On Evidential Combination Rules for Ensemble Classifiers

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

Ronnie Johansson
School of Humanities and Informatics
University of Skövde
Skövde, Sweden
Email: ronnie.johansson@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 each individual classifier of which they consist. One approach to classifier fusion is to apply Shafer's theory of evidence. While most approaches have adopted Dempster's rule of combination, a multitude of combination rules have been proposed. A number of combination rules as well as two voting rules are compared when used in conjunction with a specific kind of ensemble classifier, known as random forests, w.r.t. accuracy, area under ROC curve and Brier score on 27 datasets. The empirical evaluation shows that the choice of combination rule can have a significant impact on the performance for a single dataset, but in general the evidential combination rules do not perform better than the voting rules for this particular ensemble design. Furthermore, among the evidential rules, the associative ones appear to have better performance than the non-associative.**

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., have 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 training examples given to the learning algorithm or randomizing the process for generating each classifier [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 [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]. 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].

To the best of our knowledge, there has been no previous study that compares various combination rules on large numbers of datasets for any type of ensemble classifier. In this work, we provide some light on this problem by investigating the use of eight different combination rules on 27 datasets from the UCI repository [17], for a specific type of ensemble classifier, *random forests* [18], which is widely considered to be among the most powerful predictive methods, see e.g. [19].

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 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 $\mathbf{x}$ (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. Furthermore, the *reliability* of a classifier e is denoted r_e and is in this study an estimate of the classification *accuracy* (or recognition rate), i.e., the percentage of examples that are correctly classified.

B. Random Forests

The basic strategy that is employed when generating classification trees from training examples is called recursive partitioning, or divide-and-conquer. It works by partitioning the examples by choosing a set of mutually exclusive conditions on an independent variable, or attribute, e.g., the variable has a value less than a particular threshold, or a value greater or equal to this threshold, and the choice is usually made such that the error on the dependent variable (or class variable) is minimized within each group. The process continues recursively with each subgroup until certain conditions are met, such as that the error cannot be further reduced (e.g., all examples in a group have the same the same class label). The resulting classification tree is a graph that contains one node for each subgroup considered, where the node corresponding to the initial set of examples is called the root, and for all nodes there is an edge to each subgroup generated from it, labeled with the chosen condition for that subgroup. An example is classified by the tree by following a path from the root to a leaf node, such that all conditions along the path are fulfilled by the example. The estimated class probabilities at the reached leaf node are used to assign the most probable class to the example.

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 vote 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 [20]. 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 constructing the tree. When combined with bagging, the resulting models are referred to as random forests [18], and these are widely considered to be among the most competitive and robust of current methods for predictive data mining [19].

C. Classificer Output Combination

Xu et al. [10] suggest that the output of individual classifiers can be divided into three different levels of information content: *propositional*, *relational* and *confidence*.[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 degree to which the classifier believes the class to represent the true class for the novel example. The confidence output is 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 *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.

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

In 1976, Glenn Shafer published his seminal book entitled *"A Mathematical Theory of Evidence"* [15], often referred to as *Evidential theory* or *Dempster-Shafer theory*. The idea in evidential theory 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*, $m : 2^\Theta \mapsto \mathbb{R}$,

[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*.

satisfying the three 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 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 weights and *discount* the sources in the following way:

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

where $0 \leq \alpha \leq 1$ is the reliability weight 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 below we 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 have 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. In spite of this, in our experiments in Section V, we use some non-associative rules for comparison, but with arbitrary ordering of the mass functions to combine.

1) Associative Rules: Dempster's rule was the rule originally proposed [15]:

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

$\forall X \subseteq \Theta, X \neq \emptyset$, 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.

The disjunctive rule,

$$m_{12}(X) = \sum_{\substack{A,B \subseteq \Theta \\ A \cup B = X}} m_1(A)\, m_2(B), \tag{13}$$

$\forall X \subseteq \Theta$, has been suggested to be used when one of the sources (which one not known) is expected to be incorrect [16, pp. 391].

2) Non-Associative Rules: The two non-associative rules we use in our comparison are the Yager and Dubois-Prade rules [16]. Yager's rule assigns conflicting mass to the frame of discernment Θ (instead of normalizing as in Dempster's rule in Eq. 9):

$$m_{12}(X) = \begin{cases} \displaystyle\sum_{\substack{A,B \subseteq \Theta \\ A \cap B = X}} m_1(A)\, m_2(B), & \forall X \subset \Theta, X \neq \emptyset \\ m_1(\Theta)\, m_2(\Theta) + K, & \text{if } X = \Theta \end{cases} \tag{14}$$

where K is the same conflict as in Eq. 10, and $m_{12}(\emptyset) = 0$.

The Dubois-Prade rule, instead, assigns the conflicting mass to the union of the non-intersecting focal elements:

$$\begin{aligned} m_{12}(X) \;=\; & \sum_{\substack{A,B \subseteq \Theta \\ A \cap B = X}} m_1(A)\, m_2(B) + \\ & \sum_{\substack{A,B \subseteq \Theta \\ A \cap B = \emptyset \\ A \cup B = X}} m_1(A)\, m_2(B) \end{aligned} \tag{15}$$

$\forall X \subset \Theta$ and $X \neq \emptyset$, and $m_{12}(\emptyset) = 0$.

B. Decision Making

Deciding on a most likely state, given a mass function, is non-trivial as each state $\theta_i \in \Theta$ may be interpreted as an 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, incomparable. A mass function can, however, be "transformed" into a probability measure which can be used for comparison. A common way to construct a probability measure from a mass function is the *pignistic transform* [22]:

$$BetP(\theta) = \sum_{B \subseteq \Theta} \frac{m(B)}{|B|} d(\theta, B), \qquad (16)$$

where $d(\theta, B) = 1$ if $\theta \in B$ (zero otherwise), and $BetP(\cdot)$ is the resulting probability measure. From Eq. 16, the θ which maximizes $BetP$ can be selected as the most likely state.

IV. Evidence-Based Ensemble Classifiers

Constructing 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. In Section II-C, we presented three different levels of output. Note that the type of information represented by a mass function is of the most general level, i.e., confidence. Also, existing classifiers with confidence output frequently output a probability measure. Hence, the mass function is typically more expressive than most classifier outputs, and to utilize this extended expressiveness, meta-information about the classification is often incorporated into the mass functions. One simple way of utilizing this expressiveness is to discount (Eq. 8) the mass function with some reliability measure [13, Sec. 4.3.2]. A similar approach is to assign the reliability or recognition rate ϵ_r to the propositional output class $c \in C$, e.g., $m(\theta_c) = \epsilon_r$ and its misclassification rate ϵ_s to the complement of θ_c, i.e., $m(\neg\theta_c) = \epsilon_s$ [9], [10], where $\neg\theta_c = \Theta_C \setminus \{\theta_c\}$. Another approach [14] uses, instead of recognition rate as reliability, the difference between the confidence output for a novel example $\mathbf{x}$ and a reference output (learned from some training examples). A *proximity measure* is then used to decide the reliability of the classifier output and this is reflected in the resulting mass function.

Another approach is to construct mass functions directly in the classifier. In [23], a similar approach to [14] is adopted, but instead of utilizing a confidence output from each classifier, the mass functions are constructed directly from the comparison of an example and reference examples. The reference examples represent typical attribute values for members of the corresponding class. The mass function is then constructed by assigning mass according to a proximity measure between the novel example and the references.

For the combination of ensemble classifier mass functions, the most common combination rule in 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. Although the mass function is a confidence measure, it represents confidence intervals (where its endpoints are given by Eq. 6 and 7) rather than confidence points (as in the case of probability measures). One approach is to select the class c^* which maximizes $Bel(\theta_c)$ [9]. Another considers both ends of the confidence interval [10]. Yet another approach is to transform the mass function to a probability measure using the pignistic transform in Eq. 16 (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 [20], 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 [18]). 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 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:

$$\begin{array}{rcl} m_e(\{\theta_e\}) & = & 1 \\ m_e(A) & = & 0, \quad \forall A \subseteq \Theta, A \neq \{\theta_e\} \end{array} \qquad (17)$$

To take into consideration that the different trees have different reliability in their outputs, we also discount the mass functions (using Eq. 8) with the reliability value r, i.e., creating the updated mass function m_e^r. The reliability is 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.

The evidential combination rules (see Section III-A) that are to be compared for random forests are: Dempster (*DS*), modified Dempster, the disjunctive rule (*Disjunction*), *Yager*, and *Dubois-Prade*. The modified Dempster's rule requires a specified common prior, and although all classifiers are based on the same (or similar) dataset, it is difficult to specify a common prior. For our study, we try two different priors: uniform (*MDS-u*) and based on relative frequencies of classes in the training set (*MDS*).

As a comparison to the evidential-based combination rules, we use unweighted voting of the output of all trees in the forest (*voting*) and voting where each tree's vote is weighted by the classifier's reliability (*w. voting*).

Finally, we use the pignistic transform (Eq. 16) 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. There has, however, recently been a growing interest in also the ranking performance, which can be evaluated by measuring the area under the ROC curve (AUC) [25]. 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. This is required, e.g., when calculating the expected utility of different alternatives that depend on the class membership of the example. Different measures for the correctness of the class probabilities have been proposed, but the mean squared error of the predicted class probabilities, referred to as the *Brier score* [27], is one of the most commonly employed.

The methods are compared w.r.t. accuracy, AUC and Brier score using stratified ten-fold cross-validation on 27 data sets from the UCI Repository [17], where the average scores obtained for the ten folds are calculated.[2] The names of the data sets together with the number of classes are listed in the first column of Table I.

3) Test hypotheses: There are actually a number of hypotheses to be tested. The null hypotheses can be formulated as that there for each pair of combination rules is no difference in predictive performance (i.e., as measured by accuracy, AUC and Brier score) when used in conjunction with the selected ensemble design.

B. Experimental Results

1) Accuracy: The accuracies obtained for all methods on the 27 data sets are shown in Table I. The number of wins and losses for each pair of methods with respect to accuracy is shown in Table II, where results for which the p-value (double-sided binomial tail probability) is less than 0.05 are marked with bold-face. It can be seen that the three best performing methods w.r.t. accuracy are weighted voting, Dempster and modified Dempster with uniform prior (*MDS-u*), which all perform on the same level (i.e., about equal number of wins and losses when compared pairwise). These are slightly ahead of unweighted voting and modified Dempster with prior based on relative frequencies (*MDS*), although the number of wins and losses do not deviate significantly from what can be expected if the null hypotheses were true. Far behind come (in decreasing order of performance) *Dubois-Prade*, *Yager* and *Disjunction*, all significantly outperformed by the other combination rules.

2) The area under the ROC curve: The AUC values are shown in Table III and the number of wins and losses for each pair of methods with respect to AUC is shown in Table IV. In contrast to when comparing the methods w.r.t. accuracy, there is a single clear winner when comparing them w.r.t. AUC: weighted voting significantly outperforms all other combination rules. There is also a clear second-best method: voting, which again significantly outperforms the inferior methods. Of the remaining methods, *DS* significantly outperforms *MDS-u* and *Disjunction*, and is clearly ahead of *MDS* and *Yager*. Interestingly, in contrast to the results for accuracy, *MDS* is clearly ahead of *MDS-u* with respect to AUC.

3) Brier score: The Brier scores are shown in Table V, and the number of wins and losses for each pair of methods regarding Brier score is shown in Table VI. The two best methods are weighted voting and voting, significantly ahead of the other methods. Again, DS comes in third place.

C. Discussion

One conclusion from the empirical investigation is that the choice of combination rule indeed can have a significant impact on the predictive performance of random forests, independently of whether the performance is measured by accuracy, AUC or Brier score.

Furthermore, the results show that if accuracy is to be optimized for this particular ensemble design, three of the combination rules are equally good candidates: weighted voting, Dempster and modified Dempster with uniform prior. If AUC is to be optimized, weighted voting is the rule of choice, and finally, if Brier score is to be optimized, voting and weighted voting are the best candidates. If several of these criteria are to be optimized, weighted voting is the overall best choice of combination rule. However, it should be noted that these rankings are based on the general tendency on all datasets, and for a particular dataset and performance criterion there are obviously exceptions to these general rules of what combination rule to choose.

The experiment also clearly demonstrates the benefit of having estimates of each classifier's performance (i.e., accuracy on out-of-the-bag examples in our case) in addition to the vote. However, the experiment also demonstrates that for our ensemble design, it is hard for the evidence-based methods to utilize this information more effectively than what is done by the straightforward weighted voting method. Hence, the general conclusion from this study is that for the selected design of ensembles and mass functions, the expressiveness

[2] The AUC was calculated according to [25], and for data sets with more than two classes, the total (weighted) AUC is reported.

Table I
ACCURACY FOR THE EIGHT COMBINATION RULES

Data set	**voting**	**w. voting**	**DS**	**MDS-u**	**MDS**	**Disj.**	**Yager**	**Dubois-Prade**
balance-scale (3 cl.)	85.13	85.45	85.45	85.45	83.04	55.87	83.60	83.84
breast-cancer (2 cl.)	72.73	72.73	72.77	71.95	72.41	70.64	72.04	72.04
breast-cancer-wisconsin (2 cl.)	95.85	95.85	96.02	96.34	95.55	81.69	95.54	95.57
car (4 cl.)	96.12	96.18	96.18	96.01	96.30	70.72	94.21	94.73
cleveland-heart-disease (5 cl.)	55.76	55.42	55.42	55.42	55.11	54.12	56.43	57.11
crx (2 cl.)	86.37	86.37	86.22	86.22	85.79	59.79	85.98	85.65
cylinder-bands (2 cl.)	79.26	79.26	79.63	79.63	80.00	45.37	75.56	75.56
dermatology (6 cl.)	97.80	97.80	98.08	97.80	98.36	66.07	96.98	96.65
ecoli (8 cl.)	87.18	87.18	87.18	87.18	86.00	67.56	84.22	86.30
glass (6 cl.)	76.45	77.86	77.86	77.86	76.95	40.22	71.41	72.29
hepatitis (2 cl.)	86.42	86.42	85.79	85.79	84.46	53.46	85.17	85.17
house-votes (2 cl.)	96.31	96.31	96.31	96.31	96.32	88.24	96.09	96.09
image-segmentation (7 cl.)	91.90	92.86	92.86	92.86	92.06	58.57	86.19	88.57
ionosphere (2 cl.)	93.75	93.75	93.25	93.75	93.75	65.54	92.31	92.31
iris (3 cl.)	94.67	94.67	95.33	95.33	95.33	90.67	96.67	96.67
kr-vs-kp (2 cl.)	98.62	98.62	98.44	98.65	98.54	70.79	97.43	97.36
lung-cancer (3 cl.)	41.67	46.67	50.00	50.00	50.00	28.33	50.00	41.67
lymphography (4 cl.)	85.14	85.14	85.14	85.14	85.81	56.90	80.43	80.48
new-thyroid (3 cl.)	95.37	95.37	94.91	94.91	94.91	83.70	94.42	94.42
pima-indians-diabetes (2 cl.)	76.68	76.68	76.40	76.68	75.78	65.10	72.91	71.92
post-operative-patients (3 cl.)	70.00	68.89	68.89	69.14	69.14	71.11	68.89	68.89
promoters (2 cl.)	80.18	80.18	80.27	80.18	80.18	50.57	68.18	68.18
spectf (2 cl.)	90.27	90.27	90.27	90.27	89.69	42.14	87.39	86.95
tae (3 cl.)	54.25	54.25	54.25	54.25	54.25	32.50	49.67	49.67
tic-tac-toe (2 cl.)	97.18	97.18	97.08	97.18	96.35	48.09	90.81	90.81
wine (3 cl.)	97.71	97.71	97.16	97.71	97.71	64.02	95.49	96.22
yeast (10 cl.)	60.92	60.98	60.98	60.98	60.99	38.07	57.41	59.23

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

	voting	w. voting	DS	MDS-u	MDS	Disj.	Yager	Dubois-Prade
voting	-	2/6	10/12	7/9	13/11	**26/1**	**24/3**	**24/3**
w. voting	6/2	-	8/7	5/6	14/10	**26/1**	**23/4**	**24/3**
DS	12/10	7/8	-	4/7	14/10	**26/1**	**23/3**	**24/3**
MDS-u	9/7	6/5	7/4	-	13/8	**26/1**	**23/3**	**24/3**
MDS	11/13	10/14	10/14	8/13	-	**26/1**	**21/5**	**21/6**
Disj.	**1/26**	**1/26**	**1/26**	**1/26**	**1/26**	-	**1/26**	**1/26**
Yager	**3/24**	**4/23**	**3/23**	**3/23**	**5/21**	**26/1**	-	6/11
Dubois-Prade	**3/24**	**3/24**	**3/24**	**3/24**	**6/21**	**26/1**	11/6	-

Table III
AUC FOR THE EIGHT COMBINATION RULES

Data set	**voting**	**w. voting**	**DS**	**MDS-u**	**MDS**	**Disj.**	**Yager**	**Dubois-Prade**
balance-scale (3 cl.)	94.49	94.55	94.25	92.78	92.85	83.69	82.88	93.40
breast-cancer (2 cl.)	68.18	68.32	65.10	49.03	65.24	58.89	65.89	65.89
breast-cancer-wisconsin (2 cl.)	98.72	98.73	87.35	87.13	86.63	96.03	86.22	98.34
car (4 cl.)	99.77	99.77	99.15	97.75	98.46	91.55	96.54	99.43
cleveland-heart-disease (5 cl.)	80.32	80.51	80.35	77.14	77.19	72.56	78.38	78.31
crx (2 cl.)	92.86	92.95	91.73	91.20	90.83	73.49	81.05	91.72
cylinder-bands (2 cl.)	87.87	87.76	87.38	87.06	87.12	55.42	81.24	81.24
dermatology (6 cl.)	99.91	99.91	99.65	99.47	99.47	95.67	99.54	92.04
ecoli (8 cl.)	96.28	96.39	94.80	92.87	93.46	92.72	94.45	95.53
glass (6 cl.)	91.48	91.51	88.98	87.18	89.23	76.37	87.33	89.94
hepatitis (2 cl.)	87.81	87.66	81.60	77.63	79.95	68.14	80.75	80.75
house-votes (2 cl.)	98.95	98.95	97.31	97.33	96.97	96.06	98.16	98.16
image-segmentation (7 cl.)	98.72	98.72	98.36	96.46	90.86	92.92	97.00	98.10
ionosphere (2 cl.)	98.21	98.30	78.19	95.11	95.16	80.20	96.56	96.56
iris (3 cl.)	99.10	99.10	97.80	96.93	96.93	98.70	98.47	98.47
kr-vs-kp (2 cl.)	99.87	99.88	86.22	99.38	90.87	74.78	99.53	95.69
lung-cancer (3 cl.)	71.25	75.42	71.25	69.58	72.08	53.33	65.83	63.33
lymphography (4 cl.)	92.39	92.59	89.40	86.23	89.38	70.43	88.73	89.43
new-thyroid (3 cl.)	99.55	99.51	97.40	96.08	96.68	93.66	98.07	98.33
pima-indians-diabetes (2 cl.)	82.29	82.32	71.70	79.65	80.45	60.71	78.42	70.20
post-operative-patients (3 cl.)	51.94	49.88	48.38	45.90	47.34	51.27	52.23	52.19
promoters (2 cl.)	89.17	89.93	90.60	90.57	90.40	45.00	77.13	77.13
spectf (2 cl.)	95.70	95.82	94.95	94.75	95.29	60.23	92.29	81.27
tae (3 cl.)	72.29	72.01	71.44	71.73	71.67	59.44	67.15	69.72
tic-tac-toe (2 cl.)	99.35	99.36	98.87	98.93	98.85	61.37	96.16	96.16
wine (3 cl.)	99.87	99.90	99.15	98.77	98.77	89.61	99.53	90.20
yeast (10 cl.)	83.13	83.08	81.83	79.26	80.60	73.43	79.12	81.74

Table IV
PAIRWISE AUC COMPARISON (ROW WINS/COLUMN WINS)

	voting	w. voting	DS	MDS-u	MDS	Disj.	Yager	Dubois-Prade
voting	-	**6/17**	**25/2**	**26/1**	**25/2**	**27/0**	**26/1**	**26/1**
w. voting	**17/6**	-	**26/1**	**26/1**	**26/1**	**26/1**	**26/1**	**26/1**
DS	**2/25**	**1/26**	-	**21/6**	19/8	**23/4**	18/9	15/12
MDS-u	**1/26**	**1/26**	**6/21**	-	8/16	**23/4**	12/15	10/17
MDS	**2/25**	**1/26**	8/19	16/8	-	**23/4**	14/13	9/18
Disj.	**0/27**	**1/26**	**4/23**	**4/23**	**4/23**	-	**3/24**	**2/25**
Yager	**1/26**	**1/26**	9/18	15/12	13/14	**24/3**	-	8/11
Dubois-Prade	**1/26**	**1/26**	12/15	17/10	18/9	**25/2**	11/8	-

Table V
BRIER SCORE FOR THE EIGHT COMBINATION RULES

Data set	**voting**	**w. voting**	**DS**	**MDS-u**	**MDS**	**Disj.**	**Yager**	**Dubois-Prade**
balance-scale (3 cl.)	0.218	0.218	0.277	0.284	0.333	0.661	0.280	0.274
breast-cancer (2 cl.)	0.416	0.416	0.508	0.509	0.543	0.500	0.476	0.476
breast-cancer-wisconsin (2 cl.)	0.063	0.063	0.076	0.077	0.085	0.315	0.088	0.094
car (4 cl.)	0.068	0.068	0.074	0.075	0.072	0.621	0.121	0.103
cleveland-heart-disease (5 cl.)	0.537	0.537	0.815	0.852	0.852	0.800	0.600	0.638
crx (2 cl.)	0.195	0.195	0.267	0.270	0.271	0.496	0.235	0.243
cylinder-bands (2 cl.)	0.288	0.288	0.377	0.385	0.382	0.500	0.371	0.371
dermatology (6 cl.)	0.049	0.049	0.038	0.041	0.033	0.715	0.096	0.071
ecoli (8 cl.)	0.218	0.218	0.248	0.250	0.261	0.865	0.288	0.264
glass (6 cl.)	0.361	0.362	0.431	0.438	0.444	0.833	0.436	0.426
hepatitis (2 cl.)	0.209	0.210	0.272	0.272	0.300	0.497	0.299	0.299
house-votes (2 cl.)	0.061	0.061	0.072	0.072	0.073	0.346	0.097	0.097
image-segmentation (7 cl.)	0.142	0.142	0.141	0.139	0.144	0.843	0.256	0.216
ionosphere (2 cl.)	0.098	0.097	0.128	0.120	0.121	0.481	0.129	0.129
iris (3 cl.)	0.069	0.069	0.093	0.093	0.093	0.415	0.068	0.059
kr-vs-kp (2 cl.)	0.034	0.034	0.027	0.027	0.028	0.413	0.056	0.055
lung-cancer (3 cl.)	0.625	0.624	0.946	0.996	0.990	0.667	0.672	0.705
lymphography (4 cl.)	0.244	0.243	0.297	0.297	0.276	0.747	0.349	0.345
new-thyroid (3 cl.)	0.071	0.071	0.102	0.102	0.096	0.534	0.089	0.082
pima-indians-diabetes (2 cl.)	0.331	0.331	0.438	0.457	0.462	0.500	0.435	0.448
post-operative-patients (3 cl.)	0.530	0.531	0.610	0.606	0.623	0.667	0.573	0.571
promoters (2 cl.)	0.306	0.302	0.354	0.368	0.368	0.500	0.446	0.446
spectf (2 cl.)	0.148	0.148	0.181	0.185	0.201	0.499	0.213	0.218
tae (3 cl.)	0.637	0.636	0.862	0.888	0.890	0.667	0.722	0.749
tic-tac-toe (2 cl.)	0.096	0.095	0.054	0.054	0.067	0.495	0.168	0.168
wine (3 cl.)	0.059	0.059	0.048	0.045	0.045	0.614	0.094	0.076
yeast (10 cl.)	0.541	0.542	0.731	0.764	0.757	0.900	0.641	0.634

Table VI
PAIRWISE BRIER SCORE COMPARISON (ROW WINS/COLUMN WINS)

	voting	w. voting	DS	MDS-u	MDS	Disj.	Yager	Dubois-Prade
voting	-	12/15	**22/5**	**22/5**	**23/4**	**27/0**	**26/1**	**26/1**
w. voting	15/12	-	**22/5**	**22/5**	**23/4**	**27/0**	**26/1**	**26/1**
DS	**5/22**	**5/22**	-	**21/6**	**21/6**	**23/4**	16/11	15/12
MDS-u	**5/22**	**5/22**	**6/21**	-	17/9	**23/4**	14/13	14/13
MDS	**4/23**	**4/23**	**6/21**	9/17	-	**23/4**	13/14	13/14
Disj.	**0/27**	**0/27**	**4/23**	**4/23**	**4/23**	-	**2/25**	**2/25**
Yager	**1/26**	**1/26**	11/16	13/14	14/13	**25/2**	-	7/13
Dubois-Prade	**1/26**	**1/26**	12/15	13/14	14/13	**25/2**	13/7	-

of evidence theory has no advantage compared to the simple weighted voting approach.

In a comparison between the evidential combination rules, the disjunctive rule stands out as poor. This result should perhaps not be surprising as the rule is very cautious and avoids conflicts by making the resulting mass function increase its represented uncertainty (i.e., it is in a sense striving for a uniform pignistic probability output). For random forest classifiers, which are constructed from the same examples, the classifiers can be expected to be similar, which violates the justification for the disjunctive rule (i.e., that at least one of the sources is known to be incorrect).

Compared to the original Dempster's rule, the more recently introduced non-associative rules have, in general, worse performance (for accuracy and AUC) in our study. Hence, our results do not motivate the use of these rules for our ensemble design. It should be reiterated, however, that the order in which mass functions are combined for non-associative rules affects the performance. In our case, using 25 classifiers means that there are 25! possible ways to combine the classifier outputs. Hence, it would be interesting to more carefully consider the order in which the classifiers are combined.

VI. CONCLUDING REMARKS

We have compared six different evidential combination rules to each other and to two common ensemble rules (voting and weighted voting) for a specific ensemble design involving fusion of propositional output of random forest classification trees. The combination rules have been evaluated w.r.t. accuracy, area under ROC curve and Brier score using ten-fold cross-validation on 27 datasets from the UCI repository. The evaluation shows that weighted voting, Dempster and modified Dempster with uniform prior result in the most accurate predictions, while weighted voting significantly outperforms all other methods w.r.t. AUC and Brier score, except (unweighted) voting, which performs almost as good w.r.t. Brier score. For any particular dataset, though, the experiment demonstrates that it may be worthwhile to consider evaluating multiple combination rules.

In the current study, we have focused on one specific type of classifier, i.e., random forests, in which a single class is output for each tree. One direction for future research is to consider random forests in which each tree outputs a probability measure, e.g. as considered 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. Note also that successful applications of evidential combination rules (in comparison to others) have been reported for classifiers which have the ability to abstain from classifying some examples [8], [10] .

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.

Acknowledgement

We wish to thank Prof. Sten F. Andler (University of Skövde) for his constructive comments. 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.

[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] A. Asuncion and D. Newman, "UCI machine learning repository," 2007. [Online]. Available: http://www.ics.uci.edu/~mlearn/MLRepository.html

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

[19] 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.

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

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

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