International Journal of Approximate Reasoning 53 (2012) 584–607

Contents lists available at SciVerse ScienceDirect

International Journal of Approximate Reasoning

journal homepage: www.elsevier.com/locate/ijar

The impact of diversity on the accuracy of evidential classifier ensembles ☆

Yaxin Bi

School of Computing and Mathematics University of Ulster at Jordanstown, Co Antrim, BT37 0QB, UK

ARTICLE INFO

Article history:
Received 7 June 2010
Received in revised form 23 December 2011
Accepted 29 December 2011
Available online 9 January 2012

Keywords:
Ensemble learning
Diversity
Belief functions
Triplet evidence structure

ABSTRACT

Diversity being inherent in classifiers is widely acknowledged as an important issue in constructing successful classifier ensembles. Although many statistics have been employed in measuring diversity among classifiers to ascertain whether it correlates with ensemble performance in the literature, most of these measures are incorporated and explained in a non-evidential context. In this paper, we provide a modeling for formulating classifier outputs as triplet mass functions and a uniform notation for defining diversity measures. We then assess the relationship between diversity obtained by four pairwise and non-pairwise diversity measures and the improvement in accuracy of classifiers combined in different orders by Demspter's rule of combination, Smets' conjunctive rule, the Proportion and Yager's rules in the framework of belief functions. Our experimental results demonstrate that the accuracy of classifiers combined by Dempster's rule is not strongly correlated with the diversity obtained by the four measures, and the correlation between the diversity and the ensemble accuracy made by Proportion and Yager's rules is negative, which is not in favor of the claim that increasing diversity could lead to reduction of generalization error of classifier ensembles.

1. Introduction

The combination of multiple classifiers (ensemble approach) is a rather powerful decision making and classification technique that has been successfully used for modeling many practical problems such as text categorization [1], remote sensing [2], person recognition (face, finger print) [3], and the handling of imperfect information composed of missing, noisy and fuzzy values in classification [4,5]. In the modeling of classifier combination, many researchers believe that the success of classifier ensembles not only depends on a set of appropriate classifiers, but also on the diversity being inherent in the member classifiers. A good diversity measure would have the ability to find the extent of diversity among classifiers and estimate the improvement or deterioration in accuracy of individual classifiers when they have been combined. Unfortunately to date there has been no widely perceived concept of diversity and there exists no general accepted theoretical framework underpinning the development of methods for capturing diversity among classifiers [6,7]. Although many statistics have been employed to measure diversity with the intention to ascertain whether it correlates with ensemble performance in the literature, results are often varied. Most commonly these measures are incorporated and explained in the context of majority voting, linear sum and other non-evidential frameworks [8]. Presently there is a little effort concerning how diversity measured by statistics imparts ensemble performance in the framework of the Dempster–Shafer (DS) theory of evidence [9], especially, the research on this aspect in the blend of the DS theory and ensemble learning is in its infancy. This study carries out an empirical analysis on the relationship between diversity and accuracy of classifiers based on four pairwise and non-pairwise diversity measures and four evidential combination rules.

☆ This paper is the full version of [20].

E-mail address: y.bi@ulster.ac.uk

0888-613X/$ - see front matter © 2012 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.ijar.2011.12.011

Early studies on the relationship between diversity and ensemble performance have stimulated considerable interest and they can be categorized into two contexts, regression and classification. In [10], Krogh and Vedeldby developed a seminal idea of breaking down generalization errors of a classifier ensemble into a simple linear relationship $E = \bar{E} - \bar{A}$ holding for regression ensembles, where $\bar{E}$ is mean square error used to measure accuracy and $\bar{A}$ is variance used for measuring diversity. This relationship implies that the reduction in generalization errors for an ensemble is directly proportionate to the diversity in the constituent classifiers as measured by the variance of the classifier outputs. Unfortunately there is no such intuitive means for quantifying diversity in classification such that can postulate a similar linear relationship to errors reduction. Nevertheless this linear relationship provides an insight into capturing diversity in classification problems [11].

In the context of classification, Kuncheva and Whitaker carried out an experimental study on relationship between diversity and accuracy [8]. In their work ten statistical diversity measures introduced in the literature, such as Q-statistic, κ-statistic, correlation, etc. have been applied to the classifiers generated by the feature subspace, Bagging and the random weak-classifiers methods. Besides these the classifiers were combined using eight combination methods, including majority vote, Naive Bayes, the Behavior Knowledge Space, the maximum operator, to name but a few. Their results show that although there are proven connections between diversity and accuracy in some special cases, there is no strong linear and non-linear correlation between diversity and accuracy. In [12], Tang et al. conducted a follow-up comprehensive study. They investigate the correlation among the six statistical measures used in [8] and relate these measures to the concept of margin proposed in [13], which is explained as a key factor to the success of Boosting algorithms. The various experimental results demonstrate that large diversity may not consistently correspond to a better ensemble performance and the information perceived by varying diversity cannot provide a consistent guidance on making a classifier ensemble to achieve good generalization performance.

Most recently, Brown et al. proposed a mutual information formulation that measures the mutual dependence between two classifiers [6], meanwhile Zhou and Li developed a multi-information measure that can be used to detect the dependence among multiple classifiers [7]. Both of them are defined on the information theory, but the former formulation is mathematically similar to the latter [7]. However compared with the mutual information, the multi-information formulation is simpler and decomposable over constituent classifiers. These studies would provide a step towards the understanding of ensemble diversity.

A similar concept – conflict – has been covered in the DS theory literature [14,15]. The rationale of studying conflict of evidence sources is rooted in the criticism on the counterintuitive results of applying Dempster's rule of combination to conflicting evidence items, where an almost impossible decision (with a very low degree of confidence) by both evidence sources comes up as the most possible outcome (with certainty). In [14], Smets proposed to use the combined mass values assigned to the emptyset before normalization as a measure of conflict (ME). However in [16], Liu used examples to show that ME might not always be accurate and proposed an alternative method to measure conflict among evidence items by incorporating ME with a constituent measure called distance between betting commitments (DBC), namely ME-DBC. In the present context, if we model classifier decisions/outputs as pieces of evidence, the difference between conflict and conventional diversity is that the former not only considers disagreement between classifiers on decisions like diversity, but also accounts for difference between confidence values quantifying the support for the decisions made by classifiers. Thus measures of conflict cannot be directly converted or reinterpreted in terms of diversity.

In the previous studies [17,18], we have developed new evidence structures called a *triplet* and *quartet* and a formalism for modeling classifier outputs as triplet and quartet mass functions, and we also established a range of formulae for combining these mass functions in order to arrive at a consensus decision. However in those studies we did not address the issues of how diversity impacts the performance of combined classifiers using Dempster's rule of combination. This study extends the work in [20], covering the important aspects of diversity effects on the performance of ensemble classifiers that are independently generated by 13 machine learning algorithms and are combined using purely evidential combination functions in decreasing and mixed orders. We use the triplet as an underlying evidence structure for representing classifiers outputs and study well-known alternative combination rules by incorporating this structure, including Smets' conjunctive rule [14], Dubois and Prade's disjunctive rule [23], Yager's combination [22], and the Proportional rule [21]. We analytically compare these classical combination rules so as to examine their suitability for the triplet structure, and justify the treatment of the empty set for absorbing the conflict with and without normalization operators in the course of combining classifiers and the final classification decision making. Moreover due to the approximation in constructing triplet mass functions, it would breach the associative law held in Dempster's rule of combination, we thereby assess two decreasing and random orders to see what role the orders of classifiers play in interacting with the accuracy of classifier ensembles and the diversity among the member classifiers.

Presently there is a general lack of theory, and there is no general agreement about the notion of diversity and how to quantify the diversity among classifiers. In this study we define the concept of diversity as disagreement among classifiers. We employ statistical measures as diversity measures that are characterized into two types of pairwise and non-pairwise, and develop a uniform notation to define these styles of measures. We also select and implement four typical and commonly used pairwise and non-pairwise measures of Q-statistic [32], disagreement [29], κ-statistic [31] and Kohavi–Wolpert variance [30]. In finding overall diversity of a group of classifiers, pair-wise measures average diversity over all classifiers, while non-pairwise measures attempt to measure diversity among classifiers directly based on, for example, variance or entropy of classifiers that fail on randomly selected patterns. In both cases, classifier outputs at the final stage are formulated in the binary form of correct and incorrect, which is widely used by researchers.

The paper is organized as follows. In Section 2, we describe the representation of classifier outputs. In Section 3, we shortly summarize basics of the DS theory of evidence and the triplet mass function, and then analytically compare evidential combination rules. We formulate pairwise and non-pairwise diversity measures in a uniform way and define four statistical diversity measures in Section 4. Finally, we present our experimental methodology and experimental results in Section 5 and summarize the findings in Section 6.

2. Representation of classifier outputs

In ensemble learning, a learning algorithm is provided with a training data set made up of $D \times C = \{\langle d_1, c_1\rangle, \ldots, \langle d_{|D|}, c_q\rangle\}$ $(1 \leq q \leq |C|)$ for deriving some unknown function f such that $f(d) = c$. Instance $d_i \in D$ is characterized by a vector in the form of $(d_{i_1}, \ldots, d_{i_n})$ where d_{i_j} is typically either a nominal or ordinal value, and c_i is typically drawn from a set of categorical classes C in terms of class labels. Given a set of training data $D \times C$, a learning algorithm is aimed at learning a function φ in terms of classifier, where classifier φ is an approximation to an unknown function f.

Given a new instance d, a classification task is to make decision for d using φ about whether d belongs to class c_i. Instead of single-class assignment, we regard such a classification process as a mapping:

$$\varphi : D \to C \times [0,\ 1], \tag{1}$$

where $C \times [0,\ 1] = \{(c_i,\ s_i) \mid c_i \in C,\ 0 \leq s_i \leq 1\}$, s_i is a numeric value that can be in different forms, such as a similarity score, a class-conditional probability or other measurements, depending on the types of learning algorithms. This numeric value represents the degree of support or confidence about the proposition of that instance d is assigned to class c_i. The greater the value of class s_i, the greater the amount of belief given to the proposition of instance d belonging to class c_i. Simply we denote a classifier output by $\varphi(d) = \{s_1, \ldots, s_{|C|}\}$. Given a group of classifiers, $\varphi_1, \varphi_2, \ldots, \varphi_M$, all the classifier outputs on instance d can be organized into a matrix as illustrated in formula (2).

$$\begin{pmatrix} \varphi_1(d) \\ \varphi_2(d) \\ \vdots \\ \varphi_M(d) \end{pmatrix} = \begin{pmatrix} s_{11} & s_{12} & \ldots & s_{1|C|} \\ s_{21} & s_{22} & \ldots & s_{2|C|} \\ \vdots & \vdots & \ldots & \vdots \\ s_{M1} & s_{M2} & \ldots & s_{M|C|} \end{pmatrix}. \tag{2}$$

3. Basics of the Dempster–Shafer (DS) theory of evidence, triplet mass function and combination rules

The Dempster–Shafer theory of evidence remedies the limitations of the traditional Bayesian belief model to allow the explicit representation of uncertainty and management of conflict information involved in the decision making process [9]. The advantages of the DS theory over other approaches are the ability to (1) model the narrowing of the hypothesis set with the accumulation of evidence (via the evidence combination operation) [25], (2) explicitly represent uncertainty in the form of ignorance, and (3) handle the reliability of information sources by means of the discounting operation.

The DS theory formulates decision making process as pieces of evidence and propositions, and subjects these to a strict formal process in order to infer conclusions from the given uncertain evidence, avoiding human subjective intervention to some extent [18]. Formally DS formulates a proposition set as a frame of discernment, denoted by $\Omega = \{c_1, \ldots c_{|\Omega|}\}$. The power set 2^Ω is all the subsets of Ω. The basis of a belief measure for focal elements in 2^Ω is the basic probability assignment, called mass function.

Definition 1. Let Ω be a frame of discernment. Let m be a mass function, which is defined as a assignment function assigning a numeric value in $[0, 1]$ to $X \in 2^\Omega$ with two conditions below.

$$(1)\ m(\emptyset) = 0, \quad (2)\ \sum_{X \subseteq \Omega} m(X) = 1,$$

where $X \subseteq \Omega$ is called a focal element, focus or singleton if $m(X) > 0$. It represents a proposition of interest.

Since mass functions are defined on all the subsets of the frame of discernment Ω, instead of reckoning the individual propositions themselves as in probability theory, DS is capable of precisely apportioning the probability mass to propositions that are supported by evidence without considering assignments to those levels of detail that there is no knowledge about other propositions. Such a mechanism allows us to model any particular subset of Ω that is uncertain or unknown in classification decision processes.

Given the general representation of classifier outputs in formula (2), on the basis of Definition 1, we define an application-specific mass function below.

Definition 2. Let Ω be a frame of discernment and let $\varphi(d)$ be a list of scores, an application-specific mass function is defined a mapping function, $m : 2^{\Omega} \rightarrow [0, 1]$ as follows:

$$m(\{c_i\}) = \frac{s_i}{\sum_{j=1}^{|\Omega|} s_j}, \tag{3}$$

where $c_i \in \Omega$ for $1 \leq i \leq |\Omega|$.

This mass function expresses the degrees of belief with respect to determining class labels to which a given instance could belong.

Definition 3. Let Ω be a frame of discernment. Let *bel* be a belief function for $X \subseteq \Omega$, which is defined as the sum of probability masses supporting all the subsets of X.

$$bel(X) = \sum_{A \subseteq X} m(A). \tag{4}$$

On the basis of the mass function and belief function, a plausibility function for any focus $X \subseteq \Omega$ is further defined as the probability mass not supporting $\bar{X}$ as follows.

$$pls(X) = 1 - bel(\bar{X}) = \sum_{A \cap X \neq \emptyset} m(A) \quad \text{where } A \subseteq \Omega. \tag{5}$$

The difference $pls(X) - bel(X)$ is regarded as a measure of ignorance about X, denoted by $ign(X)$. The ignorance $ign(X) = 0$ indicates that the degree of belief about X is the same as that of plausibility, while $ign(X) = 1$ means that no probability mass is assigned to X (or its subsets), and equally no mass is assigned to $\bar{X}$. In addition we utilize masses allocated to Ω as a measurement for quantifying the degree of ignorance about a frame of discernment.

Definition 4. Let Ω be a frame of discernment. Let m_1 and m_2 be two mass functions defined for X, $Y \subseteq \Omega$. Dempster's rule of combination (or Dempster's rule) denoted by $\oplus$, is defined as

$$(m_1 \oplus m_2)(A) = \frac{\sum_{X \cap Y = A} m_1(X) m_2(Y)}{1 - E}, \tag{6}$$

where operator $\oplus$ is also called the *orthogonal sum* and $E = \sum_{X \cap Y = \emptyset} m_1(X) m_2(Y)$ is called the *conflict factor*. This rule strongly emphasizes the agreement between multiple independent sources and ignores all the conflicting evidence through a normalization factor. Specifically, this rule can calculate the mass for a new proposition $A \subseteq \Omega$ from the probability masses assigned to X and Y, resulting in the accumulation of knowledge from the evidence sources supporting X and Y, respectively. As new pieces of evidence become available, masses committed to only those propositions that are supported by evidence, are therefore combined using Demspter's rule of combination to give a new set of propositions that are supported by the *combined* evidence.

3.1. Triplet mass function and computation

Given the formulation of classifier outputs in formula (2), by formula (3), we can rewrite $\varphi(d)$ as $\varphi(d) = \{m(\{c_1\}), m(\{c_2\}), \ldots, m(\{c_{|\Omega|}\})\}$, referred to as a list of decisions — a piece of evidence. By formula (6) two or more pieces of evidence can then be combined to make the final classification decision. To improve the efficiency of computing the orthogonal sum operation and the accuracy of the final decision on the basis of the combined results, a new structure, called a triplet, has been developed [17]. The following details this novel structure, partitioning a list of decisions $\varphi(d)$ into three subsets.

Definition 5. Let Ω be a frame of discernment and $\varphi(d) = \{m(\{c_1\}), m(\{c_2\}), \ldots, m(\{c_{|\Omega|}\})\}$, where $|\Omega| \geq 2$, an expression of the form $Y = \langle \{u\}, \{v\}, \Omega \rangle$ is defined as a *triplet*, where $\{u\}$, $\{v\}$ are singletons, Ω is the whole set of classes C, and they satisfy

$$m(\{u\}) + m(\{v\}) + m(\Omega) = 1.$$

Based on the number of singleton decisions, we also refer to a triplet as a structure of *two-point focuses*, and call the associated mass function a *two-point mass function*. To obtain triplet mass functions, we define a focusing operation in terms of the *outstanding rule* and denote it by m^{σ} as follows:

$$\{u\} = arg\max(\{m(\{c_1\}), m(\{c_2\}), \ldots, m(\{c_{|\Omega|}\})\}), \tag{7}$$

$$\{v\} = arg\max(\{m(\{c\}) | \, c \in \{c_1, \ldots, c_{|\Omega|}\} - \{u\}\}), \tag{8}$$

$$m^{\sigma}(\Omega) = 1 - m^{\sigma}(\{u\}) + m^{\sigma}(\{v\}). \tag{9}$$

We refer to m^{σ} as a *triplet mass function* or as a *two-point mass function*, simply m. By applying formulas (3), (7), (8), and (9), formula (2) is simply rewritten as formula (10) below.

$$\begin{pmatrix} \varphi_1(d) \\ \varphi_2(d) \\ \vdots \\ \varphi_M(d) \end{pmatrix} = \begin{pmatrix} m_1(\{u_1\}) & m_1(\{v_1\}) & m_1(\Omega) \\ m_2(\{u_2\}) & m_2(\{v_2\}) & m_2(\Omega) \\ \vdots & \vdots & \vdots \\ m_M(\{u_M\}) & m_M(\{v_M\}) & m_M(\Omega) \end{pmatrix}. \tag{10}$$

From the above formulation it can be seen that a triplet mass function is a support function to a set of possible classes $\{\{u\}, \{v\}, \Omega\}$, which in turn can be easily proved as a belief function [18]. Thus for a given instance belief value $bel\{u\}$ represents the maximum of quantitative judgments, indicating class u will be assigned to the instance. Thus we use the maximal selection as the classification decision making rule, i.e. class with maximal belief value will be assigned to unseen instances.

3.2. Alternative combination rules

This section reviews several well-known alternative combination rules and analyzes their common features in dealing with conflict encountered in independent evidence sources. To make the rationale behind these rules more clear, let us first consider a well-known example, showing what is a counterintuitive effect caused by the normalization of Dempster's rule of combination.

Example 1. Let two mass functions m_1 and m_2 be defined on $\Omega = \{x_1, x_2, x_3\}$. Let

$$m_1(\{x_1\}) = 0.80, \quad m_1(\{x_2\}) = 0.20;$$

$$m_2(\{x_3\}) = 0.95, \quad m_2(\{x_2\}) = 0.05.$$

The mass values computed by Dempster's rule are as follows.

$$E = \textstyle\sum_{X\cap Y=\emptyset} m_1(X)m_2(Y) = 0.99, \quad \text{where } X, Y \in \{\{x_1\}, \{x_2\}, \{x_3\}\},$$
$$(m_1 \oplus m_2)(\{x_1\}) = 0/(1 - 0.99) = 0,$$
$$(m_1 \oplus m_2)(\{x_2\}) = 0.01/(1 - 0.99) = 1,$$
$$(m_1 \oplus m_2)(\{x_3\}) = 0/(1 - 0.99)$$

and

$$bel(\{x_2\}) = (m_1 \oplus m_2)(\{x_2\}) = 1,$$
$$pls(\{x_2\}) = 1 - bel(\{\bar{x_2}\}) = 1 - bel(\{x_1, x_3\} = 1,$$

where the conflict factor $E = 0.99$, indicating that the two pieces of evidence of supporting m_1 and m_2 are largely in conflicting. As a result, the application of Dempster's rule to these pieces of evidence makes x_2 with a full support. As illustrated in the example, x_2 is weakly supported by the respective evidence sources, but after combining the two pieces of evidence, it is fully supported. This combined effect is *counter-intuitive*.

Such a counter-intuitive result has led a great deal of debate in the past decades. Many researchers believe that the counter-intuitive result is due to the normalization operation in Dempster's rule of combination, whereas others defend Dempster's rule in the sense that the counterintuitive result can be avoided if the respective masses are apportioned in an appropriate way. In line with avoiding the counter effect caused by the normalization, several alternatives have been proposed and well documented in the literature. In [22], Yager proposed to allocate the combined masses of conflict to the frame of discernment Ω. Dubois and Prade [23] proposed a disjunctive combination rule. Smets [14] proposed an unnormalized combination rule, which is known as the conjunctive combination rule. In [21], Anand et al. developed a Proportion rule by taking into account average values of columns and rows in an intersection table. More recently, Denoeux proposed a cautious rule that was claimed to design for combining non distinct (dependent) items of evidence by accounting for dependence and overlapping of evidence bodies [26]. As opposite to the non distinct evidence, this study is focused on formulating distinct (independent) items of evidence to be combined by Dempster's rule and its classical alternatives. As illustrated in Section 3.1, we formulate classifier outputs as triplet mass functions, in which the classifiers are generated by different learning algorithms. In this way, the classifier outputs are treated as distinct (independent) bodies of evidence as different algorithms were developed on the basis of the different theories and the classifiers generated thus utilize their

own mechanisms to make classification decisions on any instances without use of "overlapping experience". The detailed discussion on this aspect can be referred to our previous work in [18].

In the next subsections we therefore focus our study on the classical rules, i.e. comparatively analyze these alternatives and examine their suitability for combining multiple classifiers whose outputs are represented in triplets.

Definition 6. Suppose m_1 and m_2 are two mass functions on the frame of discernment Ω. Let X and Y be subsets of Ω. Yager's combination rule is defined as:

$$(m_1 \textcircled{Y}\, m_2)(A) = \begin{cases} 0 & \text{if } A = \emptyset, \\ \sum_{X\cap Y=A} m_1(X)m_2(Y) & \text{if } \emptyset \subset A \subset \Omega, \\ \sum_{X\cap Y=\Omega} m_1(X)m_2(Y) + \sum_{X\cap Y=\emptyset} m_1(X)m_2(Y) & \text{if } A = \Omega. \end{cases} \tag{11}$$

Yager's rule keeps the condition of $m(\emptyset) = 0$ and adds masses allocated to both the empty set and the frame of discernment together into Ω. The mass $m(\Omega)$ represents the degree of ignorance about Ω.

Definition 7. Let m_1, m_2 be two mass functions defined on the frame of discernment Ω. Let X and Y be subsets of Ω. Dubois and Prade's rule, denoted by $\textcircled{\sigma}$, is then given below:

$$(m_1 \textcircled{\sigma}\, m_2)(A) = \begin{cases} 0 & \text{if } A = \emptyset, \\ \sum_{X\cap Y=A} m_1(X)m_2(Y) \sum_{X\cup Y=A,\ X\cap Y=\emptyset} m_1(X)m_2(Y) & \text{if } A \subseteq \Omega. \end{cases} \tag{12}$$

This rule is often referred to as the disjunctive combination rule. It keeps the condition of $m(\emptyset) = 0$ as in Dempster's rule and transfers masses resulting from pairs of conflicting elements ($X \cap Y = 0$) to the union of these elements.

Definition 8. Let m_1, m_2 be two mass functions defined on the frame of discernment Ω. Let X and Y be subsets of Ω. Smets' rule of combination is given by

$$(m_1 \textcircled{\cap}\, m_2)(A) = \begin{cases} \sum_{X\cap Y=\emptyset} m_1(X)m_2(Y) & \text{if } A = \emptyset, \quad \text{and } X, Y \subseteq \Omega, \\ \sum_{X\cap Y=A} m_1(X)m_2(Y) & \text{if } A \subseteq \Omega. \end{cases} \tag{13}$$

Smets' rule is known as the conjunctive rule. This rule adds up all the masses resulting from the empty intersections of subsets into $m(\emptyset)$ as the degree of conflict, but it violates the condition of $m(\emptyset) = 0$ as specified in Dempster's rule.

Definition 9. Suppose m_1 and m_2 are two mass functions on the frame of discernment Ω. Let m_1 have n focal elements: $X_1, \ldots X_n$, and m_2 have m focal elements: $Y_1, \ldots, Y_m$, where X_i, Y_j be subsets of Ω ($1 \le i \le n$, $1 \le j \le m$). The Proportion combination rule is defined as:

$$(m_1 \textcircled{a}\, m_2)(A) = \begin{cases} 0 & \text{if } A = \emptyset, \\ \left(\sum_{X_i\cap Y_j=A} m_c(X_{ij}) + \sum_{X_i\cap Y_j=A} m_r(Y_{ij})\right) \Big/ 2 & \text{if } A \neq \Omega, \end{cases} \tag{14}$$

where $m_r(X_{ij})$ and $m_c(Y_{ij})$ are the average of mass functions based on rows and columns in an intersection table, which can be calculated below:

$$m_r(X_{ij}) = \begin{cases} m_1(X_i) \times \frac{m_2(Y_j)}{\sum_{X_i\cap Y_k\neq\emptyset} m_2(Y_k)}, & \text{where } 1 \le k \le m, \\ m_1(X_i), & \text{otherwise.} \end{cases}$$

$$m_c(Y_{ij}) = \begin{cases} m_2(Y_j) \times \frac{m_1(X_i)}{\sum_{X_k\cap Y_j\neq\emptyset} m_1(X_k)}, & \text{where } 1 \le k \le n, \\ m_2(Y_j), & \text{otherwise.} \end{cases}$$

3.3. A comparison of the combination rules on triplet

The literature review on information fusion with the DS theory shows that much of the research has been devoted to the theoretical justification of the alternative combination rules and their selection [28], and the conditions of applying conjunctive or disjunctive combination [15]. To avail the strength of these alternatives in constructing ensemble classifiers,

it would be very helpful to have an appropriate understanding of their properties and interoperability with the triplet structure.

Theoretically, the rules given in Eqs. (6) and (11)–(13) share the same basis [15]. Suppose we have two mass functions m_1 and m_2 defined on the frame of discernment Ω and the mapping function $m^\tau : 2^\Omega \to [0, 1]$ with:

$$m^\tau(A) = \sum_{X \cap Y = A} m_1(X) m_2(Y), \quad A \subseteq \Omega. \tag{15}$$

Formula $m^\tau(A)$ is the core part of these alternative rules, which is also called the conjunctive rule. This notation discloses the fact that the difference between the alternatives is the way of apportioning the remaining masses committed to the empty intersections (non-intersections) in terms of the conflict factor E. Specifically, Dempster's rule reallocates them as a normalization factor, Yager's rule puts them into the frame of discernment, Dubois and Prade's rule distributes them into the unions of conflicting subsets and Smets' rule retains them in the emptyset. These combination rules employ their own strategies to handle the remaining masses and make up their own strengths. However it could be envisaged that none of them offers a generalized solution to combining pieces of evidence that are largely in conflict.

As described in Section 3.1, the classifier outputs are modeled as triplet mass functions which can be obtained by Eqs. (7)–(9). To see how these rules can be used to combine triplet mass functions, it is necessary to examine the relations between any two pairs of focal elements in two triplets. Given a setting in which there are two triplets $\langle\{u_1\}, \{v_1\}, \Omega\rangle$ and $\langle\{u_2\}, \{v_2\}, \Omega\rangle$ where $u_i, v_i \in \Omega$ $(i = 1, 2)$, and the associated triplet mass functions m_1 and m_2. The relation between $\{u_1\}, \{v_1\}$ and $\{u_2\}, \{v_2\}$ can be permutated into three situations: completely different focal points (completely inconsistent), one focal point equal(partially consistent) and two focal points equal (totally consistent).

Now let us consider the case, for instance, where $\{u_1\}, \{v_1\}$ in one triplet are completely inconsistent to $\{u_2\}, \{v_2\}$ in another triplet, i.e. $u_1 \neq u_2, u_1 \neq v_2, v_1 \neq u_2$, and $v_1 \neq v_2$. By applying Yager's rule to combine m_1 and m_2, we have

$$m^\tau(\{u_1\}) = (m_1 \textcircled{Y} m_2)(\{u_1\}) = m_1(\{u_1\}) m_2(\Omega), \tag{16}$$

$$m^\tau(\{v_1\}) = (m_1 \textcircled{Y} m_2)(\{v_1\}) = m_1(\{v_1\}) m_2(\Omega), \tag{17}$$

$$m^\tau(\{u_2\}) = (m_1 \textcircled{Y} m_2)(\{u_2\}) = m_1(\Omega) m_2(\{u_2\}), \tag{18}$$

$$m^\tau(\{v_2\}) = (m_1 \textcircled{Y} m_2)(\{v_2\}) = m_1(\Omega) m_2(\{v_2\}), \tag{19}$$

$$\begin{aligned}(m_1 \textcircled{Y} m_2)(\Omega) = {} & m_1(\Omega) m_2(\Omega) + m_1(\{u_1\}) m_2(\{u_2\}) + m_1(\{u_1\}) m_2(\{v_2\}) + m_1(\{v_1\}) m_2(\{u_2\}) \\ & + m_1(\{v_1\}) m_2(\{v_2\}).\end{aligned} \tag{20}$$

Eqs. (16)–(19) constitute the core part of calculation that are commonly required by these alternatives in calculating masses associated with intersected subsets. However Eq. (20) provides its own way in computing masses resulted from the non-intersections. This way cannot be generalized to calculating masses associated with the emptyset as required by the other rules. Thus we need to work out how the masses associated with non-intersections can be computed by Dubois and Prade's rule and Smets' rule, separately.

By applying Dubois and Prade's rule to combine m_1 and m_2, we have

$$(m_1 \textcircled{D} m_2)(\{u_1\} \cup \{u_2\}) = m_1(\{u_1\}) m_2(\{u_2\}), \tag{21}$$

$$(m_1 \textcircled{D} m_2)(\{u_1\} \cup \{v_2\}) = m_1(\{u_1\}) m_2(\{v_2\}), \tag{22}$$

$$(m_1 \textcircled{D} m_2)(\{v_1\} \cup \{u_2\}) = m_1(\{v_1\}) m_2(\{u_2\}), \tag{23}$$

$$(m_1 \textcircled{D} m_2)(\{v_1\} \cup \{v_2\}) = m_1(\{v_1\}) m_2(\{v_2\}), \tag{24}$$

$$(m_1 \textcircled{D} m_2)(\Omega) = m_1(\Omega) m_2(\Omega). \tag{25}$$

Intuitively the results given by Eqs. (21)–(25) on non-intersections are different from those produced by Eq. (20). But if we transform these results into a triplet mass function, we only need to preserve two singletons whose masses are the largest and second largest, and merge the rest subsets along with the emptyset into the frame of discernment Ω. In this way, the resulting triplet mass function is the same as the one obtained from the results produced by Yager's rule.

Similarly, by applying Smets' rule, we can obtain the respective calculation on non-intersections below,

$$(m_1 \textcircled{\cap} m_2)(\Omega) = m_1(\Omega) m_2(\Omega), \tag{26}$$

$$(m_1 \textcircled{\cap} m_2)(\emptyset) = m_1(\{u_1\}) m_2(\{u_2\}) + m_1(\{u_1\}) m_2(\{v_2\}) + m_1(\{v_1\}) m_2(\{u_2\}) + m_1(\{v_1\}) m_2(\{v_2\}). \tag{27}$$

From the above derivations, it can be seen that the results obtained Eqs. (26) and (27) are different from those with Eqs. (21)–(25) unless the masses calculated by the latter are transferred to $m(\emptyset)$. Theoretically the emptyset can be treated as a valid focal element in Ω, naturally in the triplet structure. By contrast Smets' rule permits $m(\emptyset) \neq 0$, which violates the condition of mass functions. Consequently the triplet mass functions obtained with Smets' rule differ from those produced by Dubois and Prade's rule, and Yager's rule, respectively.

What is gained by the preceding analysis is the different formulations with the triplet functions for these alternative rules in calculating masses for non-intersections. The differences between these formulas raise a question of how the emptyset should be handled in the context of classification decision making. If we assume that the emptyset is a valid focal element in triplets, it is likely that the emptyset will be assigned to a given instance. However such an assignment would result in a meaningless class for the instance, violating the closed world assumption made in supervised machine learning tasks. To avoid a meaningless assignment a realistic approach could thereby amalgamate the emptyset with the frame of discernment. As such, the role played by Yager's rule in combining triplet mass functions is tantamount to that as Dubois and Prade's rule, and Smets' rule, respectively. Meanwhile, to examine the effect of the emptyset in the course of combining evidence by Smets's rule, we keep the emptyset in the process of combining evidence and use it to absorb the conflict being inherent in evidence sources, but the final classification decision will be made on the basis of the largest combined mass value committed to a non-emptyset focal element, instead of the emptyset itself. Such a treatment not only conforms to the closed world assumption held in ensemble learning tasks, but also preserves the conflict in the emptyset, reducing the impact of the conflict in combining more pieces of evidence to some extent. Therefore to investigate the typical situations of normalization, non-normalization with and without the emptyset, and proportion in the process of accumulating pieces of evidence, we select Dempster's rule, Smets' rule, Yager's rule and the Proportional rule for the further empirical study.

The above comparative analysis is based on the case where two triplets are completely inconsistent. In the same way we can generalize the analysis to the other two cases and accordingly derive their formulas.

4. Diversity measures

Statistical diversity measures can be divided into pairwise and non-pairwise measures. The pairwise measures calculate the average of a particular agreement/disagreement metric between all possible pairings of classifiers in ensemble classifiers. Thus the metric characterizes diversity measures. The non-pairwise measures either use the idea of entropy or calculate a correlation of each ensemble member with the (weighted) arithmetic mean of the individual outputs. To date there has been no convincing theory or experimental study to suggest which of statistical measures can be best and reliably used to improve ensemble performance [8,12,6,7]. In this study, based on the way of measuring agreement and disagreement between classifier outputs, we select four measures, which have been widely discussed in the literature, to measure diversity among classifiers in the form of binary outputs as done in most studies.

Formally suppose we are given M classifiers denoted by $\varphi_1, \ldots, \varphi_M$, a set of classes $\Omega = \{c_1, \ldots, c_{|\Omega|}\}$ and a test set $T = \{x_1, \ldots, x_{|T|}\}$. For any instance $x \in T$, each classifier produces an output vector $\varphi_i(x)$. Conventionally, classifying x means assigning it into one class in Ω, i.e., deciding if x belongs to c_k, $k = 1, \ldots, |\Omega|$ according to $\varphi_i(x)$. For sake of binary outputs of classifiers, we model the final output of $\varphi_i(x)$ as a class label or a binary output, denoted by ψ_i. For the former, $\varphi_i(x) = c$ where $c \in \Omega$. In the latter case, $\psi_i(x) = 1$ if ψ_i correctly classifies x, whereas $\psi_i(x) = 0$ if ψ_i incorrectly classifies x. We also denote $\widehat{\psi}(x) = \{\psi(x) | \psi_i(x) = 1,\ 1 \leq i \leq M,\ x \in T\}$. With this notation, The four statistical diversity measures are defined below.

4.1. Kappa (κ) statistic

The κ statistic is the most widely used pairwise method to measure the level of agreement between classifiers [31]. It can be thought of as chance-corrected Proportional agreement [33]. Given two classifiers ψ_i and ψ_j and a test data set T, we can construct a global contingency table based on a set of classes Ω. The table entry $n(c_h, c_k)$ contains the number of instances $x \in T$ for $\psi_i(x) = c_h$ and $\psi_j(x) = c_k$. If ψ_i and ψ_j are identical on the data set, then all non-zero counts will appear along the diagonal of the table, otherwise there will be a number of counts off the diagonal. Now we define

$$\mu_1 = \frac{\sum_{h=1}^{|\Omega|} n(c_h, c_h)}{|T|},$$

$$\mu_2 = \sum_{h=1}^{|\Omega|} \left(\sum_{k=1}^{|\Omega|} \frac{n(c_h, c_k)}{|T|} \times \sum_{k=1}^{|\Omega|} \frac{n(c_h, c_k)}{|T|} \right),$$

where μ_1 is an estimation of the probability that two classifiers agree and μ_2 is a correction term for μ_1, estimating the probability that the two classifiers agree simply by chance. Then the $\kappa_{i,j}$ statistic over T is defined as follows:

$$\kappa_{i,j} = \frac{\mu_1 - \mu_2}{1 - \mu_2}. \tag{28}$$

The average κ statistic over the whole set of classifiers over T is then defined as follows:

$$\kappa = \frac{2}{M(M-1)} \sum_{i=1}^{M} \sum_{j=i+1}^{M} \kappa_{i,j}, \tag{29}$$

where κ is within the range from 0 to 1. If $\kappa = 0$ then it means that the agreement of classifiers equals that expected by chance, while $\kappa = 1$, indicating that classifiers agree on all the test instances, and negative values of κ mean that the agreement is less than expected by chance.

4.2. Disagreement measure

The disagreement measure is used to characterize the diversity between one classifier and its complementary classifier [29]. In [34], Ho employed it to assess the diversity in decision forests. It is the ratio between the number of binary outputs on which one classifier is correct and the other is incorrect to the total number of classifier outputs in a binary form. Formally, given two classifiers ψ_i and ψ_j, a and b, and a test instance $x \in T$, we can construct a contingency table based on the binary outputs of ψ_i and ψ_j as detailed at the beginning of Section 4, where a takes value on 1 if ψ_i correctly classifies x, 0, otherwise; and b performs in the same way for ψ_j. The table entry $n(a, b)$ is the total number of binary outputs over all the test instances in T. The disagreement between two classifiers is measured by:

$$dis_{i,j} = \frac{n(0,1) + n(1,0)}{n(0,0) + n(0,1) + n(1,0) + n(1,1)}. \tag{30}$$

The disagreement diversity among the whole set of classifiers over T is then defined as an average over all the pairs of disagreement below:

$$dis = \frac{2}{M(M-1)} \sum_{i=1}^{M} \sum_{j=i+1}^{M} dis_{i,j}. \tag{31}$$

Since for any pair of classifiers: $n(0, 0) + n(0, 1) + n(1, 0) + n(1, 1) = |T|$, we thus have:

$$dis = \frac{2}{|T|M(M-1)} \sum_{i=1}^{M} \sum_{j=i+1}^{M} (n_{i,j}(0,1) + n_{i,j}(1,0)). \tag{32}$$

The diversity increases with increasing values of the disagreement measure in the range from 0 to 1.

4.3. Q-statistic

The Q-statistic (qs) is a well studied measure in statistics [8,12]. As introduced in Section 4.2, a contingency for two classifiers ψ_i and ψ_j over a test set T can be constructed. The disagreement between ψ_i and ψ_j is then measured by:

$$Q_{i,j} = \frac{n(0,0)n(1,1) - n(1,0)n(0,1)}{n(0,0)n(1,1) + n(1,0)n(0,1)}, \tag{33}$$

where $Q_{i,j}$ is a measurement of diversity, and the notion of $n(a, b)$ is the same as in Section 4.2. When $Q_{i,j} = 1$ indicates that all the class labels assigned by ψ_i for instances $x \in T$ are exactly the same as ones assigned by ψ_j. $Q_{i,j} = -1$ means that all the class labels recognized by ψ_i for instances $x \in T$ are entirely different from those that ψ_j recognizes. For a set of classifiers, the averaged Q statistic over all the pairs of classifiers with T is measured by

$$Q = \frac{2}{M(M-1)} \sum_{i=1}^{M} \sum_{j=i+1}^{M} Q_{i,j}, \tag{34}$$

where Q varies in the range from -1 to 1. A positive Q means that a classifier ensemble tends correctly to classify the same instance, otherwise it incorrectly classifies the instance.

4.4. Kohavi–Wolpert variance

Kohavi and Wolpert [30] proposed a formula for representing the classification errors of classifiers. This formula is built on the basis of the bias-variance decomposition of errors of classifiers. The expression of the variability of a predicted class

label $c \in \Omega$ for an instance $x \in T$ is

$$variance_x = \frac{1}{2}\left(1 - \sum_{i=1}^{|\Omega|} P(c = c_i|x)^2\right), \tag{35}$$

where $|\Omega|$ is the total number of classes and $P(c = c_i|x)$ is the posterior probability of $c = c_i$ given the evidence of x. In [8], Kuncheva and Whitaker adopted formula (34) to the problem of oracle outputs in classifier ensembles, where $|\Omega| = 2$ and $P(c = 1|x) + P(c = 0|x) = 1$, formula (34) is thus rewritten as:

$$\begin{aligned} variance_x &= \frac{1}{2}(1 - P(c = 1|x)^2 - P(c = 0|x)^2), \\ &= P(c = 1|x)P(c = 0|x), \end{aligned} \tag{36}$$

where $P(c = 1|x)$ is estimated by $\left|\widehat{\psi}(x)\right| \Big/ M$ and $P(c = 0|x)$ is estimated by $\left(M - \left|\widehat{\psi}(x)\right|\right) \Big/ M$. Averaged *variance* over the whole set of testing data T, we have a revised measure, denoted by kw, which can be used to measure the diversity among the whole set of classifiers:

$$kw = \frac{1}{|T|M^2}\left(\sum_{i=1}^{|T|} \left|\widehat{\psi}(x_i)\right| \left(M - \left|\widehat{\psi}(x_i)\right|\right)\right). \tag{37}$$

The diversity increases with increasing values of the kw variance in the range from 0 to 1.

5. Experimental analysis

5.1. Experimental settings

In our experiments, we used twelve data sets downloaded from the UCI machine learning repository [36]. All the selected data sets have at least three or more classes as required by the triplet structure. The details about these data sets can be found in Table 1.

For generating individual (base) classifiers, we used thirteen learning algorithms which are taken from the Waikato Environment for Knowledge Analysis (Weka) version 3.4 (see Table 2). These algorithms were simply chosen on the basis of the performance over three data sets which were randomly picked. They can make up various ensembles of classifiers. Parameters used for each algorithm in this empirical study were set at the default settings. Detailed description of these algorithms can be found in [37].

The experiments were performed using a ten-fold cross validation to avoid overfitting to some extent. We divided each of the data sets into 10 mutually exclusive subsets – 10 folds. Each of the 10 subsets was in turn used as a test set and all the remaining subsets were used for generating classifiers by the 13 learning algorithms (see Table 2). In this way, each of the learning methods generated 10 classifiers to be combined, and each of the classifiers was tested once. Accordingly each of the combined classifiers was tested once as well. The performance of a classifier ensemble was the average of the ten testing results of the combined classifiers, which is more robust than any single tests. The detailed evaluation methodology can be found in [37].

Due to the approximation in transforming the combined results into triplet functions, the associativity of triplets may not be held in the combining process. To faithfully reflect the performance of combined classifiers, combining classifiers has been carried out in two orders of decreasing and mixture by Dempster's rule and its alternatives. For the combination of classifiers in decreasing order, we first rank all the 13 classifiers generated by the 13 learning algorithms, and then we combine the best classifier with the second best as a classifier ensemble, denoted by 2C, and combine the combined result 2C

Table 1
The general description about the datasets.

Dataset	Instance	No classes	Attribute
Anneal	798	6	38
Audiology	200	23	69
Balance	625	3	4
Car	1728	4	6
Glass	214	7	9
Autos	205	6	25
Iris	150	3	4
Letter	20000	26	16
Segment	1500	7	19
Soybean	683	19	35
Wine	178	3	13
Zoo	101	7	17

Table 2
General description of the thirteen learning algorithms.

No	Classifier	Description
0	AOD	Perform classification by averaging over all of a small space of alternative Naive–Bayes-like models that have weaker independence assumptions than naive Bayes
1	NaiveBayes	The Naive Bayes classifier using kernel density estimation over multiple values for continuous attributes, instead of assuming a simple normal distribution
2	SMO	Sequential minimal optimization algorithm for training a support vector classifier using polynomial kernels
3	IBk	A instance-based learning algorithm
4	IB1	The IBk instance-based learner with K = 1 nearest neighbors, in order to offset KStar with a maximally local learner
5	KStar	The K instance-based learner using all nearest neighbors and an entropy-based distance
6	DecisionStump	Building and using a decision stump, but it is not used in conjunction with a boosting algorithm
7	J48	Decision tree induction, a Java implementation of C4.5
8	RandomForest	Constructing random forests for classification
9	DecisionTable	A decision table learner
10	JRip	A propositional rule learner – a Java implementation of Ripper
11	NNge	Nearest neighbor-like algorithm using non-nested generalized exemplars
12	PART	Generating a PART decision list for classification

with the third best as another ensemble, denoted by 3C, and so forth, until combine the combined result of the 12 classifiers with the 13th classifier, denoted by 13C. With respect to the combination of classifiers in mixed order, the order of classifiers is random. We first pick up two classifiers to combine, and then combine the combined result with the third classifier that is randomly chosen, until combine the resulting ensemble with the last classifier. The notation used in mixed order is the same as in decreasing order. Additionally, we use 1C to represent the best classifiers in decreasing order and the first randomly picked classifier in mixed order.

To assess how the accuracy of classifier ensembles and diversity among constituent classifiers is actually correlated, we carried out correlation analyses over the 12 data sets, resulting in a set of pairing correlation coefficient $r \in [-1, 1]$ and $p\text{-}value \in [0, 1]$. A positive correlation coefficient r indicates a positive correlation between the ensemble accuracy and diversity among its member classifiers, whereas a negative number indicates a negative correlation. In particular, a negative correlation indicates that ensemble accuracy increases while diversity decreases. The closer the value of is to 0, the smaller the correlation. The perfect relationship exists with a value of 1 or -1 whereas no correlation exists with a value of 0. On the other hand, *p-value* indicates the degree of that the correlation is statistically significant.

To further quantify the relationship between the diversity and an improvement in accuracy over different groups of classifiers, we calculate the mean accuracy of the different groups of classifiers that make up the corresponding classifier ensembles, and then calculate differences between the ensemble accuracy and the mean accuracy as suggested in [19,8]. For example, given two classifiers φ_1 and φ_2, the accuracy of φ_1 and φ_2 is denoted by $F(\varphi_1)$ and $F(\varphi_2)$, respectively, and its mean accuracy is calculated by $[F(\varphi_1) + F(\varphi_2)]/2$, denoted by F_M, the accuracy of ensemble φ_1 and φ_2 built by Dempster's rule is denoted by $F_{DS} = F(\varphi_1 \oplus \varphi_2)$. Thus the difference $F_{DS} - F_M$ is regarded as an improvement in accuracy over two classifiers φ_1 and φ_2. In our experiment, we have 13 classifiers generated in terms of $\varphi_1, \ldots, \varphi_{13}$. These classifiers are permuted to form 2×12 groups of classifiers in decreasing and mixed orders, which in turn make up 2×12 classifier ensembles denoted by $2C, \ldots, 13C$, respectively. For each group of classifiers in decreasing order, for instance, the diversity among the member classifiers and an improvement in accuracy over the classifier group are calculated, finally ending up with 12 pairs of diversity and improved accuracy in total. For each of the 12 pairs correlation analyses are performed over the 12 data sets. The following subsections detail our experimental results.

5.2. Combinations of classifiers using the evidential rules in decreasing order

In this section, we study the performance of 12 classifier ensembles constructed by the four combination rules in decreasing order over the 12 data sets. The purpose of this study is to investigate the impact of classifier order on ensemble performance and find the extent of the impact with the different combination rules. The experimental results are presented in Fig. 1.

From these curves, we can see that the accuracy decrease with increasing number of classifiers in the ensembles and with the change of the combination rules, respectively. Roughly the smaller the number of classifiers in the ensembles, the better the ensemble performance, and these curves converge to the combination of two classifiers – the best classifier and the second best one. The performance of the ensembles built with Dempster's rule and Smets' rule is better than the Proportion rule and Yager's rule. There is no statistical performance difference between Dempster's rule and Smets' rule, but the Proportion rule outperforms Yager's rule.

Specifically, for Dempster's rule, the curves on the ten data sets (*anneal*, *balance*, *car*, *autos*, *iris*, *letter*, *segment*, *soybean*, *wine*, *zoo*) show that the order of the classifiers has an impact on the ensemble performance, but not dramatic. They show that the maximum number of combination of classifiers results in the lower ensemble accuracy and the minimum number of classifier combination leads to the highest accuracy, and the accuracy margins between the two ends of the ensembles are very small along with a monotonic trend. When the combinations of classifiers reach 6C from 13C to 2C, the accuracy

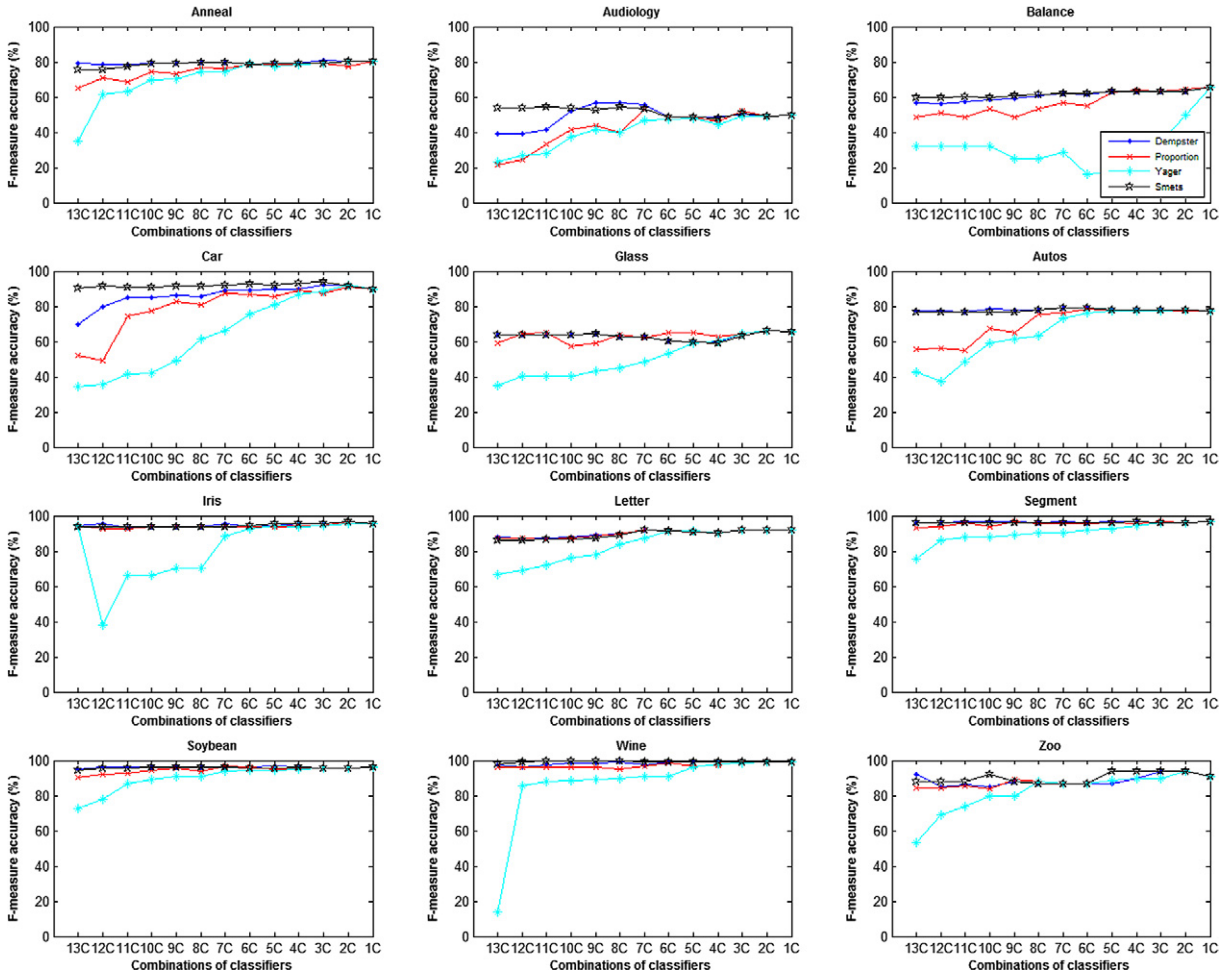

Fig. 1. Combinations of 13 classifiers in decreasing order over the 12 data sets (12 graphs share the same legend)

of the ensembles is towards flat. The similar phenomena occur to Smets' rule, but the combined accuracy is towards flat throughout the combinations of classifiers, and Smets' rule is better than Dempster's rule over the five data sets of *audiology*, *car*, *glass*, *wine*, and *zoo*.

With reference to the Proportion rule, the curves of seven data sets (*anneal*, *iris*, *letter*, *segment*, *soybean*, *wine*, *zoo*) demonstrate that this rule performs very similar to Dempster's rule and the two rules approximately fit each other when the accuracy of the ensembles exceed 85%. In the cases of *audiology*, *glass* and *autos*, although the ensemble performance contrasts with that of Dempster's rule, both of them approximately end up at the same point — the combination of two classifiers. In some cases, it appears that the Proportional rule does not favor the combination of more classifiers, such as more than six, in which the ensemble accuracy drops quickly.

Regarding Yager's rule, the performance of the ensembles is in stark contrast with that of the other three rules. Except the cases of *balance*, *autos*, *iris* and *segment*, the combination of 13 classifiers results in the lowest accuracy and the combination of the best and second best classifiers achieves the highest accuracy, and the curves of the ensemble accuracy show a monotonic trend. The average difference between the values at the two ends is about 34.11% — a large margin. This result suggests that Yager's rule might not be able to accumulate pieces of evidence derived from multiple classifiers, which is not suitable for combining multiple classifiers.

Therefore, the order of classifiers has a different impact on the ensemble performance across the 12 data sets. In general, its impact is positive on the performance of the ensemble classifiers built by Dempster's and Smets' rules, and negative on the ensemble performance obtained by the Proportion and Yager's rules.

5.3. Diversity of combinations of classifiers using the evidential rules in decreasing order

In accordance with the accuracy of the 12 classifier ensembles in decreasing order, this section studies diversity among the member classifiers of each ensemble and then assesses the relationship between the accuracy of the ensembles and diversity being inherent in the member classifiers.

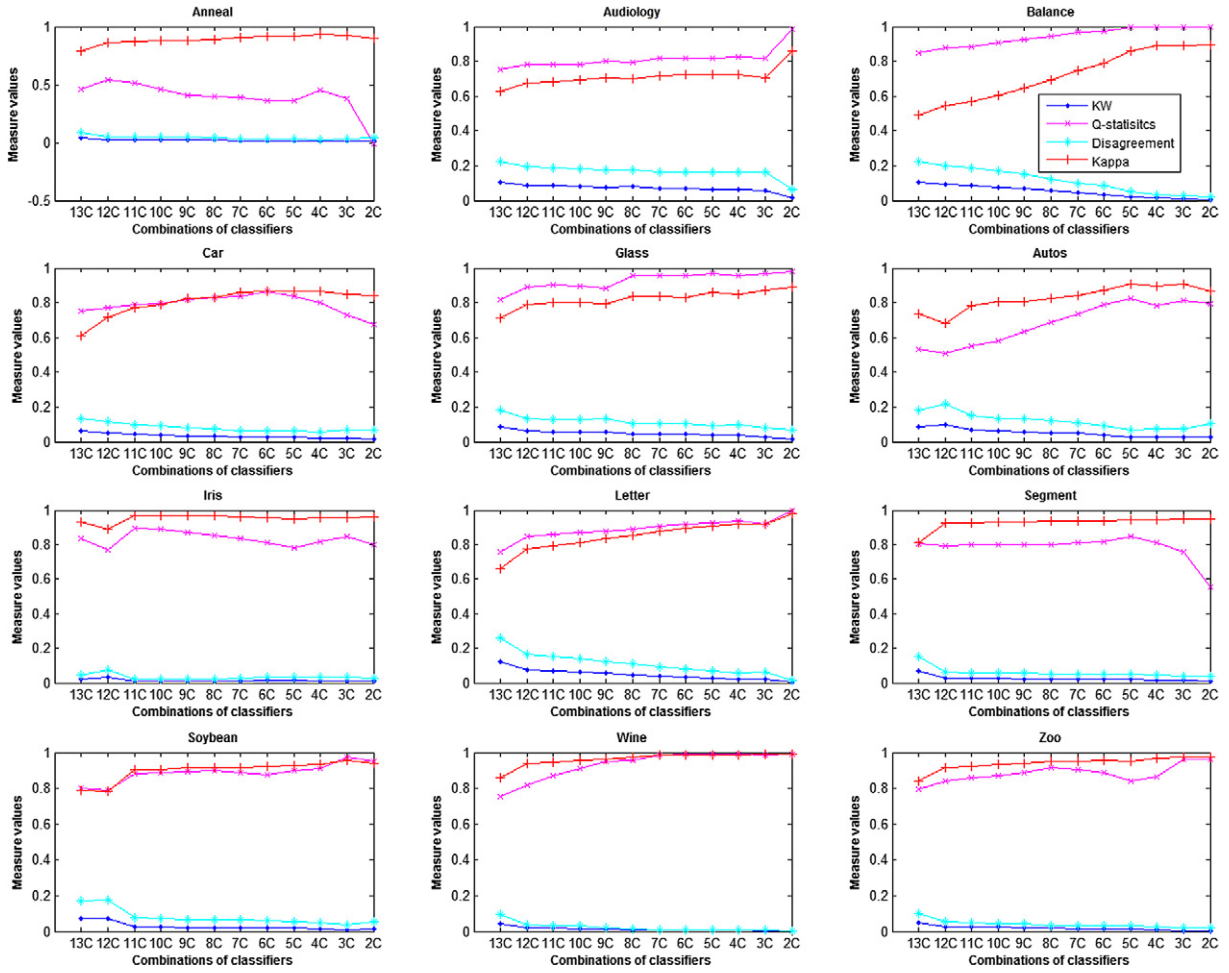

Fig. 2. Diversity of the 12 corresponding groups of classifiers in deceasing order (making up 12 classifier ensembles as seen Fig. 1) over the 12 data sets (12 graphs share the same legend).

Fig. 2 presents the curves of the diversity among the member classifiers making up the 12 classifier ensembles 2C, . . ., 13C, which are measured by *kw*, *qs*, *dis* and κ over the 12 data sets. According to the behaviors of the curves and the nature of the four diversity measures, these curves can be characterized into two groups: one is measured by *qs* and κ, and the other is measured by *kw* and *dis*. These curves show that the fitness between *kw* and *dis* is better than that between *qs* and κ, and the curve margins between *qs* and κ are larger than those between *kw* and *dis*. As introduced previously, the smaller values of *qs* and κ means that the agreement among the member classifiers is smaller, in turn representing a larger diversity, while the larger values of *kw* and *dis* represents the larger diversity among the member classifiers. As depicted in Fig. 2, the curves of the four measures are generally corresponding to the accuracy of the ensembles, when the two groups of the curves become closer to each other, the accuracy of the classifier ensembles is low. Conversely, when the accuracy of the ensembles is high, the two groups of the curves are separating from one to another, which are evident in the cases of *iris*, *letter*, *segment*, *soybean*, *wine* and *zoo*. However when the accuracy of the ensembles is lower than 63%, the curves of *qs* in the cases of *audiology*, *balance* and *glass* is above those of κ. Exceptionally the highest accuracy of the combination of two classifiers 2C corresponds to the smallest value of *qs* in the data sets of *anneal*, *car*, *iris* and *segment*, respectively. These phenomena indicate that measure *qs* is less sensitive than κ in picking up the diversity among the constituent classifiers.

Bringing together the results presented in Figs. 1, 2, and Table 3 presents the averaged accuracy on the classifier ensembles constructed by the four combination rules and the average diversity measured by *kw*, *qs*, *dis* and κ among the member classifiers on each of the data sets. The table is roughly divided into two groups, the first six data sets and the second six data sets. The accuracy in the second group is better than that in the first group. Correspondingly the diversity measured by *qs* and κ in the second group is roughly larger than that in the first group, as opposite to this, the diversity obtained by *kw* and *dis* in the second group is less than that in the first group. As mentioned previously the larger *qs* and κ values indicates the smaller diversity and the larger *kw* and *dis* values means the larger diversity. Therefore there is the same correspondence between the accuracy and the diversity as illustrated in Figs. 1 and 2, but the averages in the bottom row of the table does not indicate that the accuracy is strongly associated with the diversity.

Table 3
Average ensemble accuracy of 12 classifier ensembles and average diversity among the 12 groups of member classifiers over the 12 data sets in decreasing order.

Dataset	Dempster	Smets	Proportion	Yager	*kw*	*qs*	*dis*	κ
Anneal	79.30	78.55	74.81	70.30	0.0177	0.3921	0.0423	0.8878
Audiology	48.84	51.71	41.82	40.10	0.0706	0.8155	0.1678	0.7115
Balance	60.44	61.57	55.70	28.67	0.0502	0.9428	0.1143	0.7177
Car	86.05	91.81	78.74	62.85	0.0339	0.7917	0.0811	0.8080
Glass	62.93	62.80	62.86	49.66	0.0476	0.9271	0.1132	0.8225
Autos	77.67	77.60	69.93	64.33	0.0516	0.6859	0.1223	0.8271
Iris	94.61	94.61	94.00	76.94	0.0138	0.8342	0.0324	0.9514
Letter	89.88	89.21	89.62	82.55	0.0482	0.8925	0.1106	0.8527
Segment	96.42	96.27	95.40	89.90	0.0247	0.7826	0.0591	0.9264
Soybean	95.98	95.69	94.50	89.76	0.0283	0.8870	0.0797	0.9010
Wine	98.20	99.00	96.85	85.63	0.0131	0.9326	0.0236	0.9643
Zoo	88.36	89.94	87.57	81.61	0.0195	0.8823	0.0407	0.9399
Av	81.56	82.40	78.48	68.52	0.0349	0.8139	0.0822	0.8592

Table 4
Correlation between diversity and combined accuracy of classifiers using Dempster's rule in decreasing order.

Dataset	*kw*	*qs*	*dis*	κ
Anneal	−0.3893	−0.5307	−0.2593	0.2705
Audiology	−0.3332	0.2215	−0.3321	0.3303
Balance	**−0.9731**	**0.9858**	**−0.9720**	**0.9768**
Car	**−0.9530**	0.0664	**−0.9330**	**0.9593**
Glass	0.0253	−0.2486	0.0793	−0.0797
Autos	−0.3120	0.4156	−0.3657	0.3447
Iris	0.2170	−0.6261	0.2825	−0.2934
Letter	**−0.8324**	**0.7868**	**−0.8221**	**0.8349**
Segment	0.0975	**0.5778**	0.0312	−0.0309
Soybean	−0.4667	0.2020	−0.4647	0.4360
Wine	**−0.6079**	**0.7458**	**−0.5881**	**0.5899**
Zoo	-0.2592	0.4188	-0.0848	0.0839
Av	-0.3989	0.2512	-0.3691	0.3685
Abs(Av)	0.4556	0.4855	0.4346	0.4359

Table 5
Correlation between diversity and combined accuracy of classifiers using Proportion rule in decreasing order.

Dataset	*kw*	*qs*	*dis*	κ
Anneal	**−0.9440**	−0.4611	**−0.9373**	**0.9383**
Audiology	**−0.7191**	0.5341	**−0.6425**	**0.6260**
Balance	**−0.9407**	**0.8872**	**−0.9449**	**0.9403**
Car	**−0.9388**	0.1642	**−0.9362**	**0.9213**
Glass	**−0.6461**	**0.6588**	**−0.6569**	**0.6361**
Autos	**−0.9051**	**0.9429**	**−0.8981**	**0.8943**
Iris	−0.4702	−0.1756	−0.4196	0.4089
Letter	**−0.8278**	**0.7818**	**−0.8170**	**0.8305**
Segment	**−0.7033**	−0.1690	**−0.6818**	**0.6830**
Soybean	**−0.8357**	**0.7115**	**−0.8297**	**0.8302**
Wine	−0.5125	0.5068	−0.4691	0.4714
Zoo	**−0.7678**	**0.7064**	**−0.6823**	**0.6798**
Av	−0.7676	0.4240	−0.7429	0.7383
Abs(Av)	0.7676	0.5583	0.7429	0.7383

To assess the quantitative correlation between the accuracy and the diversity, a correlation analysis has been performed by using Spearman's rank method and the results are presented in Tables 4–7 respectively. The cell values in the tables under each measure are correlation coefficients, which are shown in *bold* provided that they are statistically significant ($p \leq 0.05$). Compared with the curves shown in Figs. 1 and 2, if the accuracy curves follow a similar trend with the diversity curves measured by *qs* and κ, the correlation coefficients under *qs* and κ are positive and the coefficients under *kw* and *dis* are negative, vice versa otherwise. In Tables 4 and 7, the correlation between the accuracy of the ensembles constructed by Dempster's and Smets' rules and the diversity obtained by the four diversity measures among their member classifiers is not very strong, where the correlation coefficients on 4–6 of the 12 data sets are statistically significant. This quantitative result supports the above affirmation made from Table 3. By contrast, the correlation coefficients in Tables 5 and 6 demonstrates a strong correlation between the accuracy of the ensembles and the diversity among the ensemble members with the exception of measure *qs*. These results reveal the fact that the diversity rated by the four measures has a varying correlation

Table 6
Correlation between diversity and combined accuracy of classifiers using Yager's rule in decreasing order.

Dataset	*kw*	*qs*	*dis*	κ
Anneal	**−0.9663**	−0.4450	**−0.9669**	**0.9664**
Audiology	**−0.7894**	**0.6132**	**−0.7080**	**0.6903**
Balance	0.0468	−0.1624	0.0501	−0.0885
Car	**−0.9280**	−0.1239	**−0.8326**	**0.7769**
Glass	**−0.9358**	**0.8463**	**−0.8899**	**0.8871**
Autos	**−0.9721**	**0.9693**	**−0.9625**	**0.9648**
Iris	**−0.6717**	−0.0985	−0.5726	0.5655
Letter	**−0.9283**	**0.8806**	**−0.9176**	**0.9285**
Segment	**−0.9598**	−0.3601	**−0.9201**	**0.9218**
Soybean	**−0.9601**	**0.8655**	**−0.9560**	**0.9575**
Wine	**−0.9427**	**0.7960**	**−0.9436**	**0.9434**
Zoo	**−0.9626**	**0.7604**	**−0.9779**	**0.9776**
Av	−0.8308	0.3785	−0.7998	0.7909
Abs(Av)	0.8386	0.5768	0.8081	0.8057

Table 7
Correlation between diversity and combined accuracy of classifiers using Smets' rule in decreasing order.

Dataset	*kw*	*qs*	*dis*	κ
Anneal	**−0.7785**	**−0.6064**	**−0.7339**	**0.7393**
Audiology	**0.6578**	−0.5308	0.5599	−0.5548
Balance	**−0.9716**	**0.9601**	**−0.9732**	**0.9759**
Car	**−0.7494**	0.1169	**−0.7600**	**0.7269**
Glass	0.1359	−0.3306	0.1793	−0.1808
Autos	−0.4799	**0.6363**	−0.5072	0.4972
Iris	−0.2051	−0.4485	−0.1187	0.1067
Letter	**−0.8606**	**0.8159**	**−0.8498**	**0.8627**
Segment	0.3355	0.4240	0.3325	−0.3283
Soybean	**−0.6389**	0.4533	**−0.6307**	**0.6174**
Wine	−0.5388	0.4075	−0.5510	0.5498
Zoo	−0.5174	0.2645	−0.4212	0.4300
Av	−0.3843	0.1802	−0.3728	0.3702
Abs(Av)	0.5725	0.4996	0.5514	0.5475

with the ensemble performance obtained by the four combination rules, i.e. the higher ensemble accuracy corresponds to the lower diversity, and the larger diversity corresponds to the lower ensemble accuracy.

To investigate the relationship between the diversity and the difference between the ensemble accuracy and the mean accuracy of its member classifiers in terms of the improved accuracy over the ensemble member classifiers, a further correlation analysis has been carried out and the resulting coefficients are presented in Tables 8–11 respectively. In Tables 8 and 11, the diversity obtained by *kw*, *dis* and κ are negatively and strongly correlated with the improved accuracy, where the correlation coefficients on 9–11 of the 12 data sets are statistically significant. The relationship quantified in this result appear to be opposite to the positive correlation illustrated in Tables 4 and 7. In Table 9, the correlation between the improved accuracy and the diversity detected by the four measures change from negative to positive over the data sets, where such change seems to correlate with the ensemble accuracy that is lower or higher than 85%. In Table 10, the improved accuracy is strongly correlated with the diversity obtained by *kw*, *dis* and κ, where the correlation coefficients on 10–11 data sets are statistically significant. The diversity measured by *qs* is weakly correlated to the improved accuracy achieved by the four combination rules as shown in the three tables. This phenomenon is similar to the case as shown in Tables 4–7 respectively.

As seen from Tables 8–11, there are strong negative correlation between the diversity and the improved accuracy. As previously stated the negative correlation means that either with the diversity increasing the improved accuracy decreases or the diversity decreases as the ensemble performance increasing. From the latter experiment we also develop an understanding that the ensemble accuracy obtained by Yager's rule is lower than the average accuracy of ensemble classifiers, thereby resulting in the positive correlation between the diversity and the improved accuracy. Nevertheless the findings drawn from the two correlation assessments are basically consistent.

5.4. Combinations of classifiers using the evidential rules without ordering

In this experiment, we evaluate the performance of 12 ensemble classifiers constructed by the four combination rules in mixed order. The experimental process and setting are the same as those in Section 5.2. The experimental results over the 12 data sets are graphed in Fig. 3.

As seen in Fig. 3 the performance curves of the ensemble classifiers made by Dempster's rule, Smets' rule and the Proportion rule fit each other and they are gradually separating from the curves of the ensembles built by Yager's rule as more classifiers are combined. The former three curves follow a similar pattern with the exception of *audiology*, *balance* and *car* data sets, i.e. the accuracy increases with more classifiers being combined particularly in the cases of *anneal*, *audiology*,

Table 8
Correlation between diversity and improved accuracy of the combined classifiers using Dempster's rule in decreasing order.

Dataset	*kw*	*qs*	*dis*	κ
Anneal	**0.9547**	0.4237	**0.9598**	**−0.9581**
Audiology	0.1784	−0.1935	0.1169	−0.1081
Balance	**0.7881**	**−0.6983**	**0.7879**	**−0.7697**
Car	−0.2476	0.1654	−0.3196	0.4409
Glass	**0.6321**	**−0.7698**	**0.6738**	**−0.6801**
Autos	**0.9896**	**−0.9532**	**0.9708**	**−0.9763**
Iris	**0.9137**	**−0.6328**	**0.9044**	**−0.9068**
Letter	**0.8933**	**−0.8774**	**0.8949**	**−0.8850**
Segment	**0.9946**	0.3089	**0.9884**	**−0.9883**
Soybean	**0.9934**	**−0.9253**	**0.9913**	**−0.9946**
Wine	**0.6422**	−0.4299	**0.6625**	**−0.6608**
Zoo	0.3940	−0.1094	0.5491	−0.5497
Av	0.6772	−0.3910	0.6817	−0.6697
Abs(Av)	0.7185	0.5406	0.7349	0.7432

Table 9
Correlation between diversity and improved accuracy of the combined classifiers using Proportion rule in decreasing order.

Dataset	*kw*	*qs*	*dis*	κ
Anneal	**−0.7903**	−0.3275	**−0.8100**	**0.8103**
Audiology	**−0.6263**	0.4463	−0.5589	0.5427
Balance	**−0.7082**	**0.6273**	**−0.7176**	**0.7120**
Car	**−0.8348**	0.2423	**−0.8608**	**0.8477**
Glass	−0.0132	0.0508	−0.0211	−0.0110
Autos	**−0.8010**	**0.8759**	**−0.7960**	**0.7885**
Iris	**0.6973**	**−0.6069**	**0.6766**	**−0.6824**
Letter	**0.8491**	**−0.8377**	**0.8523**	**−0.8399**
Segment	**0.8138**	0.2093	**0.8325**	**−0.8314**
Soybean	**0.8956**	**−0.8363**	**0.8973**	**−0.8948**
Wine	0.4420	−0.3791	0.4840	−0.4819
Zoo	0.1136	0.0927	0.2502	−0.2534
Av	0.0031	−0.0369	0.0190	−0.0245
Abs(Av)	0.6321	0.4610	0.6464	0.6413

Table 10
Correlation between diversity and improved accuracy of the combined classifiers using Yager's rule in decreasing order.

Dataset	*kw*	*qs*	*dis*	κ
Anneal	**−0.9559**	−0.4198	**−0.9644**	**0.9634**
Audiology	**−0.7319**	0.5624	**−0.6572**	**0.6389**
Balance	0.4149	−0.5140	0.4176	−0.4523
Car	**−0.8850**	−0.1656	**−0.7771**	**0.7097**
Glass	**−0.9048**	**0.8052**	**−0.8494**	**0.8442**
Autos	**−0.9647**	**0.9677**	**−0.9550**	**0.9575**
Iris	**−0.6504**	−0.1239	−0.5493	0.5421
Letter	**−0.8256**	**0.7712**	**−0.8112**	**0.8289**
Segment	**−0.8886**	−0.4049	**−0.8276**	**0.8303**
Soybean	**−0.8877**	**0.8044**	**−0.8823**	**0.8860**
Wine	**−0.9382**	**0.7876**	**−0.9391**	**0.9389**
Zoo	**−0.9527**	**0.7648**	**−0.9660**	**0.9655**
Av	−0.7642	0.3196	−0.7301	0.7211
Abs(Av)	0.8334	0.5910	0.7997	0.7965

balance, *car*, *glass* and *autos*, where the accuracy of the ensemble classifiers is less than 85%, otherwise the accuracy curves go towards flat. However the accuracy curves of the ensemble classifiers by Yager's rule decreases with fluctuations when more classifiers are combined except the data set of *audiology*. In fact some fluctuations also appear on the curves of the Proportion rule such as in the cases of *balance*, *car*, *glass* and *autos*.

Compared with the accuracy curves in Fig. 1, the accuracy curves with Dempster's rule, Semets's rule, and the Proportional rule have an opposite trend to those in decreasing order, the accuracy of the ensemble classifiers increases in mixed order as more classifiers are combined, whereas the accuracy of the combined classifiers decreases in decreasing order. However when the ensemble accuracy exceeds 85%, the accuracy curves go towards flat in both decreasing and mixed orders — the two groups of the curves appear to be close each other, such as in the cases of *iris*, *letter*, *segment*, *soybean*, *wine* and *zoo*. In fact the ensemble accuracy curves in the two orders are different on case by case, but the average accuracy over the 12 data sets in decreasing order is 1.10% better than that in mixed order for Dempster's rule, 1.78% for Smets' rule and 0.12% for

Table 11
Correlation between diversity and improved accuracy of the combined classifiers using Smets' rule in decreasing order.

Dataset	*kw*	*qs*	*dis*	κ
Anneal	**0.8586**	0.2798	**0.8492**	**−0.8463**
Audiology	**0.8244**	**−0.6561**	**0.7249**	**−0.7128**
Balance	**0.9720**	**−0.9437**	**0.9697**	**−0.9634**
Car	**0.9906**	0.0109	**0.9299**	**−0.9186**
Glass	**0.6625**	**−0.7780**	**0.6958**	**−0.7023**
Autos	**0.9629**	**−0.9142**	**0.9502**	**−0.9527**
Iris	**0.8019**	−0.4794	**0.7849**	**−0.7861**
Letter	**0.7862**	**−0.7792**	**0.7915**	**−0.7762**
Segment	**0.9933**	0.2778	**0.9941**	**−0.9936**
Soybean	**0.9914**	**−0.9109**	**0.9905**	**−0.9916**
Wine	**0.9719**	**−0.9403**	**0.9690**	**−0.9695**
Zoo	0.2699	−0.3258	0.3898	−0.3806
Av	0.8405	−0.5133	0.8366	−0.8328
Abs(Av)	0.8405	0.6080	0.8366	0.8328

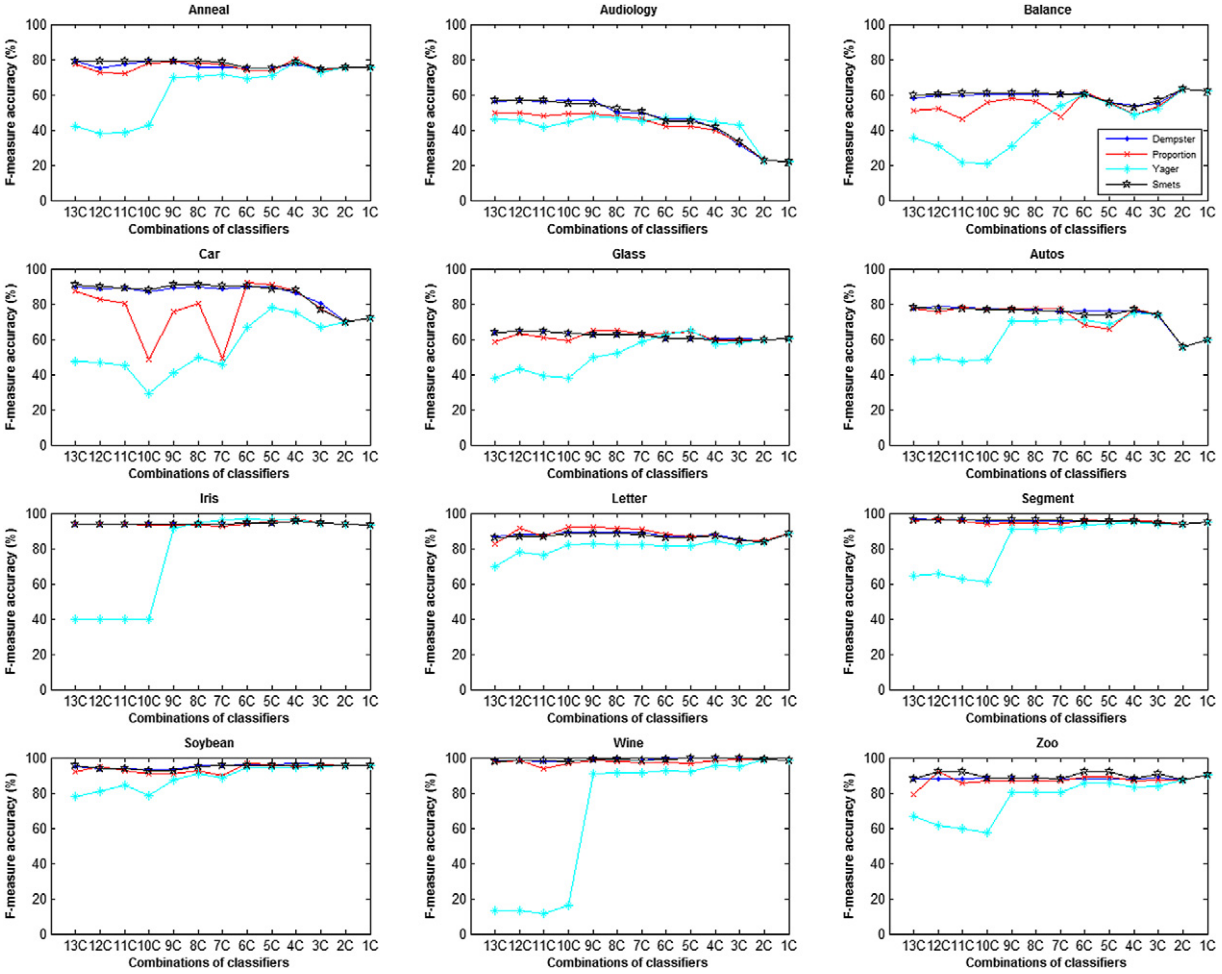

Fig. 3. Combinations of 13 classifiers over the 12 data sets without ordering (12 graphs share the same legend).

Proportion rule. For Yager's rule, the accuracy curves of ensemble classifiers in both the orders are approximately consistent as shown in Figs. 1 and 2, as more classifiers are combined, the accuracy of the ensemble classifiers goes down even though there are fluctuations in *audiology* and *autos* during the course of combination. The average accuracy over the 12 data sets in decreasing order is 2.79% better than that in mixed order. This result indicates that the order of classifiers combined by Yager's rule has the largest impact on the ensemble accuracy.

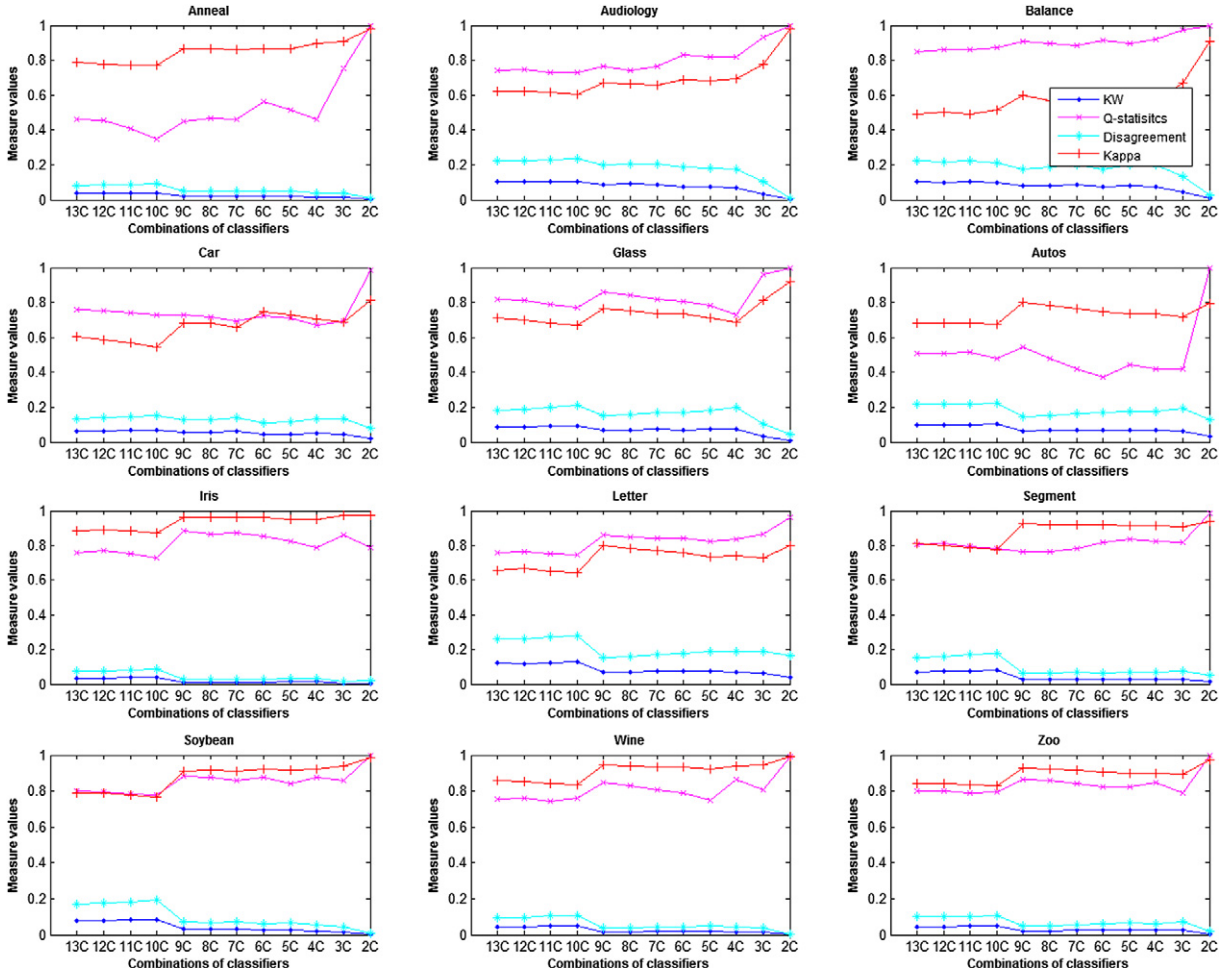

Fig. 4. Diversity of the corresponding combinations of 13 classifiers in mixed order over the 12 data sets (12 graphs share the same legend).

Table 12
Average ensemble accuracy and average diversity over the 12 data sets in mixed order.

Dataset	Dempster	Smets	Proportion	Yager	*kw*	*qs*	*dis*	κ
Anneal	76.50	77.69	75.93	61.66	0.0247	0.5286	0.0574	0.8514
Audiology	47.31	47.53	43.24	43.37	0.0781	0.8013	0.1819	0.6905
Balance	58.93	59.42	53.97	43.01	0.0771	0.9022	0.1814	0.5821
Car	86.44	86.88	76.78	55.12	0.0538	0.7418	0.1294	0.6679
Glass	62.13	61.86	61.71	51.78	0.0692	0.8322	0.1635	0.7400
Autos	74.91	74.33	73.36	62.33	0.0755	0.5089	0.1821	0.7330
Iris	94.28	94.28	94.11	76.61	0.0189	0.8103	0.0439	0.9344
Letter	87.52	86.64	88.08	80.45	0.0854	0.8239	0.2053	0.7265
Segment	95.57	95.60	95.06	82.99	0.0418	0.8145	0.0980	0.8775
Soybean	95.02	94.85	93.71	88.52	0.0425	0.8519	0.0972	0.8789
Wine	98.79	98.71	97.47	66.89	0.0253	0.8084	0.0584	0.9119
Zoo	88.13	89.62	86.94	76.04	0.0298	0.8362	0.0702	0.8910
Av	80.46	80.62	78.36	65.73	0.0518	0.7717	0.1224	0.7904

5.5. Diversity of combinations of classifiers using the evidential rules without ordering

Fig. 4 presents the diversity among the different groups of classifiers that make up the classifier ensembles in mixed order over the 12 data sets, where the diversity curves demonstrate some similarities to those in decreasing order and can hereby be grouped in the same way: one is measured by *qs* and κ, and the other is measured by *kw* and *dis*. We can see that the first group of curves decrease with more classifiers being added, while the second group of curves increases as more

Table 13
Correlation between diversity and combined accuracy of classifiers using Dempster's rule in mixed order.

Dataset	*kw*	*qs*	*dis*	κ
Anneal	0.4684	−0.4861	0.4442	−0.4444
Audiology	**0.9818**	**−0.9682**	**0.9573**	**−0.9311**
Balance	−0.1789	0.0420	−0.3588	0.3527
Car	**0.8003**	**−0.7737**	**0.6616**	−0.5170
Glass	**0.6870**	−0.3240	0.5365	−0.5362
Autos	**0.7708**	**−0.8755**	**0.5998**	−0.4814
Iris	−0.3914	0.1292	−0.3617	0.3600
Letter	0.4459	−0.5528	0.1762	−0.1735
Segment	**0.6408**	**−0.7968**	0.5554	−0.5566
Soybean	**−0.6776**	0.4193	**−0.6621**	**0.6491**
Wine	**−0.7616**	0.3411	**−0.7278**	**0.7311**
Zoo	0.3690	**−0.7391**	0.4162	−0.4070
Av	0.2626	−0.3818	0.1862	−0.1626
Abs (Av)	0.5977	0.5374	0.5381	0.5116

Table 14
Correlation between diversity and combined accuracy of classifiers using Proportion rule in mixed order.

Dataset	*kw*	*qs*	*dis*	κ
Anneal	−0.1848	−0.2054	−0.2009	0.1957
Audiology	**0.9829**	**−0.9817**	**0.9763**	**−0.9573**
Balance	**−0.6116**	0.5445	**−0.6683**	**0.6709**
Car	−0.1719	−0.0987	−0.2315	0.2923
Glass	0.1254	−0.1142	0.0955	−0.0423
Autos	**0.6627**	**−0.6714**	0.5176	−0.3946
Iris	−0.1636	−0.2032	−0.1170	0.1160
Letter	0.1433	−0.2125	−0.0673	0.0990
Segment	0.2584	−0.0963	0.2551	−0.2601
Soybean	−0.4151	0.3030	−0.3996	0.3923
Wine	**−0.6190**	**0.6097**	**−0.6207**	**0.6219**
Zoo	−0.1843	0.0633	−0.1644	0.1731
Av	−0.0148	−0.0886	−0.0521	0.0756
Abs (Av)	0.3769	0.3420	0.3595	0.3513

Table 15
Correlation between diversity and combined accuracy of classifiers using Yager's rule in mixed order.

Dataset	*kw*	*qs*	*dis*	κ
Anneal	**−0.9185**	0.4967	**−0.9060**	**0.9001**
Audiology	**0.7611**	**−0.7393**	**0.8333**	**−0.8703**
Balance	**−0.6946**	**0.6926**	**−0.6098**	**0.6170**
Car	**−0.7419**	0.1147	**−0.6224**	**0.7832**
Glass	**−0.5793**	0.2472	−0.4455	0.4663
Autos	**−0.5999**	−0.3686	**−0.6087**	**0.6310**
Iris	**−0.9726**	**0.8068**	**−0.9728**	**0.9720**
Letter	**−0.7122**	**0.6382**	**−0.6783**	**0.6848**
Segment	**−0.9860**	0.2629	**−0.9856**	**0.9847**
Soybean	**−0.9353**	**0.7403**	**−0.9250**	**0.9205**
Wine	**−0.9638**	**0.5996**	**−0.9517**	**0.9527**
Zoo	**−0.9172**	0.5284	**−0.8701**	**0.8691**
Av	−0.6884	0.3350	−0.6452	0.6592
Abs (Av)	0.8152	0.5196	0.7841	0.8043

classifiers are combined but both the groups have some fluctuations. In particular, there are sharp changes between 9*C* and 10*C* in the cases of *glass*, *autos*, *iris*, *letter*, *segment*, *soybean*, *wine* and *zoo*. We also notice that these changes appear to be correlated with the accuracy of the ensembles over all the data sets in sense that the accuracy of the ensemble classifiers made by Dempster's and Smets' rules, and the Proportion rule exceeds 85%. Meanwhile the first group of the curves show a similar phenomenon as in decreasing order that the curves of the diversity measured by *qs* are above those obtained by κ in the cases of *audiology*, *balance*, *car*, *glass*, *letter* and *heart*, provided that the accuracy of the ensemble classifiers is less than 65%. Compared with the curves presented in Fig. 2, the diversity curves have more fluctuations than those in decreasing order and the impact of the order on the diversity is not as apparent as in decreasing order. In other words, the correspondence between the diversity and the accuracy in decreasing order appears to be more apparent than that in mixed order.

Table 16
Correlation between diversity and combined accuracy of classifiers using Smets' rule in mixed order.

Dataset	*kw*	*qs*	*dis*	κ
Anneal	**0.6609**	**−0.7153**	**0.6268**	**−0.6298**
Audiology	**0.9876**	**−0.9811**	**0.9652**	**−0.9412**
Balance	−0.0777	−0.0666	−0.2737	0.2582
Car	**0.8028**	**−0.7053**	**0.6315**	−0.5211
Glass	**0.7229**	−0.3897	0.5690	−0.5526
Autos	**0.7592**	**−0.8595**	**0.5998**	−0.4785
Iris	−0.3914	0.1292	−0.3618	0.3600
Letter	0.3382	−0.4526	0.0766	−0.0716
Segment	**0.6275**	**−0.8212**	0.5375	−0.5375
Soybean	**−0.5995**	0.3914	**−0.5926**	**0.5834**
Wine	**−0.7593**	0.3795	**−0.7172**	**0.7202**
Zoo	0.3618	−0.5530	0.4125	−0.3984
Av	0.2861	−0.3870	0.2061	−0.1841
Abs (Av)	0.590735	0.537035	0.530325	0.5044

Table 17
Correlation between diversity and improved accuracy of the combined classifiers using Dempster's rule in mixed order.

Dataset	*kw*	*qs*	*dis*	κ
Anneal	**0.8880**	**−0.5945**	**0.8726**	**−0.8680**
Audiology	**0.9647**	**−0.9371**	**0.9415**	**−0.9235**
Balance	**0.6448**	**−0.7417**	0.4786	−0.4864
Car	**0.9805**	**−0.5951**	**0.8952**	**−0.9123**
Glass	**0.6071**	−0.2983	0.4980	−0.5331
Autos	**0.8243**	**−0.7748**	**0.8695**	**0.8268**
Iris	**0.9504**	**−0.8902**	**0.9662**	**−0.9655**
Letter	**0.9571**	**−0.9106**	**0.9674**	**−0.9611**
Segment	**0.9951**	−0.2869	**0.9976**	**−0.9974**
Soybean	**0.9722**	**−0.8225**	**0.9740**	**−0.9743**
Wine	**0.9656**	**−0.8162**	**0.9765**	**−0.9757**
Zoo	**0.9881**	**−0.7554**	**0.9866**	**−0.9855**
Av	0.8948	−0.7019	0.8686	−0.8675
Abs (Av)	0.8948	0.7019	0.8686	0.8675

Table 18
Correlation between diversity and improved accuracy of the combined classifiers using Proportion rule in mixed order.

Dataset	*kw*	*qs*	*dis*	κ
Anneal	**0.6487**	**−0.6401**	**0.6291**	**−0.6281**
Audiology	**0.9658**	**−0.9548**	**0.9819**	**−0.9850**
Balance	−0.1930	0.12232	−0.27412	0.2756
Car	−0.1722	0.01253	−0.2073	0.2023
Glass	0.4393	−0.3069	0.3783	−0.35812
Autos	**0.60242**	−0.4374	**0.6558**	**−0.6006**
Iris	**0.8249**	**−0.9211**	**0.8583**	**−0.8579**
Letter	**0.7567**	**−0.7256**	**0.7068**	**−0.6734**
Segment	**0.9400**	−0.1434	**0.9602**	**−0.9612**
Soybean	**0.8448**	**−0.6960**	**0.8502**	**−0.8506**
Wine	**0.678**	−0.3677	**0.6734**	**−0.6727**
Zoo	**0.5857**	−0.4695	**0.5965**	**−0.5893**
Av	0.5767	−0.4606	0.5674	−0.5582
Abs (Av)	0.6376	0.4831	0.6476	0.6379

Table 12 presents the average accuracy on the classifier ensembles and the average diversity among the corresponding groups of member classifiers on each of the data sets. As in Table 3, the data sets can also be divided into two groups and the results confirm that there is an correspondence between the accuracy and the diversity. Compared with the averaged accuracy and diversity in Table 3, an increase in diversity corresponds to a decrease in accuracy. For example, the average diversity, 0.0518, 0.7717, 0.1224 and 0.7904 in bottom row obtained by the four measures within Table 12, are greater than 0.0349, 0.8139, 0.0822 and 0.8592 in Table 3, respectively, but the average accuracy, 80.46%, 78.36% and 65.73% in Table 12, are correspondingly smaller than 81.56%, 78.48% and 68.52% in Table 3.

Tables 13–16 present the correlation coefficients quantifying the relationship between the diversity under the four measures and the accuracy of the ensemble classifiers made by the four combination rules. For Dempster's rule, although the negative correlation between the ensemble accuracy and the diversity among their member classifiers is stronger than the

Table 19
Correlation between diversity and improved accuracy of the combined classifiers using Yager's rule in mixed order.

Dataset	*kw*	*qs*	*dis*	κ
Anneal	**−0.9168**	0.4958	**−0.9030**	**0.8971**
Audiology	0.1683	−0.1274	0.3053	−0.4004
Balance	**−0.5932**	**0.5965**	−0.4992	0.5068
Car	**−0.7830**	0.2339	**−0.6274**	**0.7343**
Glass	**−0.5915**	0.2465	−0.4468	0.4587
Autos	**−0.7582**	−0.1496	**−0.6555**	**0.6264**
Iris	**−0.9733**	**0.8044**	**−0.9730**	**0.9723**
Letter	−0.1996	0.2104	−0.0063	0.0187
Segment	**−0.9856**	0.2986	**−0.9785**	**0.9776**
Soybean	−0.5627	0.4242	−0.54283	0.5375
Wine	**−0.9607**	**0.5917**	**−0.9482**	**0.9492**
Zoo	**−0.8764**	0.4643	**−0.8179**	**0.8165**
Av	−0.6694	0.3408	−0.5911	0.5912
Abs (Av)	0.6974	0.3869	0.6420	0.6579

Table 20
Correlation between diversity and improved accuracy of the combined classifiers using Smets' rule in mixed order.

Dataset	*kw*	*qs*	*dis*	κ
Anneal	**0.9456**	**−0.6931**	**0.9249**	**−0.9222**
Audiology	**0.9832**	**−0.9669**	**0.9633**	**−0.9492**
Balance	**0.6528**	**−0.7514**	0.4834	−0.4984
Car	**0.9640**	−0.5379	**0.8343**	**−0.8721**
Glass	**0.6454**	−0.3491	0.5306	−0.5545
Autos	**0.7944**	**−0.7330**	**0.8661**	**−0.8223**
Iris	**0.9505**	**−0.8903**	**0.9663**	**−0.9656**
Letter	**0.9309**	**−0.8742**	**0.9725**	**−0.9648**
Segment	**0.9961**	−0.3001	**0.9967**	**−0.9962**
Soybean	**0.9691**	**−0.8062**	**0.9681**	**−0.9671**
Wine	**0.9634**	**−0.7924**	**0.9766**	**−0.9760**
Zoo	**0.8848**	**−0.7760**	**0.9051**	**−0.8980**
Av	0.8900	−0.7059	0.8657	−0.8655
Abs (Av)	0.8900	0.7059	0.8657	0.8655

positive correlation, the entire correlation appears not to be strong since there are 3–7 data sets confirming that their correlation coefficients are statistically significant. The correlation with Smets' rule is slightly stronger than that of Dempster's rule since the correlation on 4–8 data sets are statistically significant. With respect to the Proportion rule, the correlation is even weaker as the correlation coefficients on only three data sets show $p \leq 0.05$ such that the hypothesis of no correlation is nearly true. For Yager' rule, the correlation between the accuracy of the ensembles and the diversity among their member classifiers is strong with exception of measure *qs* where 6 of the 12 data sets show that their coefficients are statistically significant, but for all the other measures, the correlation coefficients on 11–12 data sets are statistically significant. As witnessed in Fig. 3 and Table 12, the ensemble accuracy obtained by Yager's rule is the lowest, it is actually even lower than the average accuracy of the member classifiers. Thus the strong correlation means that the deteriorated ensemble accuracy could be caused by the influence of the larger diversity.

Tables 17–20 present the results of the correlation between the diversity and the improved accuracy of the groups of classifiers. For Dempster's and Smets' rules, the correlation is negatively strong since the correlation coefficients over 9–12 of the 12 data sets are statistically significant. For Proportion rule, the correlation under *qs* is weaker but the correlation between the improved accuracy and the diversity is negatively strong because 9 of the 12 cases demonstrates that the correlation coefficients are statistically significant. For Yager's rule, with the different measures, the correlation between the diversity and the accuracy is positive and stronger than that of the other two combination rules. Compared with Tables 8–11, the correlation between the diversity and accuracy are stronger than those in decrease order. It is also important to note that an improvement in the average accuracy in mixed order are 5.52% for Dempster's rule, 5.60% for Smets' rule, 3.49% for the Proportion rule and −8.37% for Yager's rule, which are correspondingly greater than 1.89%, 2.67%, −1.18% and −10.4% in decrease order although the average accuracy of all the member classifiers in decrease order is better than that in mixed order. The comparative analysis reveals the fact that the larger improvement in the average accuracy of all the member classifiers could imply the poorer ensemble accuracy, reflecting the negatively strong correlation between the diversity and the improved accuracy. This fact in turn implies that an increase in diversity corresponds to a decrease in accuracy, or vice versa, which is consistent with the above finding drawn from Tables 8–11.

6. Conclusion

Aiming to ascertain the impact of diversity on ensemble accuracy, we have investigated two possible orders to build classifier ensembles, assessed the diversity among the member classifiers, and quantified the correlation between the

Table 21
Decreasing order: summary of correlation between diversity and ensemble accuracy along with correlation between diversity and improved accuracy (↑: positive correlation; ⇑: strongly positive correlation; ↓: negative correlation; ⇓: strongly negative correlation; ↕: neutral correlation).

	Ensemble accuracy				Improved accuracy			
	kw	*qs*	*dis*	κ	*kw*	*qs*	*dis*	κ
Dempster's rule	↑	↑	↑	↑	⇓	↓	⇓	⇓
Smets' rule	↑	↑	↑	↑	⇓	↓	⇓	⇓
Proportion rule	⇑	↑	⇑	⇑	↕	↕	↕	↕
Yager's rule	⇑	↑	⇑	⇑	⇑	↑	⇑	⇑

Table 22
Mixed order: summary of correlation between diversity and ensemble accuracy along with correlation between diversity and improved accuracy (↑: positive correlation; ⇑: strongly positive correlation; ↓: negative correlation; ⇓: strongly negative correlation; ↕: neutral correlation).

	Ensemble accuracy				Improved accuracy			
	kw	*qs*	*dis*	κ	*kw*	*qs*	*dis*	κ
Dempster's rule	↓	↓	↓	↓	⇓	⇓	⇓	⇓
Smets' rule	↓	↓	↓	↓	⇓	⇓	⇓	⇓
Proportion rule	↑	↓	↑	↑	↓	↓	↓	↓
Yager's rule	⇑	↑	⇑	⇑	⇑	↑	⇑	⇑

diversity and ensemble accuracy. The empirical results show that the increase in diversity makes the ensemble accuracy decrease or vice versa. In a sense, our finding is consistent to the findings reported in [8,12], i.e. an increase in diversity does not consistently correspond to an improvement in the ensemble accuracy. However in a more general sense, our observation appears to support the conjecture that the larger the conflict between two evidence sources, the larger the counterintuitive effects produced, which is commonly believed in the research of the belief functions theory.

The experiments with the two orders of classifier combination have an important role to play in the study of the diversity and the ensemble accuracy. Such orders allow us to vary the combination rules to examine the effects of the decreasing and mixed orders on the classifier ensembles constructed, while the diversity among the member classifiers remains constant. In this way, we have analyzed the ensemble performance made by the four combination rules in the two orders along with their effects, the results show that the order of classifiers is an important factor affecting the ensemble performance and decreasing order would be recommended as a better way to construct classifier ensembles.

Specifically Table 21 presents a summary of the correlation in decreasing order, where the left column is the correlation between the diversity and the ensemble accuracy, and the right column summarizes the correlation between the diversity and the improved accuracy. The former shows the correlation change from weak to strong with the change of the combination methods from Dempster's and Smets' rules to the Proportion and Yager's rules as occurred in all the diversity measures with the exception of *qs*. The enhancement of the correlation with the change of the combination rules indicates that the ensemble accuracy obtained by the Proportion and Yager's rules is more dependent with the diversity than Dempster's and Smets' rules, however the accuracy of the ensembles made by the Proportion and Yager's rules is worse than that obtained by Dempster's and Smets' rules. This suggests that the best ensemble accuracy appears not to depend on the diversity. On the other hand, the latter shows the correlation is varied from negative to positive with the change of the combination methods from Dempster's and Smets' rules to the Proportion and Yager's rules except *qs* as well. The negative correlation means that the improved accuracy increases as the diversity decreases, instead, an increase in the diversity leads to an increase in the improved accuracy. This observation is similar to that made in [35].

Table 22 summarizes the correlation in mixed order. The left column shows the correlation change from negative to positive with the change of the combination methods from Dempster's and Smets' rules to the Proportion and Yager's rules for all the diversity measures except *qs*. This change highlights the major difference with the result in Table 21, which is further confirmed by the paired *t*-test results obtained from Tables 8–11 and 17–20. As discussed previously the negative correlation indicates that the behavior of the ensemble accuracy is opposite to that of the diversity among the member classifiers, which is not in favor of the claim that increasing diversity could lead to reduction of generalization error of classifier ensembles. Therefore the experimental results provide an insight into what role the diversity plays in improving the ensemble performance of classifiers, and they reinforce our belief that increasing diversity is not a good factor to generalize the performance of evidential ensemble classifiers. With respect to *qs*, both of complete agreement and disagreement are quantified on correct/incorrect assignments of classes made by classifiers, and *qs* is the radio of the difference with the sum of these quantities. But this ratio values may not have the ability to appropriately capturing classifier diversity. As a consequence, the *qs* diversity could not well correlate with the ensemble accuracy. This result seems to be consistent to that reported in [8].

The effect of normalization and unnormalization with the evidential rules has been empirically examined. The different ways of handling the conflicting factor E results in the different performance of the evidential ensembles. With and without the influence of the orders of classifiers, the ensembles constructed by Dempster's and Smets' rules outperform these made by Yager's rule and the Proportion rule, where Smets' rule performs better. Looking at the nature of the first three rules, Dempster's rule eliminates E by redistributing it as a normalization factor and Yager's rule merges E with the mass

committed to the whole frame of classes, however, Smets' rule retains E with the emptyset, keeping the intersected focal classes in the original order resulted from the combination of evidence. Conventionally, researchers employed the whole frame to represent uncertainty and the emptyset to represent conflict, in which they characterize different aspects of evidence sources. Our empirical results recommend that keeping the separation between them is better than their amalgamation in building the ensemble classifiers. On the other hand, the normalization by redistributing E means the exclusion of the emptyset in the iteration of evidence accumulation. In some sense, the normalization may diverge the intersected elements towards reaching a consensus of focal elements originally supported by the evidence sources, consequently leading to more accumulated disagreement in the process of evidence combination and the performance deterioration of the ensemble classifiers constructed by Dempster's rule.

Although this study has not addressed the issue of how the dependence of member classifiers affects the ensemble performance, particularly generated by the cautious rule [26], a similar issue is addressed by other researchers in [27]. In opposition to our standing aspect, those authors suggest the classifiers generated by different learning algorithms on the same data sets "cannot be considered as independent sources of information". They proposed to automatically adapt the level of dependence between the classifiers by optimizing the combination rule instead of quantifying the level of dependence among the member classifiers. Their empirical results show the effectiveness of the proposed scheme for learning an optimized rule that often provides better results than any of the fixed rules investigated, including Dempster's rule and Denoeux's cautious rule. This issue along with comprehensive comparisons of the cautious rule with other combination rules remain to be addressed in our future study.

In the blend of the DS theory and ensemble learning, amalgamating diversity with conflict will require more sophisticated methods for measuring either diversity or conflict being inherent in classifiers. Although the experimental results provide an insight into what role diversity plays in improving the performance of evidential ensembles, such statistic measures used in the form of binary outputs cannot offer an effective way to ascertain the what role a counter intuitive effect caused by conflicting evidence plays in constructing successful classifier ensembles by an evidential approach. In general, we need to devise a framework for uniformly formulating diversity and conflict, develop measures for capturing diversity and evaluating its usefulness, and design a better mechanism that will be used to build successful evidential classifier ensembles without scarifying accuracy and efficiency. These research issues are to be addressed in a future paper.

Acknowledgements

The author would like to thank four anonymous reviewers for their comments which have helped the author improve the manuscript.

References

[1] P.N. Bennett, S.T. Dumais, E. Horvitz, The combination of text classifiers using reliability indicators, Journal Information Retrieval 8 (1) (2005) 67–100.
[2] P.A. Gislason, J.A. Benediktsson, J.R. Sveinsson, Decision fusion for the classification of urban remote sensing images, Pattern Recognition Letters 27 (2006) 294–300.
[3] J. Suutala, J. Roning, Methods for person identification on a pressure-sensitive floor: experiments with multiple classifiers and reject option, Information Fusion 9 (1) (2008) 21–40.
[4] P. Bonissone, J.M. Cadenas, M.C. Garrido, R. Andrés Díaz-Valladares, A fuzzy random forest, International Journal of Approximate Reasoning 51 (7) (2010) 729–747.
[5] P. Monney, M. Chan, P. Romberg. A belief function classifier based on information provided by noisy and dependent features, International Journal of Approximate Reasoning 52 (3) (2011) 335–352.
[6] G. Brown, L.I. Kuncheva, Good and bad diversity in majority vote ensembles, in: International Workshop on Multiple Classifier Systems, 2010.
[7] Z.-H. Zhou, N. Li, Multi-information ensemble diversity, in: International Workshop on Multiple Classifier Systems, 2010.
[8] L. Kuncheva, C.J. Whitaker, Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy, Machine Learning 51 (2003) 181–207.
[9] G. Shafer, A Mathematical Theory of Evidence, Princeton University Press, Princeton, NJ, 1976.
[10] A. Krogh, J. Vedelsby, Neural network ensembles, cross validation, and active learning, in: G. Tesauro, D.S. Touretzky, T.K. Leen (Eds.), Advances in Neural Information Processing Systems 7, NIPS Conference, Denver,Colorado, USA, 1995, pp. 231–238.
[11] P. Melville, R.J. Mooney, Creating diversity in ensembles using artificial data, Information Fusion 6 (2005) 99–111.
[12] E.K. Tang, P.N. Suganthan, X. Yao, An analysis of diversity measures, Machine Learning 65 (1) (2006) 247–271.
[13] R.E. Schapire, Y. Freund, P. Bartlett, W.S. Lee, Boosting the margin: a new explanation for the effectiveness of voting methods, The Annals of Statistics 26 (1998) 1651–1686.
[14] Ph. Smets, The combination of evidence in the transferable belief model, IEEE Transactions on Pattern Analysis and Machine Intelligence 12 (5) (1990) 447–458.
[15] Ph. Smets, Analyzing the combination of conflicting belief functions, Information Fusion 8 (2007) 387–412.
[16] W. Liu, Analyzing the degree of conflict among belief functions, Artificial Intelligence 170 (11) (2006) 909–924.
[17] Y. Bi, An efficient triplet-based algorithm for evidential reasoning, International Journal of Intelligent Systems 23 (4) (2008) 1–34.
[18] Y. Bi, J. Guan, D. Bell, The combination of multiple classifiers using an evidential approach, Artificial Intelligence 17 (2008) 1731–1751.
[19] Y. Bi, D. Bell, H. Wang, G. Guo, J. Guan, Combining multiple classifiers using Dempster's rule for text categorization, Applied Artificial Intelligence 21 (3) (2007) 211–239.
[20] Y. Bi, S. Wu, Measuring impact of diversity of classifiers on the accuracy of evidential ensemble classifiers, IPMU 1 (2010) 238–247.
[21] S.S. Anand, D. Bell, J.G. Hughes, EDM: a general framework for data mining based on evidence theory, Data Knowledge Engineering 18 (3) (1996) 189–223.
[22] R.R. Yager, On the Dempster–Shafer framework and new combination rules, Information Science 41 (1987) 93–137.
[23] D. Dubois, H. Prade, On the unicity of Dempster's rule of combination, International Journal Intelligent System 1 (1986) 133–142.
[24] L. Zhang, Representation, independence, and combination of evidence in the Dempster–Shafer theory, in: R.R. Yager, J. Kacprzyk, M. Fedrizzi (Eds.), Advances in the Dempster–Shafer Theory of Evidence, John Wiley & Sons, New York, 1994, pp. 51–69.

[25] J. Gordon, E.H. Shortliffe, A method for managing evidential reasoning in a hierarchical hypothesis space, Artificial Intelligence 26 (1985) 323–357.
[26] T. Denoeux, Conjunctive and disjunctive combination of belief functions induced by nondistinct bodies of evidence, Artificial Intelligence 172 (2–3) (2008) 234–264.
[27] B. Quost, M.-H. Masson, T. Denoeux, Classifier fusion in the Dempster–Shafer framework using optimized t-norm based combination rules, International Journal of Approximate Reasoning 52 (3) (2011) 353–374.
[28] E. Lefevre, O. Colot, P. Vannoorenberghe, Belief function combination and conflict management, Information Fusion 3 (2002) 149–162.
[29] D. Skalak, The sources of increased accuracy for two proposed boosting algorithms, in: Proc. American Association for Artificial Intelligence, AAAI-96, Integrating Multiple Learned Models Workshop, 1996.
[30] R. Kohavi, D. Wolpert, Bias plus variance decomposition for zero-one loss functions, in: L. Saitta (Ed.), Machine Learning: Proc. 13th International Conference, Morgan Kaufman, 1996, pp. 275–283.
[31] J.L. Fleiss, J. Cuzick, The reliability of dichotomous judgments: unequal numbers of judgments per subject, Applied Psychological Measurement 3 (1979) 537–542.
[32] G. Yule, On the association of attributes in statistics, Philosophical Transactions of the Royal Society A 194 (1900) 257–319.
[33] D.D. Margineantu, T.G. Dietterich, Pruning adaptive boosting, Proceedings of the 14th International Conference on Machine Learning, Morgan Kaufman, 1997
[34] T. Ho, The random space method for constructing decision forests, IEEE Transactions on Pattern Analysis and Machine Intelligence 20 (8) (1998) 832–844.
[35] K. Tumer and J. Ghosh, Linear and order statistics combiners for pattern classification, in: A. Sharkey (Ed.), Combining Artificial Neural Nets, 1999, pp. 127–161.
[36] C.L. Blake, C.J.E. Keogh, Uci repository of machine learning databases, <http://www.ics.uci.edu/mlearn/MLRepository.html>.
[37] I.H. Witten, E. Frank, Data Mining: Practical Machine Learning Tools and Techniques, second ed., Morgan Kaufmann, San Francisco, 2005.