classification trees is to consider only a small subset of all available attributes at each node when forming the tree. When combined with bagging, the resulting models are referred to as *random forests* [19], and these are widely considered to be among the most competitive and robust of current methods for predictive data mining [20].

C. Classifier Output Combination

Xu et al. [10] suggest that the output of individual classifiers can be divided into three different levels of information content which we refer to as *propositional*, *relational* and *confidence* in this discussion.[1] A propositional output merely states the classifier's preferred class and relational output involves an ordering or ranking of all classes from the most likely to the least likely. The propositional and relational outputs are qualitative values in contrast to the quantitative confidence output which assigns a numeric value to each class specifying the relative degree to which the classifier believes the class to represent the true class for the novel example. The confidence output is the most general since it can be transformed into a relational, which, in turn, can be transformed in a propositional output (i.e., the highest ranked class). On the confidence level, the output is often treated as a probability measure.

In the literature, different combination methods have been presented that apply to different output levels. For instance, the *weighted majority voting* method applies to propositional output and *borda count* to relational [4]. The preferred class c^* using the *weighted* majority voting method is

$$c^* = \arg\max_{c \in \mathcal{C}} \sum_{e \in E} r_e \, \delta_{e,c} \tag{1}$$

where r_e is a reliability weight for classifier e and

$$\delta_{e,c} = \begin{cases} 1, & \text{if } e \text{ outputs } c \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

Hence, the "combined vote" for a class c is the sum of the weights of the classifiers that have c as their output. The reliability is often measured as a classifier's rate of correctly classified training examples (i.e., its classification accuracy).

Since all outputs of the confidence level can be reduced to the levels of lower information content, combination methods applicable to the propositional and relational level are also applicable to the confidence level. Consequently, such methods can be applied to heterogeneous sets of classifiers by transforming the outputs of different levels to a common level.

III. Evidential Theory

The idea in evidential theory [15] is to build beliefs about the true state of a process from smaller and distinct pieces of *evidence*. The set of possible states is called the *frame of discernment* and is denoted by Θ. The frame of discernment is both *mutually exclusive* and *exhaustive*, i.e., only one state in Θ can be the true state and the true state is assumed to be in the set. Evidences are formulated as *mass functions*, $m : 2^\Theta \mapsto [0, 1]$, satisfying the following axioms:

$$m(A) \geq 0 \tag{3}$$

$$m(\emptyset) = 0 \tag{4}$$

$$\sum_{A \subseteq \Theta} m(A) = 1, \tag{5}$$

where $A \subseteq \Theta$. All subsets $A \subseteq \Theta$ for which $m(A) > 0$ are called *focal elements*. Once a mass function over the frame

[1] Although these levels are well known, the names we have chosen are unconventional. In the literature, various names are given to these levels. Propositional output is sometimes called *abstract* or *decision*, and the confidence output is sometimes called *soft*, *continuous*, *measurement* or *degree of support*.

of discernment has been obtained, the *belief* for a set $A \subseteq \Theta$ can be calculated in the following way:

$$Bel(A) = \sum_{B \subseteq A} m(B) \tag{6}$$

Another function frequently used is *plausibility* [15]:

$$Pl(A) = 1 - Bel(\bar{A}) = \sum_{B \cap A \neq \emptyset} m(B) \tag{7}$$

If mass functions are produced by sources that have different degrees of *reliability*, e.g., sensors of different quality, it is possible to account for this by utilizing reliability factors and *discount* the sources in the following way:

$$\begin{aligned} m_i^{\alpha}(A) &= \alpha\, m_i(A), \forall A \neq \Theta \\ m_i^{\alpha}(\Theta) &= \alpha\, m_i(\Theta) + (1-\alpha), \end{aligned} \tag{8}$$

where $0 \leq \alpha \leq 1$ is the reliability factor of source i.

When a number of different distinct pieces of evidence are available, these can be combined into a single mass function by applying a *combination rule*.

A. Evidential Combination Rules

Combination rules specify how two mass functions, say m_1 and m_2, are fused into one combined belief measure $m_{12} = m_1 \otimes m_2$ (we here let the binary operator $\otimes$ denote any rule for mass function combination). Many combination rules have been suggested (several are presented in [16]), and we below briefly discuss the ones we use in our study.

To combine multiple mass functions, the combination rule is applied repeatedly. Most combination rules are *associative*, i.e., $(m_1 \otimes m_2) \otimes m_3 = m_1 \otimes (m_2 \otimes m_3)$, meaning that the order in which mass functions are combined does not affect the final outcome. For non-associative rules, however, that do not satisfy this algebraic property, the order matters. Hence, unless a specific order of the classifier outputs can be justified, the result of using this type of rules is ambiguous. For this reason, and due to indications of poor performance for ensemble classification (indicated in a previous article [17]), we focus on some associative rules in this study.

Dempster's rule was the rule originally proposed:

$$m_{12}(X) = \frac{1}{1-K} \sum_{\substack{A,B \subseteq \Theta \\ A \cap B = X}} m_1(A)\, m_2(B), \tag{9}$$

$\forall X \subseteq \Theta, X \neq \emptyset$, where K is the *degree of conflict* between the two mass functions:

$$K = \sum_{\substack{A,B \subseteq \Theta \\ A \cap B = \emptyset}} m_1(A)\, m_2(B) \tag{10}$$

The Modified Dempster's rule (MDS) by Fixsen and Mahler [16], [21] is derived from random set theory. It is similar to Dempster's rule, but has an additional factor β:

$$m_{12}(X) = k \sum_{\substack{A,B \subseteq \Theta \\ A \cap B = X}} \beta\, m_1(A)\, m_2(B), \tag{11}$$

$\forall X \subseteq \Theta, X \neq \emptyset$, where k is a normalization constant and

$$\beta = \frac{q(X)}{q(A)\, q(B)} \tag{12}$$

$q(\cdot)$ is a (ordinary) Bayesian prior common to both classifiers.

B. Decision Making

Deciding on a most likely state, given a mass function, is non-trivial as the evidence of each state $\theta_i \in \Theta$ may be interpreted as a *belief interval* $[Bel(\theta_i), Pl(\theta_i)]$ (rather than an exact number) which might be overlapping the interval for another state θ_j $(j \neq i)$ and, hence, be incomparable. A mass function can, however, be "transformed" into a probability measure which can be used for comparison. One way to construct a probability measure from a mass function is the *pignistic transform* [22]:

$$BetP(\theta) = \sum_{B \subseteq \Theta} \frac{m(B)}{|B|}\, d(\theta, B), \tag{13}$$

where $d(\theta, B) = 1$ if $\theta \in B$ (zero otherwise), and $BetP(\cdot)$ is the resulting probability measure. From (13), the θ which maximizes $BetP$ can be selected as the most likely state.

IV. Evidence-Based Ensemble Classifiers

The construction of ensemble classifiers can generally be divided into two parts: generation of classifiers and combination method design [11, Sec. 2]. Much of the work on ensembles has focused on the first part, i.e., constructing the ensembles: considering what classifiers to select (decision trees, artificial neural networks, etc.), how many and how to train them. As mentioned, diversity among ensembles is a key issue, but how diversity is most appropriately measured and achieved is an ongoing research problem.

The second part is what we focus on in this article. For mass function combination, there are three issues to consider: 1) how to construct mass functions from the classifiers, 2) how to combine the mass functions, and 3) decide on an ensemble output. Let, for the following discussion, the frame of discernment be the set $\Theta_C = \{\theta_c | c \in C\}$, where C is a set of classes and θ_c represents the hypothesis that a novel example belongs to class c.

In the literature, there are basically two different proposals on how to construct mass functions. One is to construct mass functions from classifier output, as done in, e.g., [13, Sec. 4.3.2]. Another approach is to construct mass functions directly in the classifier [23].

For the combination of ensemble classifier beliefs, the most common combination rule is the original Dempster's rule, e.g., [10], [23]. Some approaches do have an extended combination scheme which inspects the mass functions before combination and to avoid combining conflicting masses [10].

The final issue to consider is that of ensemble output. One approach is to select the class c^* which maximizes $Bel(\theta_c)$ [9]. Another considers both ends of the belief interval [10, p. 428]. Yet another approach is to transform the mass function to a probability measure using the pignistic transform in (13) (that and other decision approaches for mass functions are presented in [10], [24]).

V. Empirical Evaluation

A. Experimental Setting

1) Ensemble Design: In Section IV, we describe different parts of the ensemble construction procedure. Below, we present the specific design details of the ensembles that we use in our experiments.

The ensemble classifiers are constructed using the random forest technique presented in Section II-B. For each ensemble, 25 trees are constructed. Each tree is generated from a bootstrap replicate of the training set [18], and at each node in the tree generation, only a random subset of the available attributes are considered for partitioning the examples, where the size of this subset is equal to the square root of the number of available attributes (as suggested in [19]). The entire set of training examples is used for determining which class is the most probable in each leaf. All compared ensembles are identical except for the combination rule and belief representation that is used when classifying novel instances.

In this study, we consider random forests for which each tree has propositional output (i.e., each tree provides only its best class for a novel example). From this output, a mass function m_e for each constituent classifier e with output class proposition θ_e is constructed in the following way:

$$\begin{array}{rcl} m_e(\{\theta_e\}) & = & 1 \\ m_e(A) & = & 0, \quad \forall A \subseteq \Theta, A \neq \{\theta_e\} \end{array} \tag{14}$$

To take into consideration that the different classifiers have different reliability in their outputs, we also discount the mass functions, using (8), with the reliability value r, i.e., creating the updated mass function m_e^r. We consider two different types of reliability measures for a classifier: 1) *average* and 2) *class-specific classification accuracy* (also called precision). The reliabilities are estimated by measuring the accuracy of each tree on training examples that are *out-of-the-bag*, i.e., which have not been used to generate the tree. For the former type $r = m/n$, where n is the number of out-of-the-bag examples and m is the number of examples correctly classified by classifier e. Since the accuracy of classifier may differ between classes, it would seem to be an improvement to specify the class-specific accuracy for each classifier and class. This is the latter type, which could assign reliability $r_c = m_c/n_c$ to classifier e if n_c is the number of examples that e classifies as belonging to class c and m_c is the number correctly classified. There is, however, a problem. For some datasets that we use, the number of examples for each class may be low, resulting in poor estimates of the class-specific accuracy. To become less sensitive to the number of examples we apply the Laplace correction resulting in

$$r_c = \frac{m_c + 1}{n_c + 2} \tag{15}$$

The Laplace correction has the appealing property that it assumes a uniform probability when there are no test examples for a specific class.

Table I

The fusion configurations are based on both combination rule and reliability measure

Rel./Comb.rule	**WV**	**DS**	**MDSu**	**MDS**
Average	*WV-a*	*DS-a*	*MDSu-a*	*MDS-a*
Class-specific	*WV-p*	*DS-p*	*MDSu-p*	*MDS-p*

The evidential combination rules (see Section III-A) that are to be compared for random forests are: Dempster (*DS*), and modified Dempster's rule. As shown in (11), the modified rule requires a specified common prior. Although all classifiers are based on the same training set, it is non-obvious how this fact can be translated into a common prior. For our study, we try two different priors: uniform (*MDSu*) and based on the relative frequencies of classes in the training set (*MDS*). As a comparison to the evidential-based combination rules, we use weighted voting (1) of the output of all trees in the forest where each tree's vote is weighted by the classifier's estimated reliability (*WV*).

Since all four combination rules can be used with both types of reliability, we end up comparing the eight configurations shown in Table I.

Finally, we use the pignistic transform (13) to generate the ensemble output.

2) Methodology and data sets: Accuracy (i.e., the percentage of correctly classified examples) is by far the most common criterion for evaluating classifiers, and this is the criterion chosen also for this study. It should, however, be noted that there are several other possible criteria for evaluating the predictive performance. There has recently been a growing interest in the ranking performance, which can be evaluated by measuring the area under the ROC curve [25] (AUC). The AUC can be interpreted as the probability of ranking a true positive example ahead of a false positive when ordering examples according to decreasing likelihood of being positive [26]. A third important property when evaluating classifiers that output class probabilities is the correctness of the probability estimates. This is of particular importance in situations where a decision is to be made that is based not on which class is the most likely for an example, or the relative likelihood of class membership compared to other examples, but on the likelihood of a particular class being the true class for the example. However, in this study, we will only consider the accuracy criterion.

The methods are compared w.r.t. accuracy using stratified ten-fold cross-validation on 27 data sets from the UCI Repository [27], where the average scores obtained for the ten folds are calculated. The names of the data sets together with the number of classes are listed in the first column of Table II.

3) Test hypotheses: The null hypotheses can be formulated as that there for each pair of fusion configurations (summarized in Table I) is no difference in predictive performance (i.e., as measured by accuracy when used in conjunction with the selected ensemble design). A null hypothesis for a pair of configurations is rejected if the probability of obtaining the observed number of wins and losses, given the null hypothesis, is less than 0.05. We refer to such observed differences in performance as being statistically significant.

B. Experimental Results

The accuracies obtained for all methods on the 27 data sets are shown in Table II. The number of wins and losses for each pair of method with respect to accuracy is shown in Table III, where results for which the p-value (double-sided binomial tail probability) is less than 0.05 are marked with bold-face. As can be seen, the results only show *MDS-p* to outperform *DS-a*. However, it should be noted that for all combination rules (except *DS*) the classification performance improves when exchanging the average accuracy reliability to the class-specific one.

C. Discussion

First of all, improving the ensemble classification accuracy, over the easily implemented and computed weighted voting (*WV*), by applying an evidential approach appears to be challenging. Among the evidential rules, however, the *MDS* (with class-specific belief, i.e., *MDS-p*) appears advantageous, not losing any of the pairwise competitions.

One of the motivations for this study is, furthermore, the class-specific belief representation (explained in Section V) which allows classifiers to have different reliability values depending on the class which they output. The accuracy performance of all classifiers improved slightly for the class-specific belief representation, except for the *MDS* which experienced a drastic improvement, and *DS* which declined.

It should be stressed that the above findings concern the general tendency on all datasets, and for a particular dataset other combination rules may be advantageous. Note, e.g., the image-segmentation dataset in Table II for which *DS-a*, having relatively poor performance in general in the study, excels.

VI. Concluding Remarks

Our study shows that by using a class-specific reliability measure, instead of one based on overall accuracy, the predictive performance of applying combination rules in an evidential framework may be substantially improved. In contrast to the results reported in [17], our results reported here indicate that evidential combination rules may in fact perform better than the straightforward weighted voting approach, given that suitable reliability measures are provided. Advantages with the evidential framework for other ensemble designs and datasets have been reported [8], [10].

One direction for future research is to consider random forests with confidence output (e.g., a probability measure over classes) as discussed in [28]. One specific issue that needs to be considered in such a study is whether or not accuracy is a suitable criterion to use for discounting in that case, since the accuracy does not reflect the correctness of the probability estimates, in contrast to, e.g., the Brier score [29].

The design of the mass function is of course also important. In this study, we construct mass functions from propositional output and meta-information (i.e., reliability values). A natural approach to exploit the potential of the mass function is to construct mass functions directly in the classifiers as in, e.g., [23]. Another fascinating approach (described in [13, Sec. 4.5]) is to build meta-combiners which combine the outputs from several different combination rules.

An additional direction for future research is to empirically compare the combination rules for other types of classifiers. In principle, a similar experiment as the one presented in this study could instead have considered ensembles in which each classifier is built from data from a specific sensor that captures the specific sensor properties and environmental context. Note that classifier diversity, which is necessary for effective ensembles, is inherent when each classifier stems from a different sensor.

Acknowledgment

We wish to thank Prof. Sten F. Andler (University of Skövde) for his constructive comments and support. This work was supported by the Information Fusion Research Program (www.infofusion.se) at the University of Skövde, Sweden, in partnership with the Swedish Knowledge Foundation under grant 2003/0104, and participating partner companies.

References

[1] E. L. Waltz, "Information understanding: Integrating data fusion and data mining processes," in *Proceedings of the IEEE international symposium on circuits and systems*, 1997.

[2] V. Torra, Ed., *Information Fusion in Data Mining*, ser. Studies in Fuzziness and Soft Computing. Springer Verlag, 2003, ISBN-10:3540006761 and ISBN-13:978-3540006763.

[3] D. Wolpert and W. Macready, "No free lunch theorems for optimization," *IEEE Transactions on Evolutionary Computation,*, vol. 1, no. 1, pp. 67–82, 1997.

[4] R. Polikar, "Ensemble based systems in decision making," *IEEE Circuits and systems magazine*, vol. 6, no. 3, pp. 21–45, 2006.

[5] L. K. Hansen and P. Salamon, "Neural network ensembles," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 12, no. 10, pp. 993–1001, October 1990.

[6] E. Bauer and R. Kohavi, "An empirical comparison of voting classification algorithms: Bagging, boosting, and variants," *Machine Learning*, vol. 36, no. 1-2, pp. 105–139, 1999. [Online]. Available: http://citeseer.ist.psu.edu/bauer99empirical.html

[7] L. I. Kuncheva, "A theoretical study on six classifier fusion strategies," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 2, pp. 281–286, February 2002.

[8] D. Bahler and L. Navarro, "Methods for combining heterogeneous sets of classifiers," in *17th National Conference on Artificial Intelligence, Workshop on New Research Problems for Machine Learning*, 2000.

[9] D. Ruta and B. Gabrys, "An overview of classifier fusion methods," *Computing and Information Systems*, vol. 7, pp. 1–10, 2000.

[10] L. Xu, A. Krzyzak, and C. Y. Suen, "Methods of combining multiple classifiers and their applications to handwriting recognition," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 22, no. 3, pp. 418–435, May/June 1992.

[11] M. Reformat and R. R. Yager, "Building ensemble classifiers using belief functions and OWA operators," *Soft Computing - A Fusion of Foundations, Methodologies and Applications*, vol. 12, no. 6, pp. 543–558, April 2008.

[12] A. Al-Ani and M. Deriche, "A new technique for combining multiple classifiers using the Dempster-Shafer theory of evidence," *Journal of Artificial Intelligence Research*, vol. 17, pp. 333–361, 2002.

Table II
ACCURACY FOR THE EIGHT COMBINATION RULES

Data set	**WV-a**	**WV-p**	**DS-a**	**DS-p**	**MDSu-a**	**MDSu-p**	**MDS-a**	**MDS-p**
balance-scale (3 cl.)	85.45	87.52	85.45	87.52	85.06	87.38	82.92	87.22
breast-cancer (2 cl.)	72.73	73.08	72.81	73.79	72.39	73.08	72.41	72.73
breast-cancer-wisconsin (2 cl.)	95.85	95.85	95.85	95.39	95.39	95.35	95.71	95.85
car (4 cl.)	96.18	96.41	96.18	95.72	96.18	96.12	96.40	96.82
cleveland-heart-disease (5 cl.)	55.42	55.76	55.42	55.11	55.42	55.76	55.11	56.73
crx (2 cl.)	86.37	86.37	86.22	87.09	86.78	86.94	85.79	86.22
cylinder-bands (2 cl.)	79.26	78.70	79.63	79.26	80.45	79.26	80.00	80.25
dermatology (6 cl.)	97.80	97.80	97.86	96.47	97.56	97.59	98.36	98.08
ecoli (8 cl.)	87.18	85.99	87.78	85.39	87.18	86.29	86.00	87.18
glass (6 cl.)	77.86	77.86	77.86	75.06	77.86	77.84	76.95	75.50
hepatitis (2 cl.)	86.42	85.75	85.79	82.54	85.79	82.54	84.46	85.79
house-votes (2 cl.)	96.31	96.31	96.31	96.41	96.31	96.31	96.32	96.31
image-segmentation (7 cl.)	92.86	92.38	94.18	91.90	92.86	91.90	92.06	91.90
ionosphere (2 cl.)	93.75	93.45	93.75	94.02	93.75	94.02	93.75	93.75
iris (3 cl.)	94.67	94.67	95.33	94.67	95.33	94.67	95.33	94.67
kr-vs-kp (2 cl.)	98.62	98.62	98.62	98.65	98.68	98.65	98.61	98.65
lung-cancer (3 cl.)	46.67	50.00	50.00	50.00	50.00	50.00	50.00	50.00
lymphography (4 cl.)	85.14	85.14	85.14	84.34	85.14	85.14	85.81	85.14
new-thyroid (3 cl.)	95.37	94.91	94.91	94.44	94.91	94.91	94.91	95.37
pima-indians-diabetes (2 cl.)	76.68	75.90	76.55	74.47	76.68	74.73	75.78	76.68
post-operative-patients (3 cl.)	68.89	70.00	68.89	71.11	68.89	71.11	68.89	68.89
promoters (2 cl.)	80.18	81.09	80.27	83.00	80.18	82.09	80.18	82.09
spectf (2 cl.)	90.27	91.40	90.27	90.83	90.27	91.39	89.69	90.83
tae (3 cl.)	54.25	55.58	54.25	54.92	53.33	54.07	54.25	55.58
tic-tac-toe (2 cl.)	97.18	96.76	97.08	96.45	96.99	96.45	96.18	97.34
wine (3 cl.)	97.71	98.30	97.16	98.30	97.71	98.30	98.61	97.46
yeast (10 cl.)	60.98	61.26	60.78	61.59	60.98	61.19	60.99	61.39

Table III
PAIRWISE ACCURACY COMPARISON (ROW WINS/COLUMN WINS)

	wv-a	wv-p	DS-a	DS-p	MDSu-a	MDSu-p	MDS-a	MDS-p
wv-a	-	8/11	7/8	12/13	8/5	11/12	14/10	5/12
wv-p	11/8	-	11/9	14/9	12/10	13/7	18/8	9/13
DS-a	8/7	9/11	-	14/12	9/6	12/11	15/8	**5/15**
DS-p	13/12	9/14	12/14	-	11/14	8/9	12/14	10/13
MDSu-a	5/8	10/12	6/9	14/11	-	11/12	13/10	7/12
MDSu-p	12/11	7/13	11/12	9/8	12/11	-	14/12	9/12
MDS-a	10/14	8/18	8/15	14/12	10/13	12/14	-	8/17
MDS-p	12/5	13/9	**15/5**	13/10	12/7	12/9	17/8	-

[13] L. A. Cuong, "A study of classifier combination and semi-supervised learning for word sense disambiguation," Ph.D. dissertation, Japan Advanced Institute of Science and Technology, March 2007.

[14] G. Rogova, "Combining the results of several neural network classifiers," *Neural Networks*, vol. 7, no. 5, pp. 777–781, 1994.

[15] G. Shafer, *A Mathematical Theory of Evidence*. Princeton, NJ, USA: Princeton University Press, 1976.

[16] P. Smets, "Analyzing the combination of conflicting belief functions," *Information Fusion*, vol. 8, pp. 387–412, 2007.

[17] H. Boström, R. Johansson, and A. Karlsson, "On evidential combination rules for ensemble classifiers," in *Proceedings of the 11th International Conference on Information Fusion*, 2008.

[18] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, no. 2, pp. 123–140, 1996. [Online]. Available: http://citeseer.ist.psu.edu/breiman96bagging.html

[19] ——, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001. [Online]. Available: http://citeseer.ist.psu.edu/breiman01random.html

[20] R. Caruana and A. Niculescu-Mizil, "An empirical comparison of supervised learning algorithms," in *Proc. of the 23rd International Conference on Machine Learning*, 2006, pp. 161–168.

[21] D. Fixsen and R. P. S. Mahler, "The modified Dempster-Shafer approach to classification," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 27, no. 1, pp. 96–104, January 1997.

[22] P. Smets and R. Kennes, "The transferable belief model," *Artificial Intelligence*, vol. 66, no. 2, pp. 191–234, April 1994.

[23] G. Rogova, P. Scott, and C. Lolett, "Distributed reinforcement learning for sequential decision making," in *Proceedings of the 5th International Conference on Information Fusion*. International Society of Information Fusion, 2002, pp. 1263–1268.

[24] H. Altinçay and M. Demirekler, "Speaker identification by combining multiple classifiers using Dempster-Shafer theory of evidence," *Speech Communication*, vol. 41, pp. 531–547, 2003.

[25] T. Fawcett, "ROC graphs: Notes and practical considerations for data mining researchers," HP Laboratories, Palo Alto, Tech. Rep., 2003.

[26] A. Bradley, "The use of the area under the ROC curve in the evaluation of machine learning algorithms," *Pattern Recognition*, vol. 30, no. 6, pp. 1145–1159, 1997.

[27] C. Blake and C. Merz, "UCI repository of machine learning databases," 1998. [Online]. Available: http://www.ics.uci.edu/~mlearn/MLRepository.html

[28] H. Boström, "Estimating class probabilities in random forests," in *Proc. of the International Conference on Machine Learning and Applications*, 2007, pp. 211–216.

[29] G. W. Brier, "Verification of forecasts expressed in terms of probability," *Monthly Weather Review*, vol. 78, pp. 1–3, 1950.