Pattern Recognition ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

Pattern Recognition

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

Information theoretic combination of pattern classifiers

Julien Meynet [a,*,1], Jean-Philippe Thiran [b]

[a] Yahoo! France R&D, Parc Sud Galaxie, 38130 Echirolles, France
[b] Ecole Polytechnique Fédérale de Lausanne (EPFL), Signal Processing Laboratories (LTS5), CH-1015 Lausanne, Switzerland

ARTICLE INFO

Article history:
Received 28 October 2009
Received in revised form
17 March 2010
Accepted 19 April 2010

Keywords:
Machine learning
Pattern recognition
Classifier combination
Information theory
Mutual information
Diversity

ABSTRACT

Combining several classifiers has proved to be an effective machine learning technique. Two concepts clearly influence the performances of an ensemble of classifiers: the diversity between classifiers and the individual accuracies of the classifiers. In this paper we propose an information theoretic framework to establish a link between these quantities. As they appear to be contradictory, we propose an information theoretic score (ITS) that expresses a trade-off between individual accuracy and diversity. This technique can be directly used, for example, for selecting an optimal ensemble in a pool of classifiers. We perform experiments in the context of overproduction and selection of classifiers, showing that the selection based on the ITS outperforms state-of-the-art diversity-based selection techniques.

1. Introduction

In many machine learning problems, combining the decisions of several classifiers has shown to be an effective technique for improving the classification performances. Dietterich [1] gives three main reasons why an ensemble of classifiers might be a better choice than a single classifier. First, when the same classification accuracy can be achieved by several classifiers (in particular when the available training set is small), averaging all the decisions can reduce the risk of picking the wrong classifier. Then, many learning techniques use local searches to converge toward a solution (e.g. neural networks techniques), with the risk of staying stacked in local optima. Running several searches and combining the solutions can improve the performances. Finally, from a representational perspective, we may not be able to obtain the optimal classifier using a given training set and a given classifier architecture. Combining several classifiers can produce a better approximation of the optimal solution. In [2], Freund and Shapire also discuss why averaging classifiers can avoid overfitting.

Many techniques have been proposed in the past few years for combining classifiers. There are basically two different combination scenarios. On the first hand, outputs of several classifiers can be fused together. One can obtain several classifiers by using various learning techniques (e.g. different algorithms, different initializations, different parameters,etc.) or various training data (different training sets or feature sets). The combination can be performed using either trainable rules (e.g. a classifier as combination rule) or fixed rules (e.g. simple probability rules or voting strategies). On the second hand, the combination can be implemented in ensemble creation scheme. Classifiers are iteratively added in the ensemble such that a strong ensemble of classifiers is produced. The most known ensemble growing techniques are Bagging [3], AdaBoost [4] and Random Forests [5]. More detailed surveys on classifier combination can be found in [6,7]. From a practical perspective, the use of an ensemble is only justified if it becomes better than its best individual member. To achieve this requirement, classifiers need to commit errors on different data. This concept refers to the notion of diversity of the classifiers. However it is known (e.g. [6])—and we will also demonstrate it—that directly maximizing diversity measures does not necessarily lead to optimal ensembles.

In this paper, we propose a novel approach to tackle this issue. We develop an information theoretic framework that defines a new measure of the goodness of an ensemble of classifiers which is based on the trade-off between the individual accuracies and the diversity between the classifiers. This framework was already introduced in [38] from a practical perspective, with a specific application to AdaBoost. In this paper we provide a complete theoretical analysis with detailed experimental work. As an applicative example, we propose to use the new measure as a criterion for selecting an optimal ensemble in a predefined pool of classifiers.

The paper is organized as follows: in Section 2, we introduce and discuss the notion of diversity between classifiers and its relationship to ensemble accuracy. Then, after reviewing in

* Corresponding author. Tel.: +33 480384123; fax: +41 21 6937600.
E-mail address: julien.meynet@gmail.com (J. Meynet).
[1] Julien Meynet was at EPFL, LTS5 when this research work was conducted.

0031-3203/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2010.04.013

Section 3 some basic notions of information theoretic classification, we present information theoretic combination of classifiers in Section 4. The new score is defined in Section 5, while Section 6 presents an example of potential application of the score, in the context of overproduction and selection of classifiers. Finally we draw some conclusions in Section 7.

2. Diversity in ensembles of classifiers

It is commonly admitted that large diversity between classifiers in a team is preferred. However, diversity can be understood differently depending on the context. It can be viewed as a measure of dependence, complementarity or even orthogonality between classifiers [8]. In practice, diversity can be used in three different philosophies. First, it can be directly used as an optimization criterion for training diverse ensembles. Then, it is often used implicitly in the ensemble growing techniques, where the ensembles become progressively diverse (e.g. AdaBoost). Finally it can be used for controlling the relevance of an ensemble by checking if it is diverse enough.

In [9], Cunningham et al. claim that "any work with classification ensembles should explicitly measure diversity in the ensemble". Giacinto et al. [10] states that classifiers in an ensemble need to be "accurate and diverse". Several studies focused on understanding how diversity was handled on various ensemble creation techniques like AdaBoost or Bagging [11,12]. Finally, many techniques have been proposed for exploiting diversity for finding good ensembles [13–18]. It was even proposed to voluntarily overtrain the classifiers in order to create diversity between them [19]. In all these studies, various diversity measures have been proposed. We give hereafter a general overview of the most significant ones.

2.1. Diversity measures

Let us consider that we want to associate an example $\mathbf{x} \in \mathbb{R}^d$ to a class y. Let $f_1 f_2$ be decision functions of two different classifiers such that $f_1(\mathbf{x})=y_1$ and $f_2(\mathbf{x})=y_2$.

We define the following probabilities of the respective pairs of correct/incorrect classifications: $a=P(y_1=y,\ y_2=y)$, $b=P(y_1\neq y,y_2=y)$, $c=P(y_1=y,y_2\neq y)$, $d=P(y_1\neq y,y_2\neq y)$. The most used diversity measure is certainly Yule's Q-statistic [20] (QS). It is defined by:

$$Q=\begin{cases}\dfrac{ad-bc}{ad+bc} & \text{if} \quad a,\ b,\ c,\ d<1,\\ 1 & \text{otherwise.}\end{cases} \tag{1}$$

As shown in Eq. (1), two statistically independent classifiers will have $Q=0$. Q varies between -1 and 1, the lower the value the more diverse the classifiers. Classifiers that tend to recognize the same objects correctly will have positive values of Q, and those which commit errors on different objects will render Q negative.

Then other similar diversity measures have been proposed, among them, the disagreement measure [21] corresponding to the total proportion of examples for which the two classifiers disagree, the double fault measure [10] which counts the proportion of examples misclassified by both classifiers.

On the other hand, there exists several non-pairwise diversity measures. The most used is the Kohavi–Wolpert variance [22]. The diversity measures introduced in this section clearly present correlation between them and, as pointed out in [12], there is no best diversity measure that can be used for building ensembles with minimal error. Moreover, finding a systematic relationship between these diversity measures and ensemble accuracy is a very challenging task.

2.2. Limits of diversity measures

The diversity measures introduced here above have been extensively used in many applications, particularly in the context of classifier selection from a large set of classifiers [10]. However, is has been observed in practice that explicitly maximizing diversity measures is not as successful as expected. Several studies proposed to understand, both theoretically and practically, why these diversity measures present limitations for obtaining effective ensembles. In [23], Kuncheva explains why the use of these diversity measures becomes irrelevant when combining two classifiers. Then, empirical studies [8,13], showed that the relationship between diversity measures and ensemble accuracy is somehow confusing. Kuncheva also reported in [6] that the improvement on the best individual accuracy by forcing diversity is negligible. In [14], Hadjitodorov showed that, in some particular cases, using moderate diversity can produce better ensemble than maximum measure of diversity.

Finally Tang et al. [24] recently gave theoretical insights showing that the diversity measures are in general ineffective. In particular, they proved that using diversity measures usually produces ensembles with large diversity, but not maximum diversity.

These considerations explain why diversity is usually only used for visualization (plot pairs of classifiers according to their diversity), or overproduction and selection of classifiers. More details about diversity and how to create diversity in ensemble are given in [15].

In the following sections, we will investigate this paradoxical behavior of diversity. On the one hand it is known as a major factor in ensemble design, on the other hand using state-of-the-art diversity measures dos not systematically lead to improved performances compared to the best member of the team.

3. Introduction to information theoretic classification

3.1. Motivations

In this section we will introduce an information theoretic framework for combining classifiers. We will first motivate our choice by showing how information theory (IT) can help tackling the ambiguity of diversity as explained in the previous section. We will use information theoretic tools to understand why selecting the most diverse ensemble is not necessarily the best choice.

Information theory is commonly used in coding and communication applications and more recently, is has also been used in classification area. Information theoretic classification was first introduced by Principe et al. [25]. Basically, a learner is viewed as an agent that gathers information from some external sources. Information theoretic quantities have been widely used for feature extraction and selection, e.g. Fisher et al. [26], Hild et al. [27] or Sindhwani et al. [28], who proposed a feature selection technique for support vector machines and neural networks. Recently, Butz et al. [29] proposed to apply this framework to multi-modal signal processing.

Multi-modal signals represent several signals of different modalities but coming from the same physical scene. The underlying idea is that the information contained in one signal can help for processing other modalities, and, IT offers a variety of tools for handling the exchange of information between the source and the signals, and between the signals of several modalities.

In this work, we propose to model classifier combination as a similar problem, considering that several classifiers are trained

Please cite this article as: J. Meynet, J.-P. Thiran, Information theoretic combination of pattern classifiers, Pattern Recognition (2010), doi:10.1016/j.patcog.2010.04.013

from examples coming from the same physical sample distribution. IT can provide efficient tools for measuring and analyzing dependency between classifiers and of course accuracy of the classifiers.

3.2. Information theoretic definitions

Let us first review some basic IT concepts that will be used in the remaining of the paper. More details can be found in [30]. Shannon's entropy $H_S(X)$ of a discrete random variable X with probability density function $p(x)$ is defined by $H_S(X)=-\sum_k p(x_k)\log p(x_k)$. Consider two random variables X and Y with a joint probability density function $p(x,y)$ and marginal probability density functions $p(x)$ and $p(y)$, then Shannon's mutual information $I_S(X;Y)$ between X and Y is defined by $I_S(X;Y)=\sum_k\sum_j p(x_k,y_j)\log p(x_k,y_j)/p(x_k)p(y_j)$. For notation simplicity and if not specified otherwise, Shannon's definitions $H_S(Y|X)$ and $I_S(X;Y)$ will be written $H(Y|X)$ and $I(X;Y)$. The relationships between entropy and mutual information can be represented graphically by means of Venn diagrams as shown in Fig. 1. Mutual information represents the information that is shared by both variables X and Y. It can thus be represented by the intersection between both marginal entropies $H(X)$ and $H(Y)$.

Shannon's definitions of entropy and mutual information have been extended to the more general Renyi's definitions: $H_\alpha(X)=(1/(1-\alpha))\log\sum_k p^\alpha(x_k)$ with $\alpha>0,\alpha\neq 1$ and $I_\alpha(X;Y)=(1/(1-\alpha))\log\sum_k\sum_j p^\alpha(x_k,y_j)/p^{\alpha-1}(x_k)p^{\alpha-1}(y_j)$.

3.3. Information theoretic classification

The classification process can be modeled using an information theoretic framework. Let us first introduce some notations and variables concerning information theoretic classification, that will be used in the remaining of the paper.

Let us assume that we have a set X of n examples, obtained from a physical signal that we denote S. The true class labels of the examples are represented by a random variable C. C is defined over the set of classes Ω_c ($\Omega_c=\{-1;1\}$ in a binary classification task). Let us denote F the feature vectors of the examples, obtained by feature selection and extraction. The class labels estimated by the classification process is called $\hat{C}$. The common classification problem can be summarized by a simple processing chain: acquisition of the signal, feature selection, feature extraction and classification. In information theory, this processing chain can be formulated as a first order Markov chain [25,29], as depicted in Fig. 2, along with the main classification steps.

The ultimate goal in classification is to minimize the difference between the true labels and the estimated class labels. This can be modeled by considering a random variable E taking values into {1,0}. The probability of making an error during the classification process is thus:

$$P_e=P(E=1)=P(\hat{C}\neq C). \tag{2}$$

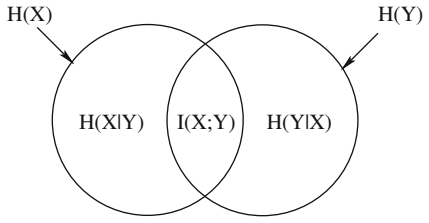

Fig. 1. Venn Diagram representing the concept of entropy and mutual information. Mutual information can be viewed as the intersection between the marginal entropies.

In this work, we will consider the classification as a general problem of classifying examples to classes. We will thus integrate the pre-processing steps (signal acquisition, feature selection and feature extraction) into the classification step, resulting in one single random variable $\hat{C}$ for the whole classification process.

The complete classification process can thus be simplified into the following first order Markov chain:

$$C\rightarrow\hat{C}\rightarrow E. \tag{3}$$

Intuitively, the estimated class labels $\hat{C}$ should contain as much information about the class labels C as possible. In other words, we would like to maximize the mutual information between the true labels and the estimated labels $I(C;\hat{C})$. This intuitive idea can be formalized by trying to minimize bounds on the error probability P_e.

In [32], Erdogmus et al. give a generalization of Fano's inequality [31] for bounding the error probability P_e:

Theorem 3.1 (*Erdogmus and Principe [32]*). *Considering the first order defined in Eq.* (3), *the probability of making an error is bounded by*:

$$\frac{H_S(C)-I_\alpha(C;\hat{C})-h_S(P_e)}{\log|\Omega_c|-1}\leq P_e\leq\frac{H_S(C)-I_\beta(C;\hat{C})-h_S(P_e)}{\min_k H_S(C|e,\hat{c}_k)}, \tag{4}$$

where $h_S(P_e)=-P_e\log P_e-(1-P_e)\log(1-P_e)$ *is the binary Shannon's entropy, and* $I_{\alpha,\beta}(C;\hat{C})$ *represents Renyi's definition of the mutual information with* $\alpha,\beta\in\mathbb{R}^+\backslash\{1\}$.

As the number of classes $|\Omega_c|$ is fixed, the entropy of the class labels $H_S(C)$ does not depend on the classification process. Bounds in Eq. (4) point out that maximizing the MI between the two random variables C and $\hat{C}$ will tend to minimize both bounds, thus increasing the chances of having a low error probability P_e. Clearly minimizing both bounds in Theorem 3.1 does not mean necessarily minimizing P_e, however, if the lower bound is high, the error we be also be high. In the next section we will extend these properties to the framework of multiple classifiers.

4. Information theoretic combination of classifiers

The information theoretic framework introduced in the previous section was referring to the general classification task. This section shows how it can be extended to the case where the classification problem is more specifically a combination of several classifiers.

Let us assume that we have a team of K given classifiers. Let us now denote $\hat{C}$ the random variable representing the estimated class labels obtained by aggregation of the individual decisions. The aim is thus to find the best combination of members in the sense that it will maximize $I(C;\hat{C})$.

Let us call C_i, $i=1,\ldots,K$, the random variables representing the decisions of classifiers $i=1,\ldots,K$. Each classifier can be modeled separately using the Markov chain in Eq. (3).

Conceptually, the combination process can be summarized by the Venn diagram shown in Fig. 3. The accuracy of an individual classifier C_i is represented by intersection between $H(C_i)$ and $H(C)$. The mutual information between two classifiers C_i and C_j is the intersection of the two corresponding marginal entropies $H(C_i)$ and $H(C_j)$.

For simplification and without loss of generality, let us consider a two class problem with labels $\{-1,1\}$. The main relationships between the classifiers and the true classes are measured by the mutual information (MI) between them:

- The MI between the output of individual classifier i and the true labels is $I_{C;C_i}$, $i\in\{1,\ldots,K\}$.

Please cite this article as: J. Meynet, J.-P. Thiran, Information theoretic combination of pattern classifiers, Pattern Recognition (2010), doi:10.1016/j.patcog.2010.04.013

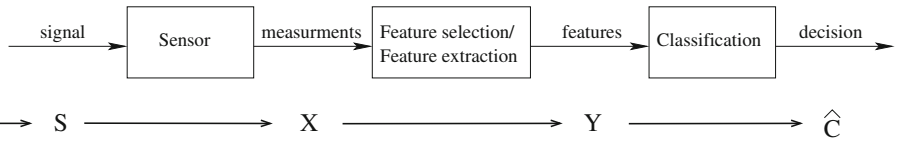

Fig. 2. Different stages of pattern recognition systems, formulated as a first order Markov chain.

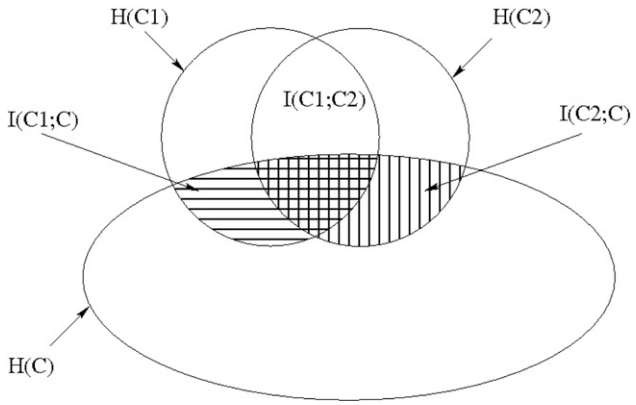

Fig. 3. Venn Diagram representing relationships between two classifiers C 1, C 2 and the true class labels C.

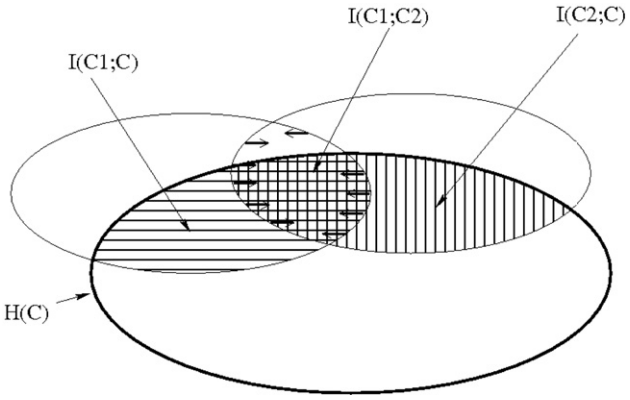

Fig. 4. Venn Diagram showing how to optimize classifier combination.

- The MI between two classifiers i and j is: $I_{C_i;C_j}$, $i,j \in \{1,\ldots,K\}$, $j > i$.

Let $P_{C;\hat{C}}$ be the joint probability density function between true classes and the ensemble decision and P_C, $P_{\hat{C}}$ the corresponding marginal probabilities. The quantity that we want to maximize is the MI between the true labels and the labels obtained by aggregation of the individual decisions, $\hat{C}$:

$$I_{C;\hat{C}} = \sum_{k=-1,1} \sum_{j=-1,1} P_{C;\hat{C}}(k,j) \log \frac{P_{C,\hat{C}}(k,j)}{P_C(k) P_{\hat{C}}(j)}. \tag{5}$$

The combination rule can be implemented using variety of strategies. The simplest combination rule is the majority voting (MV). Despite its simplicity, MV has proved to be an effective rule for many combination tasks. Moreover, MV can easily be extended to weighted majority voting which is widely used in the multiple classifiers community. For example, the decision of AdaBoost [4] is a weighted majority vote of weak classifiers. Many studies [33–36] have focused on analyzing why majority voting was as effective as more complicated schemes in improving the pattern recognition results. In the remaining of the paper, we restrict the combination rule to majority voting.

4.1. Majority voting for combining classifiers

Considering a MV combination scheme, the probability $P(\hat{C} = y)$ that $\hat{C}$ outputs y is related to each voter classifier by [37]

$$P(\hat{C} = y) \leq \sum_{i=1}^{K-1} \sum_{j=i+1}^{K} P(C_i = y, C_j = y). \tag{6}$$

Then, the following theorem gives the relationship between the ensemble accuracy and the individual accuracies as a function of the numbers of voters:

Theorem 4.1 (*Shapley and Grofman [37]*). *Consider a group of odd size K with any distribution of individual accuracies* $(p_1,\ldots,p_K)$, *where* $p_i > 0.5$ $\forall i$. *The probability to reach the correct decision, when utilizing the simple majority rule, is larger or equal to the probability* $p = (1/K)\sum_{i=1}^{K} p_i$ *of a random group member to do so.*

In our case this theorem leads to

$$P(C = y, \hat{C} = y) \geq \frac{1}{K} \sum_{i=1}^{K} P(C = y, C_i = y). \tag{7}$$

Considering bounds Eqs. (6) and (7) on each term of the mutual information in Eq. (5), we see that minimizing the MI between each pair of classifier $I_{C_i;C_j}, i \neq j$ and maximizing the MI between each single classifier and the true class labels $I_{C;C_i}$ will tend to maximize the mutual information between the ensemble decision and true labels: $I_{C;\hat{C}}$, which proves the following Theorem that we propose:

Theorem 4.2. *Let* $C_1,C_2,\ldots,C_K$ *be K random variables representing the output labels of K classifiers and C a random variable representing the true class labels. Maximizing* $I_{C;C_i}$ $\forall i \in \{1,\ldots,K\}$ *and minimizing* $I_{C_i;C_j}$ $\forall i \in \{1,\ldots,K\}$, $\forall j \in \{1,\ldots,K | j > i\}$, *will maximize* $I_{C;\hat{C}}$. $\hat{C}$ *represents the estimated class labels obtained from* $C_1,\ldots,C_K$ *by majority voting.*

As introduced in Section 3, $I_{C;C_i}$ can be viewed as a measure of accuracy of classifier i. $I_{C_i;C_j}$ measures the similarity between the two classifiers i and j. In other words, by minimizing $I_{C_i;C_j}$, we maximize the diversity between the two classifiers.

It is important to note that Theorem 4.2 represents a sufficient condition for maximizing $I(C;\hat{C})$, but it is clearly not a necessary condition. In fact, it is possible to have an accurate ensemble of classifiers which does not maximize diversity between classifiers. This will be discussed experimentally in Section 6.

Theorem 4.2 can be summarized graphically by the Venn diagram shown in Fig. 4.

4.2. Diversity/accuracy dilemma

The relationships shown in Theorem 4.2 reveal a paradox in the sense that the two measures involved are somehow contradictory. In fact, two very good classifiers will clearly have very low diversity, while two poor classifiers, say slightly better than random guessing, will be very likely diverse. This paradox can easily be seen using Venn diagrams Fig. 5(a) and (b). Fig. 5(a) shows that maximizing both individual accuracies will tend to

Please cite this article as: J. Meynet, J.-P. Thiran, Information theoretic combination of pattern classifiers, Pattern Recognition (2010), doi:10.1016/j.patcog.2010.04.013

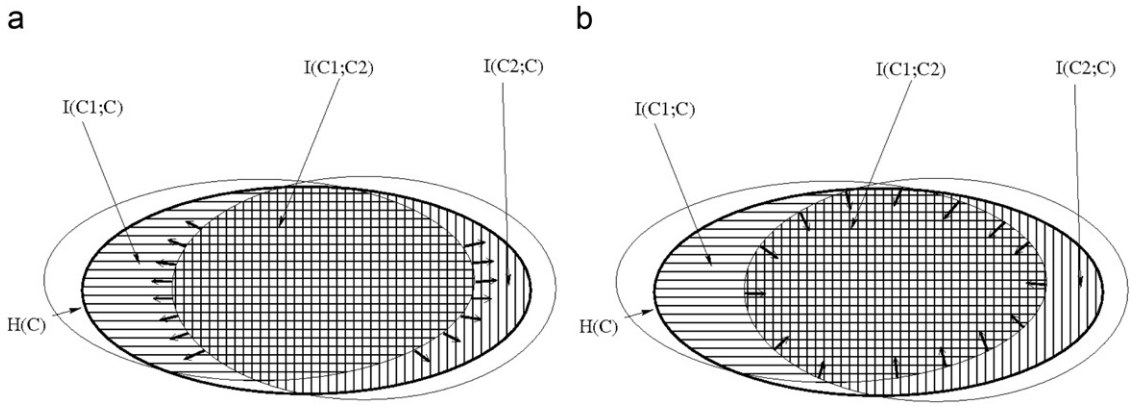

Fig. 5. Diversity accuracy dilemma: (a) maximize both accuracies; (b) maximize diversity.

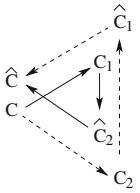

Fig. 6. Coupled Markov chains for two classifiers trained differently from the same input data.

expand the intersection between the two marginal entropies, which is equivalent to increase similarity between classifiers. Inversely, Fig. 5(b) shows that minimizing similarity between classifiers will force to decrease individual accuracies.

This phenomenon can be discussed through the following formalism. Consider two random variables C_1, C_2 representing two classifiers and $\hat{C_1}, \hat{C_2}$ their respective class labels. Let C be the true class labels.

To establish a probabilistic link between the two classifiers, a parallel is made with the work of Butz et al. in [29] concerning processing of multi-modal signals. First recall some pattern recognition definitions. We consider that the training and testing examples are generated from an unknown but fixed probability density function (*pdf*) and the task is to find a function that minimizes the risk of misclassifying new vectors drawn from the same *pdf*. We can consider that the inputs of both classifiers C_1 and C_2 come from this *pdf*. Two coupled Markov chains can be built

$$\begin{cases} C \to C_1 \to \hat{C_2} \to \hat{C} \to E, \\ C \to C_2 \to \hat{C_1} \to \hat{C} \to E. \end{cases} \tag{8}$$

These coupled Markov chains are depicted in Fig. 6. The probability densities of C_1 and $\hat{C_1}$, resp. C_2 and $\hat{C_2}$, are both estimated from the same data sequences. Therefore we can write $I(C_1;\hat{C_2}) \approx I(C_2;\hat{C_1}) \approx I(C_1;C_2)$. Then, the data processing inequality [30] gives $I(C_1;C_2) \geq I(C;C_2)$ and $I(C_1;C_2) \geq I(C;C_1)$. This implies that

$$I(C_1;C_2) \geq \frac{I(C;C_1)+I(C;C_2)}{2}. \tag{9}$$

Maximizing the individual accuracies represented by $I(C;C_1)$,$I(C;C_2)$ will consequently maximize $I(C_1;C_2)$, the similarity between the classifiers. Inversely, minimizing $I(C_1;C_2)$ (maximizing the diversity) will tend to minimize the classifiers accuracy.

This phenomenon reflects the limitations of diversity-based techniques, as presented in Section 2.2. To address the contradiction presented here, a trade-off needs to be introduced. A study of how the diversity evolves depending on the classifiers accuracies is given in the next section.

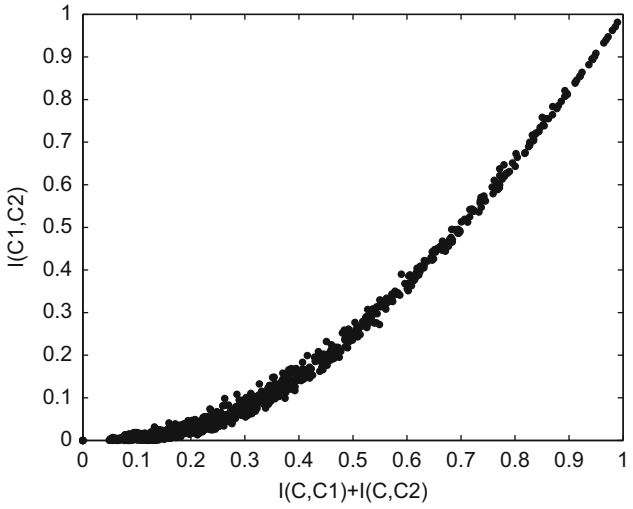

Fig. 7. The similarity of two classifiers $I(C_1; C_2)$ function of the average individual accuracy $(I(C_2;C)+I(C_1;C))/2$. The two classifiers have the same individual accuracy.

5. Information theoretic score

5.1. Estimation of the relationship between diversity and classifiers accuracy

This section proposes an empirical estimation of the relationship between diversity and accuracy in order to give a computable measure of the ensemble performance. This link is estimated with the following experiment. Outputs of two classifiers (C_1, C_2) with equal individual accuracies between 0.5 and 1 (i.e. classifiers better than random guessing) are iteratively simulated. We report in Fig. 7 the similarity between output labels $I(C_1; C_2)$ for each trial as a function of the individual accuracy $(I(C;C_1)+I(C;C_2))/2$.

A simple possible modeling of the relationship is to approximate similarity by a quadratic function of the average individual accuracy. Fig. 8 gives a graphical interpretation of this approximation. A classifier is represented by a 2-dimensional vector. Its projection onto the horizontal axis measures its individual accuracy while the difference between vertical projections of two vectors measures the diversity between them. The dash line represents the maximal diversity allowed between two classifiers with identical accuracy. This fits with the

Please cite this article as: J. Meynet, J.-P. Thiran, Information theoretic combination of pattern classifiers, Pattern Recognition (2010), doi:10.1016/j.patcog.2010.04.013

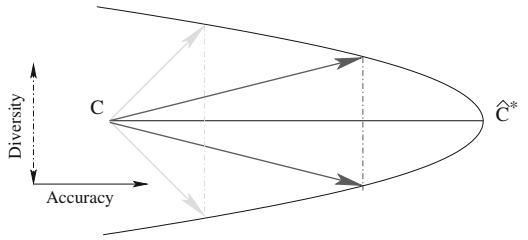

Fig. 8. Graphical representation of accuracy/diversity dilemma.

remark that two poor classifiers can have large diversity while two accurate classifiers cannot be so diverse.

In the following, we will consider two terms based on the mutual information between classifiers: one measuring average accuracy, the other measuring diversity.

Definition 5.1. The average accuracy of the K classifiers called information theoretic accuracy (ITA)

$$ITA = \frac{\sum_{i=1}^{K} I(C;C_i)}{K}. \tag{10}$$

Definition 5.2. The average diversity between the classifiers is called information theoretic diversity (ITD)

$$ITD = \frac{\binom{K}{2}}{\sum_{i=1}^{K-1}\sum_{j=i+1}^{K} I(C_i;C_j)}. \tag{11}$$

In this work we propose to use a simple first order statistic to measure the individual accuracies and diversities, for two main motivations. On the first hand, the goal is to design a simple global score that can be used in various classifier combination applications. Using other statistics could increase significantly the computational costs of the measure. On the second hand, Theorem 4.2 does not help us to discriminate between ensembles having a large variance in the individual accuracies (thus large diversity) and ensembles having low variance between the individual accuracies and possibly less diversity. In this case we propose to keep the ensembles with high average accuracy even if diversity between them is penalized. It avoids the limitations of diversity-based techniques presented in Section 2.2.

In order to design a score that reflects Theorem 4.2, we need to consider the quadratic approximation of the similarity between the classifiers and the average accuracy presented before. This relationship can be written as $ITA^2 \propto 1/ITD$. In fact, the diversity term ITD already contains relevant information about the average individual accuracy. We thus propose to compensate this information by considering the following Information Theoretic Score (ITS) as a function of ITS and ITD:

Definition 5.3. The information theoretic score (ITS) of an ensemble of K classifiers combined by majority voting is defined by

$$ITS = (1+ITA)^3(1+ITD). \tag{12}$$

ITS can also be written $ITS=(1+ITA)(1+ITA)^2(1+ITD)$. The first factor based on ITA forces to choose the most accurate classifiers. The other terms maximize diversity for ensembles with identical ITA. This score will tend to select the best ensemble of classifiers by only considering diversity when it becomes a relevant feature. Compared to standard diversity based techniques, this will penalize ensembles with low ITA and large ITD. Moreover, the ITS can be evaluated very easily. The mutual information terms involved in the ITS are estimated from the labels output by the classifiers. In other words, the dimensionality of the space is the number of classes. Considering that we have a very small number of classes compared to the number of available samples, computing the mutual information terms does not suffer from estimation in high dimensional spaces. Finally, note that the model that we propose is a choice and other similar modelings could be chosen. The next section tries to validate this definition in the context of overproduction and selection of classifiers.

5.2. Validation of the ITS

To evaluate the intrinsic behavior of the ITS, we first consider simulated classifier outputs. By generating random outputs we can explore the complete space of output labels. It presents the advantage of being completely independent of the process of feature selection and independent of the learning algorithm. We can thus perform an unbiased evaluation of the ITS. Let us consider the following simple experimental setup. We randomly generate output vectors for three classifiers. For each run, we measure the accuracy of the majority voting ensemble and the ITS. The results are shown in Fig. 9. Note that, in this experiment, we do not impose the individual accuracies to be identical, we only constraint them to fall between 0.5 and 1.

As expected, ensembles with high ITS are accurate. Moreover, an ensemble can be accurate but with a low ITS, therefore, the condition for maximizing $I(C;\hat{C})$ is sufficient but not necessary.

5.3. ITS in multi-class problems

In Section 5.1, ITS was defined according to empirical considerations based on binary classification purposes. However, we will show in this section that the ITS can also be used in multi-class problems. From a practical point of view, increasing the number of classes will increase the chances of having different outputs between the classifiers. This phenomenon is reflected by a higher diversity for a fixed individual accuracy. In order to check this multi-class behavior, we performed the same experiments as in Section 5.2, with simulated output labels but with various number of classes from 2 to 6. Results are reported in Fig. 10. As expected, the accuracy/diversity representation still holds for multi-class problems, the diversity being an increasing function of the number of classes. The global relationship between ITD and ITA does not depend on the number of classes. Nevertheless, for very large number of classes, an adaptation of the diversity term

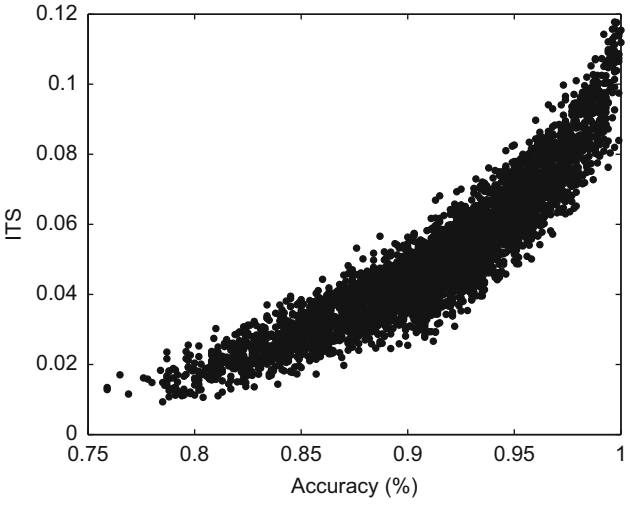

Fig. 9. Score behavior with synthetic class labels.

Please cite this article as: J. Meynet, J.-P. Thiran, Information theoretic combination of pattern classifiers, Pattern Recognition (2010), doi:10.1016/j.patcog.2010.04.013

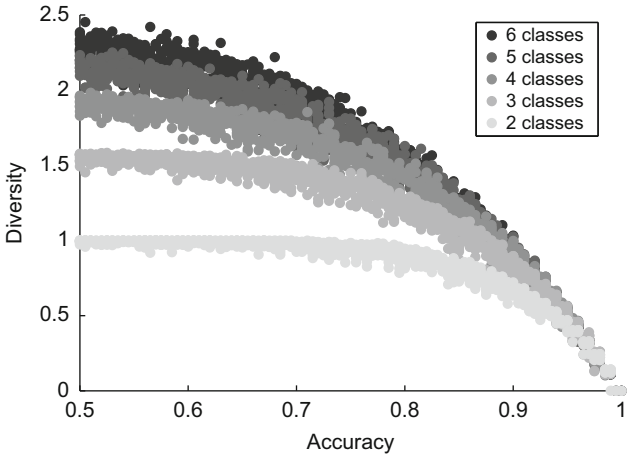

Fig. 10. ITS in multi-class problems.

can be imagined. In the remaining experiments of this paper, we will consider datasets with up to 10 classes.

5.4. Discussion

ITS fixes a trade-off between average individual accuracies and diversity. It can be used as a global measure of ensemble efficiency. However it presents some limitations. The main underlying hypothesis we made in the experiments for designing the new measure, is that the variance of individual accuracies in the ensemble is small. Clearly, if the individual accuracies in the ensemble are really balanced, approximating the individual accuracies by their average becomes irrelevant. In this case, no global relationship between accuracy and diversity can be found. For example, combining two classifiers, one very accurate and one poor cannot be optimized by measuring diversity between them. However, in such cases, the contribution of ITA term in ITS will be more important than ITD, resulting in the selection of the best classifiers even if they are not diverse.

In fact, the general idea of ITS is to adapt the contributions of both terms ITA and ITD depending on the context. If the classifiers to be combined have almost identical individual accuracies, then the contribution of ITD in the ITS will be discriminant. It there exists more difference between individual accuracies, then ITA becomes more relevant and it then preferable to choose good classifiers even if they have more redundancy between them.

The other drawback of this information theoretic framework is that the proposed criterion is not differentiable. It cannot be used directly as optimization criterion for building good ensembles. There are several alternatives to tackle this limitation. On the one hand, we will propose to use ITS as a measure for controlling the performance of classifiers ensembles. This can be done for example in the context of overproduction and selection of classifiers. As we cannot maximize ITS analytically, we can also imagine techniques for incrementally increasing the ITS of the ensemble. For example, ITS can be used in a modified version of AdaBoost called ITS-Boost (see [38]) by selecting at each iteration the weak classifier that maximizes a weighted version of the ITS. It can also be used for training iteratively ensembles of Support Vector Machines [39].

An important remark is that, we do not really need to find the best ensemble in the sense that it maximizes the ITS. As pointed out in Theorem 4.2, the conditions on the mutual information between classifiers and true classes do not imply finding the best ensemble but means finding one of the best.

In the next section, we will present a possible application of the new score: overproduction and selection of classifiers.

6. Experiments and results

6.1. A simple two-dimensional binary problem

For evaluating the relevance of the ITS defined above on a real classification task, we first consider a 2 class toy problem using the Banana dataset available in the Matlab Pattern Recognition Toolbox [40]. We generate 1000 training examples for both classes and we split this training set into 15 smaller subsets by random sampling. We then train one classifier with each subset. A first experiment (Fig. 11(a)) consists in training 15 support vector machines (SVM) with 3rd order polynomial kernels (the C parameter being evaluated by cross-validation). The 455 possible combinations of three classifiers (called triplets in the following) are exhaustively tested. For each triplet, we measure the ITS on the training set, the ensemble accuracy on a large test set and we also compute the average individual accuracy of the three classifiers. This average accuracy is represented by the gray level of the disks in Fig. 11(a).

In the second experiment, three different learning algorithms are used. We trained five SVM, five linear classifiers and five K-nearest neighbors KNN and again ITS is measured for each triplet. Results are reported in Fig. 11(b).

As expected, the triplets of classifiers with low ITA (dark disks) lead to low ensemble classification accuracy. When the three individual classifiers are accurate individually (light disks in Fig. 11(a) and (b)), the final classification is generally accurate. However, in both configuration, the white points (which means the three best classifiers combined together) do not necessarily give the best combination. This phenomenon is more visible in the case of 15 SVM as they only have slight differences in their individual accuracies. In any case, the ensembles with high ITS are very accurate. These experiments show that, at least in toy problems, the ITS can overcome the limitations of diversity as presented in Section 2.

In Fig. 12(a) and (b), we show graphical examples of classifier selection by ITS. We generated 10 linear classifiers in Fig. 12(a) and 10 SVM with 3rd order polynomial kernels in Fig. 12(b). The decision functions selected by maximal ITS are drawn in bold. The two class subspaces are represented by different gray backgrounds.

6.2. Real world datasets

In this section we report experiments on real world datasets taken from the UCI Machine Learning repository [41]. The datasets cover a wide range of applications with number of classes between 2 and 10, with small sample size and large sample size cases. A summary of the datasets used is given in the three first columns of Table 1.

In these experiments, we first trained a set of 15 decision trees (CART trees [42]) on random subspaces of the training set. The choice of the base learner was motivated by the notion of classifier stability. As in Bagging [3], unstable classifiers should be preferred in order to obtain accurate ensembles. However, it is important to recall here that the proposed approach is completely independent of the nature of the classifiers to be combined. One could combine SVMs together or SVMs with decision trees, bayesian classifiers or KNN, etc,…

For each single classifier we measure the error rates by cross-validation for small sample datasets or using a separate test set if

Please cite this article as: J. Meynet, J.-P. Thiran, Information theoretic combination of pattern classifiers, Pattern Recognition (2010), doi:10.1016/j.patcog.2010.04.013

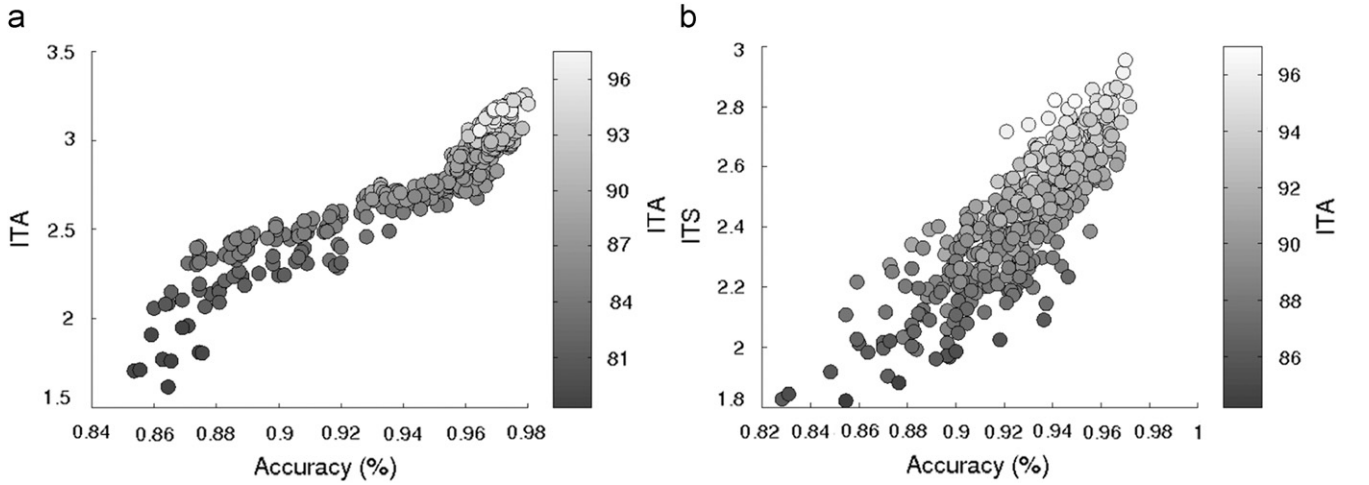

Fig. 11. Combination accuracy and ITS for each triplet of classifiers: (a) 15 SVM with RBF kernels and (b) five SVM with RBF kernels, five KNN classifiers and five linear classifiers. The color of the circle is proportional the average accuracy of the ensembles.

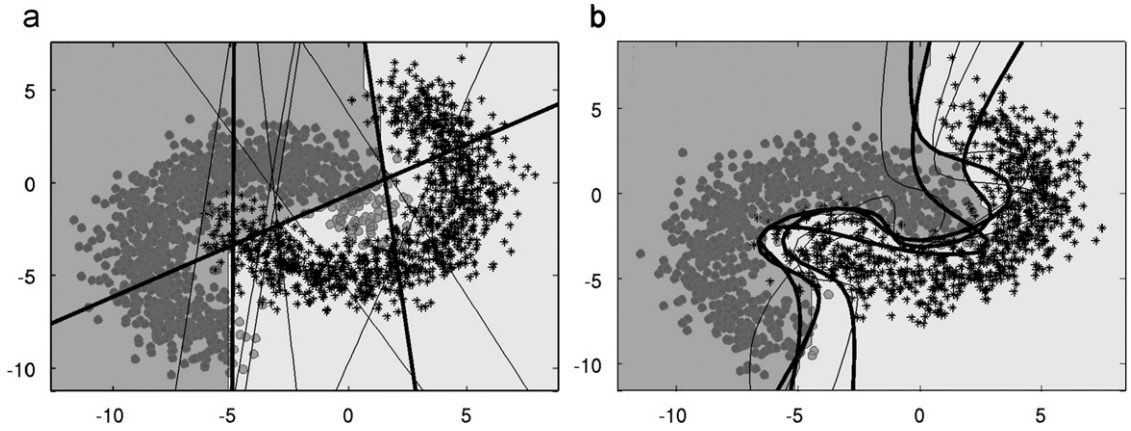

Fig. 12. Example of ensemble selection with ITS. Classifiers are generated on subsets of the complete training set. Bold lines represent the three selected candidates: (a) voting of three linear classifiers; (b) voting of three SVM polynomial, $d=3$.

Table 1
Results on UCI datasets.

Dataset	# classes	# feat.	# ex.	Best classifier	ITS, $K=3$
BreastWC	2	33	198	39.43 ± 0.05	32.21 ± 0.02
Glass	6	9	214	36.91 ± 0.04	33.74 ± 0.05
Image	7	19	2310	23.14 ± 0.01	22.60 ± 0.02
Ionosphere	2	34	351	15.92 ± 0.01	15.45 ± 0.01
Iris	3	4	150	5.21 ± 0.12	5.27 ± 0.14
Pen	10	16	10 992	4.16 ± 0.03	3.78 ± 0.04
Prima	2	8	768	33.37 ± 0.01	32.16 ± 0.01
Wine	3	13	178	29.87 ± 0.02	27.61 ± 0.03
Zoo	7	16	101	12.11 ± 0.04	12.10 ± 0.04

Summary of the datasets used. Number of samples and dimensionality of the input space. We report error rates (in %) of the best single classifier and an ensemble of $K=3$ classifiers created by maximizing ITS.

available (for Image and Pen datasets). For each possible combination of $K=3$, 5 and 7 classifiers, we measure the performance of the ensemble having the highest ITS. The sampling and training procedure has been repeated 10 times in order to obtain reliable classification statistics. In last column of Table 1, we report the mean and standard deviation of the error rates for the best individual classifier and the statistics of the ensemble of three classifiers selected by ITS. It clearly shows that even a small ensemble of 3 classifiers compares favorably with the best individual classifier in most datasets.

In order to compare with a pure diversity based ensemble selection, we also extracted, at each run, the ensemble having the largest average QS. In other words, we compare the ITS measure with the ensemble the most diverse. Results are reported in Table 2 for various numbers of classifiers: $K=3$ and 7. It some datasets, it significantly decreases performances compared to the best individual classifier. This confirms the limitation of the pure diversity-based techniques as described in Section 2.2. Then, ITS-based selection outperforms QS selection in most situations.

Finally, Table 3 shows the influence of number of members in the ensemble. As expected, increasing the number of classifiers in the ensembles increases the performances but in general, small ensembles already give significant improvements compared to the best individual member.

6.3. Iterative selection of classifiers

The purpose of the experiments presented in previous section was to show that the ITS is a relevant measure of ensemble efficiency. It can tackle the limitations of pure diversity based selection methods. For this we tested all possible combinations of $K=3, 5, 7$ classifiers in a pool of $M=15$ classifiers. This means that for each experiment, we needed to test $\binom{M}{K}$. For instance, selecting 7 classifiers in a pool of 15 means testing 6435 ensembles. As many datasets require cross-validation techniques for estimating

Please cite this article as: J. Meynet, J.-P. Thiran, Information theoretic combination of pattern classifiers, Pattern Recognition (2010), doi:10.1016/j.patcog.2010.04.013

Table 2
Results on UCI datasets.

Dataset	Best individual	Q		ITS	
		$K=3$	$K=7$	$K=3$	$K=7$
Breast	39.43 ± 0.05	38.32 ± 0.09	34.12 ± 0.12	32.21 ± 0.02	31.76 ± 0.09
Glass	36.91 ± 0.04	35.42 ± 0.05	33.32 ± 0.02	33.74 ± 0.05	29.71 ± 0.04
Image	23.14 ± 0.01	23.23 ± 0.01	21.09 ± 0.02	22.60 ± 0.02	19.94 ± 0.02
Iono.	15.92 ± 0.01	14.68 ± 0.06	14.23 ± 0.01	15.45 ± 0.01	13.72 ± 0.01
Iris	5.21 ± 0.12	25.67 ± 0.12	20.22 ± 0.11	5.27 ± 0.14	5.05 ± 0.09
Pen	4.16 ± 0.03	5.13 ± 0.07	4.36 ± 0.03	3.78 ± 0.04	3.20 ± 0.04
Prima	33.37 ± 0.01	32.87 ± 0.03	32.12 ± 0.05	32.16 ± 0.01	31.56 ± 0.01
Wine	29.87 ± 0.02	30.52 ± 0.02	28.28 ± 0.04	27.61 ± 0.03	27.03 ± 0.04
Zoo	12.11 ± 0.04	16.10 ± 0.10	9.95 ± 0.08	12.10 ± 0.04	9.75 ± 0.01

Comparison of error rates of various methods: best individual classifier, selection by maximal ITS and selection by maximal QS.

Table 3
Results on UCI datasets.

Dataset	ITS		
	$K=3$	$K=5$	$K=7$
BreastWC	32.21 ± 0.02	32.17 ± 0.01	31.76 ± 0.09
Glass	33.74 ± 0.05	29.78 ± 0.08	29.71 ± 0.04
Image	22.60 ± 0.02	20.92 ± 0.03	19.94 ± 0.02
Ionosphere	15.45 ± 0.01	13.92 ± 0.02	13.72 ± 0.01
Iris	5.27 ± 0.14	5.17 ± 0.09	5.05 ± 0.09
Pen	3.78 ± 0.04	3.25 ± 0.05	3.20 ± 0.04
Prima	32.16 ± 0.01	32.94 ± 0.01	31.56 ± 0.01
Wine	27.61 ± 0.03	27.20 ± 0.02	27.03 ± 0.04
Zoo	12.10 ± 0.04	9.81 ± 0.00	9.75 ± 0.01

Influence of the number of classifiers in the ensemble. We report error rates (mean and standard deviation) for ensembles of $K=3,5$ and 7 classifiers.

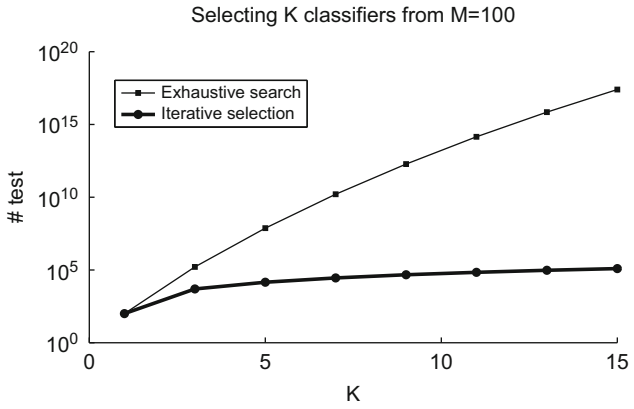

Fig. 13. Number of tests that need to be performed for classifier selection using either exhaustive search (solid) or our iterative selection (bold).

errors, ITS and QS, the exhaustive search is very computationally expensive.

Moreover, in practical applications, the number of classifiers in the pool (M) may be much larger that 15, and the optimal number of classifiers to select (K) is not known a priori. As in the problem of feature selection where we want to keep only the features that are discriminant and not redundant, selection of classifiers in a pool can be seen as selecting only classifiers that are accurate and, if possible, diverse. We can use various sub-optimal alternatives to avoid the exhaustive search of the best classifiers, mainly using greedy algorithms.

For example, we propose to use the following selection procedure:

- First select the best individual classifier C_{1^*}:

$$C_{1^*} = \underset{C_i, i=1,\ldots,M}{\operatorname{argmax}} I(C_i, C). \tag{13}$$

- Then, as majority voting requires an odd number of classifiers, we need to find two more classifiers maximizing the ITS between C_{1^*} and them:

$$(C_{2^*}, C_{3^*}) = \underset{(C_i,C_j), i,j \in \{1,\ldots,M\}\setminus 1^*}{\operatorname{argmax}} ITS(C_{1^*}, C_i, C_j). \tag{14}$$

This procedure can continue recursively until a given number of classifier is reached or until the improvements by adding two more classifiers becomes small enough. Once the first classifier has been selected, we need to extract two other classifiers from the $M-1$ remaining. Consequently, each iteration i of the procedure, we need to perform $\binom{M+1-2i}{2}$ tests in order to find the two new optimal classifiers. The total number of tests for selecting K classifiers from M is

$$N_{tests} = M + \sum_{i=1}^{\lceil (K-1)/2 \rceil} \binom{M+1-2i}{2}, \tag{15}$$

which appears to be much lower than $\binom{M}{K}$ (except when K is very close to M, but in that case, classifier selection becomes useless.). An example is shown in Fig. 13. It shows the number of tests that need to be performed using either exhaustive search (solid line) or iterative selection (bold line), for selecting classifiers from $M=100$ classifiers.

7. Conclusions

This paper presents a new ensemble learning technique in an information theoretic framework. It provides a tool for measuring the goodness of an ensemble by taking into account a trade-off between individual accuracy and diversity. This information theoretic criterion is classifier-independent and only assumes that the combination is done by voting. We propose to use this new measure for selecting an optimal ensemble in a predefined team of classifiers. The experimental results show that our new selection method outperforms standard diversity based selection techniques.

Acknowledgment

This work is supported by the Swiss National Science Foundation through the National Center of Competence in

Research on "Interactive Multimodal Information Management (IM2)".

References

[1] T.G. Dietterich, Ensemble Methods in Machine Learning, in: Lecture Notes in Computer Science, vol. 1857, 2000, pp. 1–15.
[2] Y. Freund, Y. Mansour, R. Schapire, Why averaging classifiers can protect against overfitting, in: Proceedings of the Eighth International Workshop on Artificial Intelligence and Statistics, 2001.
[3] L. Breiman, Bagging predictors, Machine Learning 24 (2) (1996) 123–140.
[4] Y. Freund, R. Schapire, A decision-theoretic generalization of on-line learning and an application to boosting, Journal of Computer and System Sciences 55 (1997) 119–139.
[5] L. Breiman, Random forests, Machine Learning 45 (1) (2001) 5–32.
[6] L.I. Kuncheva, Combining Pattern Classifiers Methods and Algorithms, Wiley, New York, NY, USA, 2004.
[7] J. Kittler, M. Hatef, R.P.W. Duin, J. Matas, On combining classifiers, IEEE Transactions on Pattern Analysis and Machine Intelligence 20 (3) (1998) 226–239.
[8] L.I. Kuncheva, C.J. Whitaker, Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy, Machine Learning 51 (2) (2003) 181–207.
[9] P. Cunningham, J. Carney, Diversity versus quality in classification ensembles based on feature selection, in: European Conference on Machine Learning (ECML), 2000, pp. 109–116.
[10] G. Giacinto, F. Roli, Design of effective neural network ensembles for image classification purposes, Image Vision Computing 19 (9–10) (2001).
[11] C. Shipp, L. Kuncheva, An investigation into how adaboost affects classifier diversity, in: Information Processing and Management of Uncertainty in Knowledge-based Systems (IPMU), Annecy, France, 2002.
[12] L. Kuncheva, C. Whitaker, Using diversity with three variants of boosting: aggressive, conservative, and inverse, in: Proceedings of International Workshop on Multiple Classifier Systems, Lecture Notes in Computer Science, vol. 2364, 2002, pp. 81–90.
[13] C.A. Shipp, L.I. Kuncheva, Relationships between combination methods and measures of diversity in combining classifiers, Information Fusion 3 (2002) 135–148.
[14] S.T. Hadjitodorov, L.I. Kuncheva, L.P. Todorova, Moderate diversity for better cluster ensembles, Information Fusion 7 (3) (2006) 264–275.
[15] G. Brown, J. Wyatt, R. Harris, X. Yao, Diversity creation methods: a survey and categorisation, Journal of Information Fusion 6 (1) (2005) 5–20.
[16] T. Windeatt, Diversity measures for multiple classifier system analysis and design, Information Fusion 6 (2005) 21–36.
[17] P. Melville, R.J. Mooney, Creating diversity in ensembles using artificial data, Information Fusion 6 (1) (2005) 99–111.
[18] A. Krogh, J. Vedelsby, Neural network ensembles, cross validation, and active learning, in: Seventh Conference on Neural Information Processing Systems, 1995, pp. 234–238.
[19] P. Cunningham, Overfitting and diversity in classification ensembles based on feature selection, Technical Report TCD-CS-2000-07, Department of Computer Science, Trinity College Dublin, 2000.
[20] G. Yule, On the association of attributes in statistics, Biometrika 2 (1903) 121–134.
[21] D. Skalak, The sources of increased accuracy for two proposed boosting algorithms, in: AAAI '96 Workshop on Integrating Multiple Learned Models for Improving and Scaling Machine Learning Algorithms, 1996.
[22] R. Kohavi, D. Wolpert, Bias plus variance decomposition for zero-one loss functions, in: ICML, 1996, pp. 275–283.
[23] L.I. Kuncheva, C.J. Whitaker, Ten measures of diversity in classifier ensembles: limits for two classifiers, in: IEE Workshop on Intelligent Sensor Processing, February 2001, IEE.
[24] E.K. Tang, P.N. Suganthan, X. Yao, An analysis of diversity measures, Machine Learning 65 (1) (2006) 247–271.
[25] J.C. Principe, D. Xu, J.W. Fisher, Learning from examples with information theoretic criteria, Journal of VLSI Signal Processing Systems 26 (2000) 61–77.
[26] J. Fisher, III, J. Principe, A methodology for information theoretic feature extraction, in: IEEE International Conference on Neural Networks (IJCNN'98), vol. 3, Anchorage, AK, 1998, pp. 1712–1716.
[27] K.E. Hild II, D. Erdogmus, K. Torkkola, J.C. Principe, Feature extraction using information-theoretic learning, IEEE Transactions on Pattern Analysis and Machine Intelligence 28 (9) (2006) 1385–1392.
[28] V. Sindhwani, S. Rakshit, D. Deodhare, D. Erdogmus, J.C. Principe, P. Niyogim, Feature selection in mlps and svms based on maximum output information, IEEE Transactions on Neural Networks 15 (2004) 937–949.
[29] T. Butz, J.-P. Thiran, From error probability to information theoretic (multimodal) signal processing, Signal Processing 85 (5) (2005) 875–902.
[30] T. Cover, J. Thomas, Elements of Information Theory, Wiley, New York, 1991.
[31] R.M. Fano, Transmission of Information: A Statistical Theory of Communication, MIT Press, Wiley, Cambridge, 1961.
[32] D. Erdogmus, J.C. Principe, Lower and upper bounds for misclassification probability based on renyi's information, Journal of VLSI Signal Processing 37 (2004) 305–317.
[33] L.I. Kuncheva, C.J. Whitaker, C.A. Shipp, R.P.W. Duin, Limits on the majority vote accuracy in classifier fusion, Pattern Analysis and Applications 6 (2003) 22–31.
[34] L. Lam, S.Y. Suen, Application of majority voting to pattern recognition: an analysis of its behavior and performance, IEEE Transactions on Systems, Man, and Cybernetics 27 (1997) 553–568.
[35] Narasimhamurthy, Theoretical bounds of majority voting performance for a binary classification problem, IEEE Transactions on Pattern Analysis and Machine Intelligence 27 (2005) 1988–1995.
[36] D. Ruta, B. Gabrys, A theoretical analysis of the limits of majority voting errors for multiple classifier systems, Pattern Analysis and Applications 5 (4) (2002) 333–350.
[37] L. Shapley, B. Grofman, Optimizing group judgemental accuracy in the presence of interdependencies, Public Choice 43 (1984) 329–343.
[38] J. Meynet, J.-P. Thiran, Information theoretic combination of classifiers with application to AdaBoost, in: 7th international Workshop on Multiple Classifier Systems (MCS), Prague, 2007, ITS.
[39] J. Meynet, J.-P. Thiran, Ensembles of SVMs using an Information Theoretic Criterion, Technical Report, 2008.
[40] R. Duin, P. Juszczak, P. Paclik, E. Pekalska, D. de Ridder, D.M.J. Tax, Prtools4, a matlab toolbox for pattern recognition, Delft University of Technology, 2004.
[41] D.J. Newman, A. Asuncion, UCI machine learning repository, 2007.
[42] L. Breiman, J. Friedman, R.A. Olsen, C.J. Stone, Classification and Regression Trees, Wadsworth International Group, CA, USA, 1984.

About the Author—JULIEN MEYNET was born in Evian-les-Bains, France, in September 1980. He received his Engineering and M.Sc. degree in Electrical Engineering from the Grenoble Institute of Technology (INPG), Grenoble, France in 2003 and his Ph.D. from Swiss Federal Institute of Technology (EPFL), Lausanne, Switzerland in 2007.

He was at EPFL from 2003 to 2008 as assistant researcher and then as postdoctoral researcher in the Signal Processing Institute. He is now R&D engineer at Yahoo!, Grenoble, France.

His main research interests include different theoretical and applied aspects of statistical machine learning with application to computer vision and web search.

About the Author—JEAN-PHILIPPE THIRAN was born in Namur, Belgium, in August 1970. He received the Electrical Engineering degree and the Ph.D. degree from the Université catholique de Louvain (UCL), Louvain-la-Neuve, Belgium, in 1993 and 1997, respectively.

Dr Thiran is Assistant Professor and the leader of the Image Analysis Group at the Signal Processing Institute of the Swiss Federal Institute of Technology (EPFL), Lausanne Switzerland. His current scientific interests include image segmentation, prior knowledge in image analysis, PDE's in image analysis, multimodal signal processing, medical image analysis, including multimodal image registration, segmentation, computer-assisted surgery, diffusion MRI, etc. Prof. Thiran is authors or co-author of more than 75 journal papers and 130 conference papers on image analysis and owns of four international patents.

Please cite this article as: J. Meynet, J.-P. Thiran, Information theoretic combination of pattern classifiers, Pattern Recognition (2010), doi:10.1016/j.patcog.2010.04.013