Accuracy and Diversity in Ensembles of Text Categorisers

Juan José García Adeva[1], **Ulises Cerviño Beresi**[2], and **Rafael A. Calvo**[1]

[1]School of Electrical and Information Engineering
University of Sydney, NSW 2006
Australia
{jjga,rafa}@ee.usyd.edu.au

[2]Instituto de Física Rosario
2000 Rosario, Santa Fe
Argentina
cervino@ifir.edu.ar

Abstract

Error-Correcting Out Codes (ECOC) ensembles of binary classifiers are used in Text Categorisation to improve the accuracy while benefiting from learning algorithms that only support two classes. An accurate ensemble relies on the quality of its corresponding decomposition matrix, which at the same time depends on the separation between the categories and the diversity of the dichotomies representing the binary classifiers. Important open questions include finding a good definition for diversity between two dichotomies and a way of combining all the pairwise diversity values into a single indicator that we call the decomposition quality. In this work we introduce a new measure to estimate the diversity between two learners and we compare it to the well-known Hamming distance. We also examine three functions to evaluate the decomposition quality. We present a set of experiments where these measures and functions are tested using two distinct document corpora with several configurations in each. The analysis of the results shows a weak relationship between the ensemble accuracy and its diversity.
Keywords: Text Categorisation, Document Classification, Ensembles, Diversity Measures

1 Introduction

Monolithic multi-category machine learning algorithms have successfully been used in a number of application areas such as text classification. More recently, there is a significant interest [3, 2, 16] in using ensembles of learning machines as a way of improving the classification accuracy by many binary classification algorithms to solve a multi-category problem.

There are several types of ensemble methodologies [18]. Two of the most popular are boosting [9] and bagging [3], where each learning machine of the ensemble is trained with a different subset of the training set. Mixture of experts [11] is also a well documented technique where the individual predictions provided by each learner are non-linearly combined. Our current work focuses on the Error Correcting Output Codes (ECOC) ensemble decomposition method [7], where the general multi-category classification problem or polychotomy is decomposed into a set of dichotomies, each one of them targeted at a particular subset of categories, with each dichotomy processed by a binary classifier. These category subsets are chosen in a way that a certain amount of prediction errors can be recovered, hence offering an error-correcting ability that helps improve accuracy.

There are two key factors that affect the quality of an ECOC ensemble decomposition: 1) keeping the categories well separated in order to maximise the decision boundaries and 2) having diverse dichotomies to represent the binary classifiers. While the separation of categories has been well studied in the literature [2, 8, 15, 17], a formal definition of diversity is still an open question [12, 13]. For example, there is no definitive definition of diversity, except for being an intuitive measure of the relationship between classifiers.

We contribute here a definition of diversity and an experimental evaluation of its use as a parameter to improve the ensemble's classification accuracy. Besides, we want this global diversity measure to be independent of the separation of categories, as opposed to some related work [12]. This decomposition quality value should provide an indication of how well it will eventually affect its classification accuracy.

The structure of this paper is as follows. Section 2 briefly describes the theoretical framework for ECOC ensembles. Section 3 reviews the pairwise diversity measures and the global diversity measures. Section 4 contains 1) the description of the the empirical design designed to provide global measures that can help indicate the overall accuracy of the ensemble, 2) the experimental framework, including the configurations, software, and document corpora and the results obtained. Section 5 concludes.

2 ECOC Ensembles

Ensembles are sets of implemented instances of machine learning algorithms (learning machines) that work together to improve the performance of the overall system. Instead of having a single and monolithic learner that is trained for all the existing categories of documents, each learner in the ensemble is independently built and specialised in a subset of categories.

Ensembles are often used to transform a multi-class problem (also known as polychotomy) into many two-class problems (also known as dichotomies or binary problems). Therefore, given a corpus $D = \{d_1, d_2, \ldots, d_{|D|}\}$ with a training set $D_t \subset D$, there is also a set of categories $C = \{c_1, c_2, \ldots, c_{|C|}\}$, and a certain number of binary learners $L = \{l_1, l_2, \ldots, l_{|L|}\}$. A bipartition separates the set C into only two categories $\{+1, -1\}$. Because each bipartition serves an individual learner, the set $B = \{b_1, b_2, \ldots\ b_{|L|}\}$ is immediately defined. This enables the establishment of the decomposition matrix $M \in \{+1, -1\}^{L \times C}$.

During the training process, each learner l_j will be presented with all the documents $d_k \in D_t$; if the category (or one of the several categories) of d_k is one of the categories labeled $+1$ in b_j, then $+1$ is used to label this document; otherwise -1 is used. Consequently, each l_j will induct a corresponding function $f_j : \Omega \to \{+1, -1\}$. For example, the first learner l_0 could be responsible for distinguishing between $\{c_0, c_1, c_2\}$ and $\{c_3, c_4\}$ using $+1$ and -1, respectively.

The categorisation process consists of having each document $d_k \in D \setminus D_t$ classified by each learner $l_i \in L$. A vector $V \in \{+1, -1\}$ containing $|L|$ binary predictions will be obtained for each individual classification. The category that best fits the document d_k will be determined by comparing V with the binary representations of each c_i in M and then choosing the closest one.

A simple decomposition scheme is One-to-All, where a bipartition has the form $(i, \bar{i})$, separating C into c_i and its complement. It is evident that in this case the number of learners equals the number of categories ($|L| = |C|$). Figure 1(a) depicts an example of One-to-All decomposition. Another well known design is Pairwise Coupling, where there is a binary learner for each possible pair of (different) categories, providing $|L| = \frac{|C|(|C|-1)}{2}$. A bipartition (i, j) with $i > j$ separates the categories c_i and c_j while ignoring all the rest. That is why in this exceptional case $M \in \{+1, -1, 0\}^{L \times C}$. Figure 1(b) shows an example of Pairwise Coupling decomposition.

	c_0	c_1	c_2	c_3	c_4
l_0	+1	−1	−1	−1	−1
l_1	−1	+1	−1	−1	−1
l_2	−1	−1	+1	−1	−1
l_3	−1	−1	−1	+1	−1
l_4	−1	−1	−1	−1	+1

(a) One-to-All

	c_0	c_1	c_2	c_3	c_4
l_0	+1	−1	0	0	0
l_1	+1	0	−1	0	0
l_2	+1	0	0	−1	0
l_3	+1	0	0	0	−1
l_4	0	+1	−1	0	0
l_4	0	+1	0	−1	0
l_6	0	+1	0	0	−1
l_7	0	0	+1	−1	0
l_8	0	0	+1	0	−1
l_9	0	0	0	+1	−1

(b) Pairwise Coupling

	c_0	c_1	c_2	c_3	c_4
l_0	+1	+1	+1	−1	−1
l_1	+1	+1	−1	−1	+1
l_2	+1	−1	+1	+1	−1
l_3	−1	+1	−1	+1	−1
l_4	+1	−1	−1	+1	+1
l_5	−1	−1	+1	−1	+1
l_6	−1	+1	+1	+1	+1

(c) ECOC

Figure 1: Decomposition matrices for a 5-category problem using different decomposition schemes.

A more general and appealing architecture is Error Correcting Codes (ECOC) decomposition. ECOC is very well known in the area of communications for correcting data errors during transmission. It is based on adding some redundant information to the block to be transmitted, hence obtaining a *codeword*. Even if during transmission the codeword is affected by some interference due to a noisy channel, the errors may still be corrected at the receiving end. This is achieved by all possible pairs of codewords having a certain separation, which is measured by the Hamming distance. This quantity measures how many bits of information are different between two binary strings of equal length. Therefore, given $a, b \in \{+1, -1\}$ with $|a| = |b|$, the Hamming distance between a and b is defined as

$$\mathrm{h}(a, b) = \sum_{i=1}^{|a|} \frac{|a_i - b_i|}{2} \tag{1}$$

In the context of ensembles, the use of ECOC-generated codewords to represent the categories in M is known as ECOC decomposition. The error correction capability can be calculated by $e = \frac{\mathrm{h}-1}{2}$. This has an important implication because while for One-to-All $e = \frac{2-1}{2} = 0$ and for Pairwise Coupling $e = \frac{2(|C|-2)-1}{2} = |C| - 3$, in the case of ECOC e varies according to h. It is precisely in this configuring ability that the main advantage of using ECOC occurs.

3 Diversity Measures

The classification accuracy provided by the ensemble depends mainly on the following factors: the individual accuracies that correspond to the binary classifiers, the complexity of the multi-category classification problem, and the quality of the decomposition. As stated before, in this work we are interested in the latest. Accordingly, the quality of the decomposition depends on the pairwise separation between the existing categories (e.g. represented by the columns of the decomposition matrix) and the pairwise diversity between the binary classifiers (e.g. represented by the rows of the decomposition matrix). Diverse dichotomies are assumed to provide diverse binary classifiers, therefore reducing the chance of having correlated predictions. On the one hand, in this section we present the two functions used to measure diversity between two dichotomies. On the other hand, using those two functions, we propose several global measures that combine all the possible

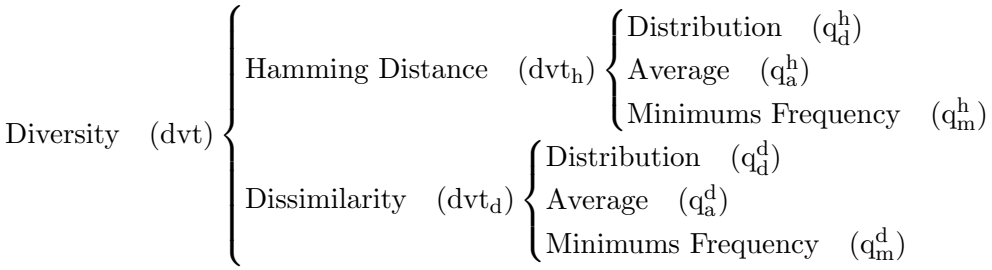


Figure 2: Summary of the pairwise diversity measures and the decomposition quality functions

pairwise diversity measures into a single general value. This value is used as an overall indication of the diversity quality of the dichotomies in the ensemble.

The first diversity function is the Hamming distance as defined by Equation 1. This function is traditionally used to measure the error-correcting capability of a code, while in the case of ECOC ensembles it is employed to maximise the pairwise distance between categories, as explained in Section 2. However, the Hamming distance can also be used to measure diversity between dichotomies [10]. Two dichotomies are similar (i.e. there is no diversity whatsoever) when the corresponding Hamming distance is either 0 or a maximum). Therefore, measuring the diversity between two dichotomies by means of the Hamming distance can be expressed by

$$\mathrm{dvt_h}(l_i, l_j) = \mod(\mathrm{h}(l_i, l_j), |C|), \tag{2}$$

where $i, j \in \{1, \ldots, |L|\}$ and the maximum diversity value corresponds to $|C| - 1$.

The second diversity function introduced is Dissimilarity, which is based on the Disagreement function found in [12, 13]. Dissimilarity takes into account the fact that two complementary dichotomies hold a null diversity while assessing how similar they are. It can be represented by

$$\mathrm{dvt_d}(l_i, l_j) = \min\{\mathrm{h}(l_i, l_j), |C| - \mathrm{h}(l_i, l_j)\}, \tag{3}$$

where $i, j \in \{1, \ldots, |L|\}$ and its maximum diversity value can be $|C|/2$.

The closer $\mathrm{dvt}(l_i, l_j)$ is to 0 the less diverse the two dichotomies l_i and l_j are. The collection of diversity values that correspond to the all possible pairs of dichotomies can be expressed by

$$D = \{\mathrm{dvt}(l_0, l_1), \mathrm{dvt}(l_0, l_2), \ldots, \mathrm{dvt}(l_1, l_2), \mathrm{dvt}(l_1, l_3), \ldots, \mathrm{dvt}(l_{|L|-2}, l_{|L|-1})\}, \tag{4}$$

where the number of elements in D is

$$|D| = \frac{|L|(|L| - 1)}{2} \tag{5}$$

We propose three different functions in order to estimate the decomposition quality using the diversity measurements between all the possible pairs of dichotomies. These three functions provide a value in the range between 0 and 1, meaning that the closer to 1 the more diverse the dichotomies of the ensemble are. Figure 2 shows a summary of these measures.

The first global measure is named $\mathrm{q_d}$, for quality of diversities' distribution, and is intended to provide a value that quantifies both the arithmetic mean diversity as well as how these diversities are distributed among all the pairs. It is expressed by

$$\mathrm{q_d}(D) = 1.0 - \frac{\mathrm{dvt}_{max} - \bar{D} + \sigma(D)}{\mathrm{dvt}_{max}}, \tag{6}$$

Table 1: Some diversity measurements and their corresponding quality values

Pairwise Diversities	Quality Values		
	$\mathrm{q_d}(D_i)$	$\mathrm{q_a}(D_i)$	$\mathrm{q_m}(D_i)$
$D_1 = \{2, 2, 3, 1\}$	0.3	0.13	0.38
$D_2 = \{1, 4, 2, 1\}$	0.23	0.13	0.44
$D_3 = \{2, 2, 2, 2\}$	0.5	0.13	0.63
$D_4 = \{3, 3, 4, 1\}$	0.37	0.17	0.38
$D_5 = \{3, 3, 3, 3\}$	0.75	0.19	0.69

where $\bar{D} = \frac{1}{|D|}\sum_{i=1}^{|D|} D_i$, $\sigma(D) = \sqrt{\frac{\sum_{i=1}^{|D|}(D_i - \bar{D})^2}{|D|-1}}$, and dvt_{max} is the maximum possible value of diversity. This measure is based on the assumption that it is better to have a high average and uniform distribution of diversities in as many dichotomy pairs as possible, as opposed to a skewed distribution of pairwise diversities where some pairs have good diversity whereas others have low values.

The second global measure is called $\mathrm{q_a}$, for quality of average diversity, and simply focuses on favouring large average diversities. It can be calculated by

$$\mathrm{q_a}(D) = \frac{\bar{D}}{|D|\,\mathrm{dvt}_{max}}, \tag{7}$$

The third global measure $\mathrm{q_m}$ calculates the quality of minimum diversities. It pays attention to the minimum diversity value in D as well as its frequency. The lower the minimum diversity value and the more times this value is present, the worse is considered to be the quality of D. This can be expressed by

$$\mathrm{q_m}(D) = \frac{\min D + |D| + \mathrm{f}(D, \min D)}{|D|\,\mathrm{dvt}_{max}}, \tag{8}$$

where $\min D$ is the minimum diversity value found in D and $\mathrm{f}(D, \min D)$ is the number of times that $\min D$ is observed in D.

Table 1 offers some examples, including the quality values corresponding to five different diversity sets. It is clear that D_5 is the best set of diversities, followed by D_3 or D_4, depending on which quality function is applied.

4 Empirical Evaluation

4.1 Approach

We have defined three functions that describe decomposition quality and applicable to the two diversity measures and we performed experiments applying these functions to distinct ensembles. Although somewhat similar empirical analyses have already been performed elsewhere [12, 4, 13] we believe that, because they also take into account the fluctuating and unknown pairwise separation between categories, they may fail to find the true relationship between diversity and ensemble accuracy. We advocate the analysis of the diversities when this measure is isolated and independent of the categories separation. This means that when creating the decomposition matrix that corresponds to the ensemble to be analysed, the separation between the categories must be a known parameter to the decomposition matrix building method, so that all the matrices can be built with the same categories' separation, and thus the diversity can be studied as an isolated measure.

Table 2: Experimental Measurement Ranges for *ModApté*

Measure	Binary Classifiers			
	63	127	255	511
q_d^h	[0.466,0.490]	[0.452,0.472]	[0.453,0.464]	[0.453,0.461]
q_a^h	[0.501,0.509]	[0.499,0.507]	[0.5,0.506]	[0.45,0.505]
q_m^h	[0.009,0.009]	[0.009,0.448]	[0.009,0.421]	[0.009,0.395]
q_d^d	[0.914,0.954]	[0.875,0.913]	[0.876,0.894]	[0.876,0.885]
q_a^d	[0.979,0.985]	[0.947,0.957]	[0.939,0.944]	[0.935,0.94]
q_m^d	[0.018,0.018]	[0.018,0.877]	[0.018,0.789]	[0.018,0.772]
F_1^M	[0.267,0.375]	[0.313,0.447]	[0.347,0.515]	[0.391,0.534]
F_1^μ	[0.751,0.792]	[0.779,0.821]	[0.808,0.835]	[0.828,0.842]

There are several approaches suitable for generating the codewords needed to build the ECOC decomposition matrices. Examples include the exhaustive method [7], randomly generated codes [17], and the algebraic-based BCH codes [14]. The exhaustive method has the limitation of the Hamming equidistance of codewords being always 8 [7]. The random method is computationally expensive and even unaffordable when the number of categories is relatively large (above 10 or so). We chose BCH codes, a method based on algebraic techniques from Galois Field theory due to its good scalability for hundreds or thousands of categories and its ability to generate ECOC codewords separated by a minimum, configurable Hamming equidistance.

A BCH code is expressed by (n, k), where k is the number of bits necessary to generate a codeword of length n with an error-correcting capability of t bits that offers a minimum Hamming distance $h_{min} = 2t + 1$. Given a number of categories $|C|$, the number of bits required to represent them is $k = log_2|C|$. This value of k is combined with a desired error-correcting capability of at least t bits (i.e. the desired Hamming distance between categories) to establish the value of n as well as the closest possible values of $\hat{k}$ and $\hat{t}$ that suit the given k and t. As a result, n determines the number of dichotomies in the decomposition matrix, which is the same as how many binary classifiers will form the ensemble. Because the number of categories that can be represented by $\hat{k}$ is $|\hat{C}| = 2^{\hat{k}}$, then, in most cases, $|\hat{C}| > |C|$. Accordingly, the *maximum* decomposition matrix that can be obtained is $\hat{M} = L \times \hat{C}$. The set of possible decomposition matrices $L \times C$ with $C \subset \hat{C}$ that can be extracted from $\hat{M}$ is $T = \{M_1, M_2, \ldots, M_{|T|}\}$, where

$$|T| = \binom{|\hat{C}|}{|C|} = \frac{|\hat{C}|!}{|C|!\ (|\hat{C}| - |C|)!} \tag{9}$$

In most cases, $|T|$ is going to be a very large value. This means that we enjoy a huge amount of different decomposition matrices where the separation between categories is virtually constant and adjusted by us. This framework allows us to apply the diversity measures depicted in Section 3 to a large number of different decomposition matrices, collect the results, and analyse them, with the certainty that in all cases the separation between categories does not influence the diversity measurements.

4.2 Experimental Configuration

We selected two very distinct document corpora as the basis of our experiments: *ModApté* [6], a well studied collection of Reuters newstories and *La Capital* [5], a collection of news in Spanish.

We selected 4 different ensemble sizes for *ModApté* and 5 for *La Capital*. These ensemble sizes

Table 3: Experimental Measurement Ranges for *La Capital*

Measure	Binary Classifiers				
	31	63	127	255	511
q_d^h	[0.324,0.451]	[0.367,0.429]	[0.374,0.415]	[0.375,0.408]	[0.375,0.405]
q_a^h	[0.471,0.552]	[0.489,0.542]	[0.499,0.536]	[0.499,0.534]	[0.499,0.532]
q_m^h	[0.06,0.427]	[0.062,0.371]	[0.062,0.312]	[0.062,0.311]	[0.062,0.25]
q_d^d	[0.619,0.785]	[0.687,0.747]	[0.697,0.725]	[0.698,0.715]	[0.7,0.709]
q_a^d	[0.83,0.907]	[0.843,0.89]	[0.848,0.874]	[0.849,0.865]	[0.85,0.862]
q_m^d	[0.121,0.743]	[0.125,0.736]	[0.125,0.623]	[0.125,0.5]	[0.125,0.499]
F_1^M	[0.441,0.637]	[0.484,0.693]	[0.514,0.686]	[0.506,0.681]	[0.497,0.698]
F_1^μ	[0.691,0.789]	[0.725,0.82]	[0.754,0.805]	[0.739,0.803]	[0.738,0.818]

(i.e. 511, 255, 127, 63, 31) were determined by both the desired separation between categories and the number of existing categories as explained in Section 4.1. For each ensemble size in each corpus (i.e. each one of the 9 different scenarios) we generated 1,000 different decomposition matrices. Each of these decomposition matrices corresponded to an ensemble used to train and test the respective corpus documents. Multinomial Naïve Bayes was the learning algorithm used for these binary learners due to the quality of the individual learners [19]. In each one of these 9,000 executions we collected eight results that include: q_d^h, q_a^h, q_m^h, q_d^d, q_a^d, q_m^d, F_1^M, and F_1^μ.

The very well-known Reuters-21578 can be considered a standard corpus used for benchmarking in Text Categorisation [6]. It has 21,578 news articles written in English that appeared in the Reuters newswire circa 1987. The total number of categories is 135, allowing for the possibility of the same document being categorised into more than one category. The *ModApté* split makes reference to a particular partitioning of the corpus into two subsets, one for training with 9,603 documents and the rest for testing including 3,299 documents. This split comprises 115 categories. Please note the existence of two additional subcollections of the original Reuters-21578 corpus: one containing only 10 categories with a higher number of training documents and another with 90 categories. These two splits have not been considered in our experiments.

The difficulty of finding document corpora in Spanish motivated us to construct a corpus derived from the original *La Capital* news articles [5]. The original database is composed of approximately 75,000 newspaper articles written in Spanish and represented in XML. The number of categories is 19 with one category per document and a very skewed distribution of documents over categories. For instance, the category *sports* contains more than 22,000 documents while *arts* contains only 20.

We did a preliminary removal of both numeric characters and stop-words to the documents in both corpora. We also applied a stemming process to the words in order to remove the most common morphological and inflectional endings. With two different languages in use, in the case of the *La Capital* corpus a dictionary-based technique was used, whereas for the *ModApté* corpus we used Porter's rule-set.

Feature selection was performed using the function $\chi^2(t_i, c_j)$ that measures the dependence of the category c_j on the occurrence of the term t_i. Using χ^2 produced better results than other feature selection functions such as Term Frequency, Document Frequency, or Information Gain. For the *La Capital* corpus, the number of features selected was 7,000 (this corresponds to a reduction factor $\xi = 0.75$). After several attempts, we determined that this number of features was close to optimum for this particular corpus. For the *ModApté* corpus the number of features selected was approximately 7,000, which corresponds to $\xi = 0.88$.

The feature vectors were built using a sparse representation that favours the efficient usage of memory. Each feature was weighted by means of the function TF/IDF(t_i, d_j), which is based on the assumption that those terms occurring in more documents have little discriminatory strength among categories.

For the *La Capital* corpus, 80% of the existing documents in an arbitrarily chosen month were used for training and 20% for testing. The documents were randomly chosen for each of several executions and eventually averaged. For the *ModApté* corpus, the two appropriate subsets of documents were used for training and testing, as explained above.

All the experiments were performed on the Awacate framework [1]. Awacate is an object-oriented software framework for Text Categorisation developed by the authors of this paper. It was written in Java and is available as Open Source. Awacate includes several learning algorithms such as Naïve Bayes, Rocchio, SVM, and kNN; ensembles using the decomposition methods One-to-All, Pairwise Coupling, and ECOC; preprocessing of documents is capable of handling texts in English, French, German, Spanish, and Basque; evaluation of results including category-specific TP_i, FP_i, FN_i, π_i, ρ_i, F_{1_i} and averaged π^μ, ρ^μ, F_1^μ, π^M, ρ^M, F_1^M, as well as partitioning of the testing space using n-fold Cross Validation.

4.3 Experimental Results

Because of the large volume of results generated, we decided to display two sets of measurement ranges, one for each corpus. Table 2 shows the ranges of diversity measurements obtained for the *ModApté* corpus while Table 3 is the equivalent for the *La Capital* corpus. Each one of these two tables contains eight ranges of measurement obtained in the experiments for each ensemble of binary classifiers. Furthermore, Figure 3 contains six different graphs regarding the results for *ModApté* with 255 binary classifiers while Figure 4 corresponds to *La Capital* with 63 binary classifiers.

In general, these graphs illustrate the relationships between classification accuracy in terms of F_1^μ and the decomposition quality functions $\mathrm{q_d}$, $\mathrm{q_a}$, and $\mathrm{q_m}$. Each one of these three functions are applied to the two diversity measures including the Hamming distance $\mathrm{dvt_h}$ and Dissimilarity $\mathrm{dvt_d}$, therefore producing six different graphs. Please, refer to Figure 2 for further details.

In the case of Figure 3, there seems to be a very slight visible relationship between diversity and accuracy when looking at both Figure 3(a) and Figure 3(d). The first graph corresponds to the relationship between F_1^μ and $\mathrm{q_d^h}$ or, in other words, the quality function Distribution (i.e. $\mathrm{q_d}$) applied to the diversity measure of Hamming Distance (i.e. $\mathrm{dvt_h}$). The second graph measures the diversity with $\mathrm{q_d^d}$, which uses again the quality function $\mathrm{q_d}$ in combination with the Dissimilarity measure (i.e. $\mathrm{dvt_d}$). The other fours graphs does not seem to provide any apparent visible relationship between accuracy and diversity. The graphs found in Figure 3(b), Figure 3(c), and Figure 3(f) show that the values of diversity are concentrated around two groups of values, hence not providing a continuous, useful relationship. There is a particular case in Figure 3(e), which focuses on the relationship between accuracy and diversity using $\mathrm{q_a^d}$, as it illustrates the same classification accuracy no matter what the diversity is.

On the other hand, and taking into account Figure 4, Figure 4(d) is the only instance that shows a slight visible relationship between diversity and accuracy. The rest of the graphs do not seem to replicate this situation. Figure 4(a), Figure 4(b), and Figure 4(e) exhibit the existence of several very narrow ranges of diversity measurements, with values centred around certain values. Figure 4(c), Figure 4(c) show that there is a small number of different diversity values and they are uniformly distributed along the range of accuracy measurements.

When comparing the two sets of results displayed in Figure 3 and Figure 4, it can be noted that the only relationship that are visually similar are those found in Figure 4(d) and Figure 3(d).

Both show the same almost linear relationship between the ensemble classification accuracy F_1^{μ} and its diversity measures q_d^d that pertains to the Distribution quality function (i.e. q_d) applied to the Dissimilarity measure (i.e. dvt_d).

5 Conclusions

The research community has indicated that the diversity of classifiers in an ensemble has an impact on the accuracy. The definition of a specific measure of diversity has yet to be achieved. The lack of such a definition has made it hard to evaluate what is the actual correlation between these variables.

This paper contributes a definition of diversity together with an experimental evaluation. In these experiments we have studied the relationship between three ways of combining pairwise values of two different diversity measures. Results suggest that, while some of the evaluation functions (i.e. q_m and q_a) do not provide any significant prediction capability on the ensemble accuracy, there is one function (i.e. q_d) that does. Moreover, both diversity measures can be applied to this function, although Dissimilarity (i.e. dvt_d) provided a slightly better estimation than Hamming Distance (i.e. dvt_h). Unfortunately q_d seems to be insufficient for predicting how well an ECOC ensemble will perform over a corpus of documents.

Our future work will include further analysis of combination of diversity values and how they are reflected on ensemble accuracy. We believe that finding a direct relationship between ensemble diversity and accuracy will have a major impact on ensemble research.

References

[1] J. J. García Adeva. Awacate: Towards a Framework for Intelligent Text Categorisation in Web Applications. Technical report, University of Sydney, 2004.

[2] A. Berger. Error-correcting output coding for text classification. In *Proceedings of IJCAI*, 1999.

[3] L. Breiman. Bagging predictors. In *Machine Learning*, volume 24, pages 123–140, 1996.

[4] Gavin Brown, Jeremy Wyatt, Rachel Harris, and Xin Yao. Diversity Creation Methods: A Survey and Categorisation. *Journal of Information Fusion*, 2004.

[5] Ulises Cerviño Beresi, J. J. García Adeva, Rafael A. Calvo, and Alejandro H. Ceccatto. Automatic classification of news articles in Spanish. In *Actas del Congreso Argentino de Ciencias de Computación (CACIC)*, pages 1588–1600, 2004.

[6] Franca Debole and Fabrizio Sebastiani. An analysis of the relative hardness of Reuters-21578 subsets. In *Proceedings of LREC-04, 4th International Conference on Language Resources and Evaluation*, pages 971–974, Lisbon, PT, 2004.

[7] Thomas G. Dietterich and Ghulum Bakiri. Solving multiclass learning problems via error-correcting output codes. *Journal of Artificial Intelligence Research*, 2:263–286, 1995.

[8] F. F. Masulli and G. Valentini. Effectiveness of Error Correction Output Coding Decomposition Schemes in Ensemble and Monolithic Learning Machines. *Pattern Analysis and Application Journal*, 6:285–300, 2003.

[9] Y. Freund and R. Schapire. Experiments with a new boosting algorithm. In *Thirteeth Int. Conf. on Machine Learning*, pages 148–156, 1996.

[10] J. J. García Adeva and Rafael A. Calvo. A Decomposition Scheme based on Error-Correcting Output Codes for Ensembles of Text Categorisers. In *Third International Conference on Information Technology and Applications*, volume Vol. I, pages 375–378. IEEE Computer Society, 2005.

[11] R. A. Jacobs. Methods for combining experts probability assessment. *Neural Computation*, 7:867–888, 1995.

[12] L. Kuncheva and C. Whitaker. Measures of Diversity in Classifier Ensembles and Their Relationship with the Ensemble Accuracy. *Machine Learning*, (51):181–207, 2003.

[13] Kuncheva L.I. Using diversity measures for generating error-correcting output codes in classifier ensembles. *Pattern Recognition Letters*, (26):83–90, 2005.

[14] Shu Lin and Daniel J. Costello. *Error Control Coding, Second Edition.* Prentice-Hall, Inc., 2004.

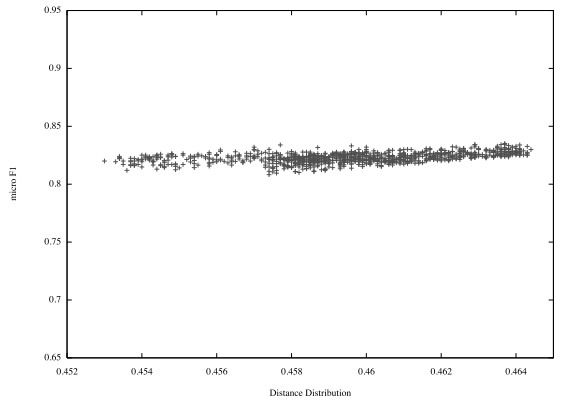
[15] F. Masulli and G. Valentini. An experimental analysis of the dependence among codeword bit errors in ECOC learning machines. *Neurocomputing*, (57C):189–214, 2004.

[16] Dietterich T. Ensemble methods in Machine Learning. In *Proceedings of the First International Workshop on Multiple Classifier Systems*, pages 1–15, 2000.

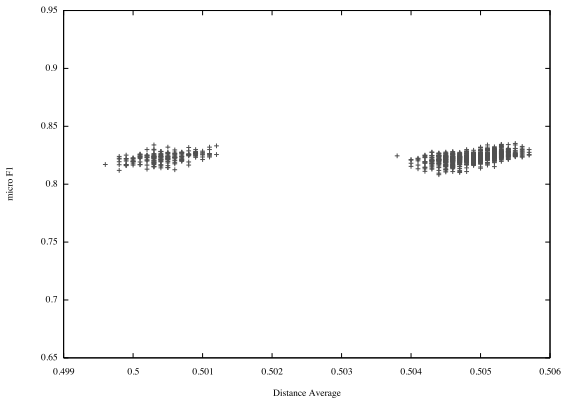
[17] Terry Windeatt T. and Ghaderi R. Coding and decoding strategies for multi-class learning problems. *Information Fusion*, (4(1)):11–21, 2003.

[18] G. Valentini and F. Masulli. Ensembles of Learning Machines. *Neural Nets WIRN Vietri-02, Series Lecture Notes in Computer Sciences*, 2486:3–19, 2002.

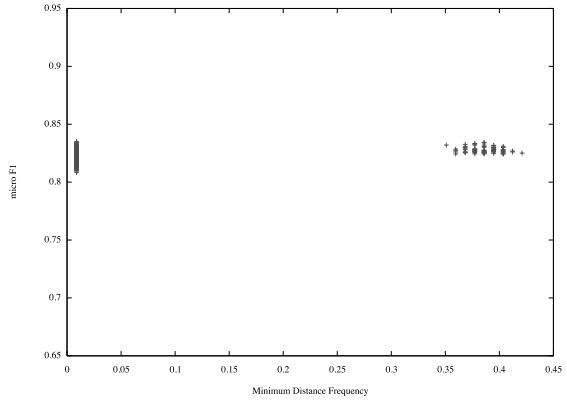
[19] Y. Yang and X. Liu. A re-examination of text categorization methods. In *22nd Annual International SIGIR*, pages 42–49, Berkley, August 1999.



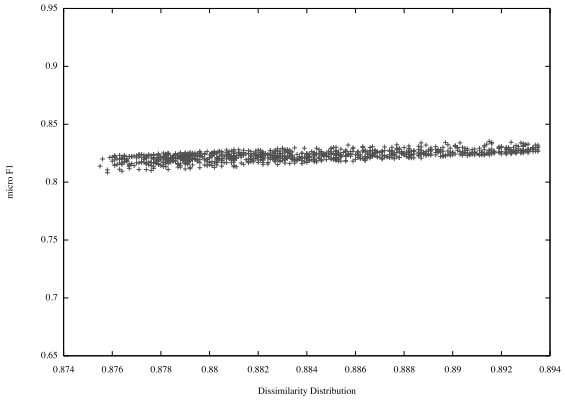
(a) Accuracy vs Diversity based on Distance Distribution



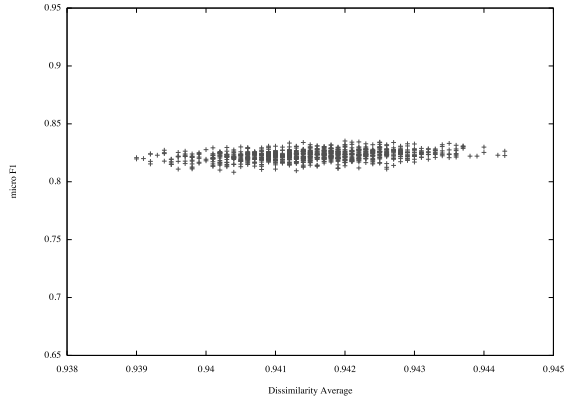
(b) Accuracy vs Diversity based on Distance Average



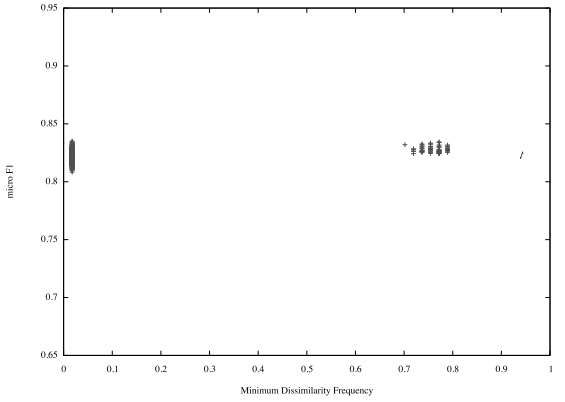
(c) Accuracy vs Diversity based on Minimum Distance Frequency



(d) Accuracy vs Diversity based on Dissimilarity Distribution

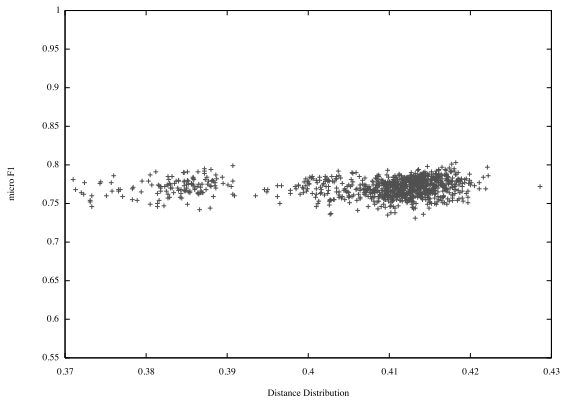


(e) Accuracy vs Diversity based on Dissimilarity Average

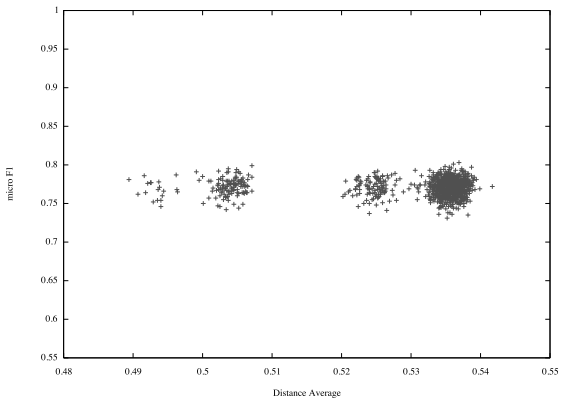


(f) Accuracy vs Diversity based on Minimum Dissimilarity Frequency

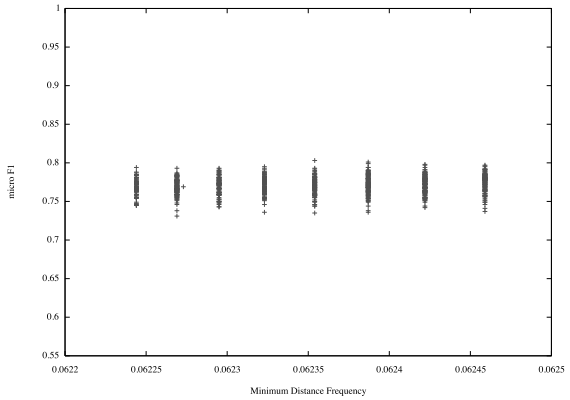
Figure 3: Results for *ModApté* using 255 learners.



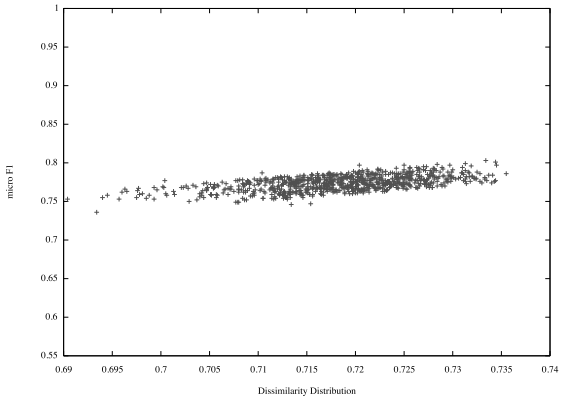
(a) Accuracy vs Diversity based on Distance Distribution



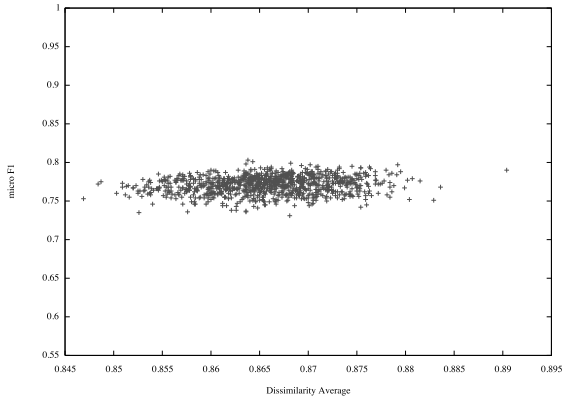
(b) Accuracy vs Diversity based on Distance Average



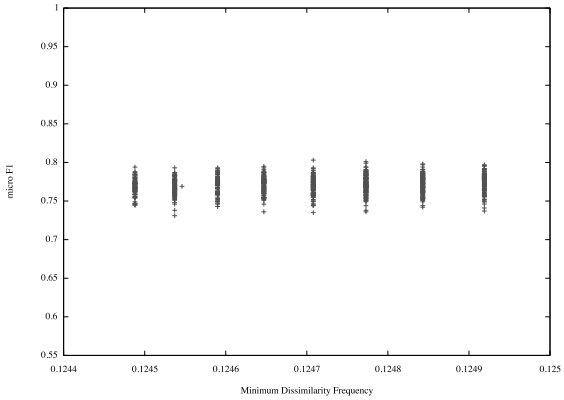
(c) Accuracy vs Diversity based on Minimum Distance Frequency



(d) Accuracy vs Diversity based on Dissimilarity Distribution



(e) Accuracy vs Diversity based on Dissimilarity Average



(f) Accuracy vs Diversity based on Minimum Dissimilarity Frequency

Figure 4: Results for *La Capital* using 63 learners.