Diversity Analysis on Boosting Nominal Concepts

Nida Meddouri[1], Héla Khoufi[1], and Mondher Sadok Maddouri[2]

[1] Research Unit on Programming, Algorithmics and Heuristics - URPAH,
Faculty of Science of Tunis - FST,
Tunis - El Manar University
{nida.meddouri,hela.khoufi}@gmail.com
[2] College of Community, Hinakiyah
Taibah University - Medinah Monawara
Kingdom of Saoudi Arabia
maddourimondher@yahoo.fr

Abstract. In this paper, we investigate how the diversity of *nominal classifier* ensembles affects the *AdaBoost* performance [13]. Using 5 real data sets from the *UCI Machine Learning Repository* and 3 different diversity measures, we show that *Q Statistic* measure is mostly correlated with *AdaBoost* performance for 2-class problems. The experimental results suggest that the performance of *AdaBoost* depend on the *nominal classifier* diversity that can be used as a stopping criteria in ensemble learning.

1 Introduction

Boosting is an adaptive approach, which makes it possible to correctly classify an object that can be badly classified by an ordinary classifier. The main idea of *Boosting* is to build many classifiers who complement each other, in order to build a more powerful classifier. *Adaboost* (***Ada**ptive **Boost**ing*) is the most known method of *Boosting* for classifiers generation and combination.

AdaBoost algorithm is iterative. At first, it selects a subset of instances from the learning data set (different subset from the training data set in each iteration). Then, it builds a classifier using the selected instances. Next, it evaluates the classifier on the learning data set, and it starts again T times.

It has been found that this ingenious manipulation of training data can favorise diversity especially for *linear* classifiers [11]. However, there is no study concerning the role of diversity on *Nominal Concepts* classifiers [13]. In this paper, we study how diversity changes according to the *nominal classifier* numbers and we show when adding new classifiers to the team can't provide further improvements.

This paper is organized as follows: section 2 presents the principle of *Classifier of Nominal Concepts* (***CNC***) used in *Boosting* [7,13]. In section 3, we discuss the diversity of classifiers and the different measures that can be exploited in the classifiers ensembles generation. Section 4 presents the experimental results that prove when the diversity can be useful in *Boosting of Nominal Concepts*.

P.-N. Tan et al. (Eds.): PAKDD 2012, Part I, LNAI 7301, pp. 306–317, 2012.

2 Boosting of CNC

CNC is a classifier based on the *Formal Concept Analysis*. It is distinguished from the other *Formal Concept Analysis* methods by handling nominal data. It generates *Nominal Concept* that is used as classification rule. Comparative studies and experimental results have proved the benefits of *CNC* compared to existing ones (*GRAND, RULEARNER, CITREC, IPR*) [13].

2.1 Nominal Concepts

A nominal classifier can be build using the whole of training instances $\mathcal{O} = \{o_1, ..., o_N\}$ described by L nominal attributes $\mathcal{AN}$ (which are not necessary binary).

$$AN = \{AN_l | l = \{1, .., L\}\}. \tag{1}$$

At first, the pertinent nominal concept AN^* is extracted from the training instances by selecting the nominal attribute which minimises the measure of *Informational Gain* [13]. Then, the associated instances are selected with each value v_j (j = {1,..,J} and J the number of different value of a nominal attribut) from this attribute as $\delta(AN^* = v_j)$. The δ operator is defined by:

$$\delta(AN^* = v_j) = \{o \in O | AN^*(o) = v_j\}. \tag{2}$$

Then, the other attributes describing all the extracted instances are determined (using the closure operator $\delta \circ \varphi$ $(AN^* = v_j)$) as follows:

$$\varphi(B) = \{v_j | \forall o, o \in B \ and \ \exists AN_l \in AN | AN_l(o) = v_j\}. \tag{3}$$

In [13], a method called *BNC* (***B**oosting **N**ominal **C**oncepts*) has been proposed. The advantage of *BNC* is to build a part of the lattice covering *the best nominal concepts* (*the pertinent*) which is used as classification rules (the *Classifier Nominal Concepts*). The *BNC* has the particularity to decide the number of *nominal classifiers* in order to control the time of application and to provide the best decision.

2.2 Learning Concept Based Classifiers

For K-class problem, let $Y = \{1, .., K\}$ the class labels, with $y_i \in Y$ is the class label associated for each instance o_i (*i=1 to N*). To generate T classifiers in *AdaBoost*, the distribution of the weight of o_i is initially determined as :

$$D_0(i) = (1/N). \tag{4}$$

The weight of o_i is:

$$w^1_{i,y} = D_0(i)/(K-1) \ for \ each \ y \ \in \ Y - \{y_i\} \tag{5}$$

On each iteration t from 1 to T, we define:

$$W_i^t = \sum_{y \neq y_i} w_{i,y}^t \quad and \quad we \quad set \quad q_t(i,y) = \frac{w_{i,y}^t}{W_i^t} \; for \; each \; y \neq y_i \tag{6}$$

The distribution of weights is calculated by:

$$D_t(i) = \frac{W_i^t}{\sum_{i=1}^{N} W_i^t} \tag{7}$$

Each generated nominal classifier h_t provides an estimated probability $p_t(o_i,y_i)$ to the class y_i from the entry o_i. Three cases are presented:

- If $p_t(o_i,y_i) = 1$ and $p_t(o_i,y) = 0$, $\forall y \neq y_i$, h_t has correctly predicted the class of o_i.
- If $p_t(o_i,y_i) = 0$ and $p_t(o_i,y) = 1$, $\forall y \neq y_i$, h_t has an opposed prediction of the class of o_i.
- If $p_t(o_i,y_i) = p_t(o_i,y)$, $\forall y \neq y_i$, the class of o_i is selected randomly (y or y_i).

The error rate of h_t is calculated on the weighted training set. If an instance o_i is correctly classified by h_t, then the weight of this instance is reduced. Otherwise, the weightis increased. The pseudo-loss of the classifier h_t is defined as:

$$\epsilon_t = 0.5 \times \sum_{i=1}^{N} D_t(i)(1 - p_t(o_i, y_i) + \sum_{y \neq y_i} q_t(i,y) p_t(o_i, y)) \tag{8}$$

The weights are then updated according to β_t:

$$\beta_t = \varepsilon_t / (1 - \varepsilon_t) \tag{9}$$

The procedure is repeated T times and the final result of *BNC* is determined via the combination of the generated classifier outputs:

$$h_{fin}(o_i) = \arg\max_{y \in \mathcal{Y}} \sum_{t=1}^{T} log(1/\beta_t) \times p_t(o_i, y_i). \tag{10}$$

The first variant of the *AdaBoost* algorithm is called *Adaboost.M1* [5,6] that uses the previous process and stops it when the error rate of a classifier becomes over 0.5. The second variant is called *AdaBoost.M2* [6] which has the particularity of handling multi-class data and operating whatever the error rate is. In this study, we use *AdaBoost.M2* since *Adaboost.M1* has the limit to stop *Boosting* if the learning error exceeds 0.5. In some experiments, Adaboost.M1 can be stopped after the generation of first classifier thus we cannot calculate the diversity of classifier ensemble in this particular case.

Recent researches have proved the importance of classifier diversity in improving the performance of *AdaBoost* [1,4,8]. We shall discuss about that in the next section.

3 Classifier Diversity

According to [4], *linear classifiers* should be different from each other, otherwise the decision of the ensemble will be of lower quality than the individual decision. This difference, also called *diversity*, can lead to better or worse overall decision [3].

In [14], the authors found a consistent pattern of diversity showing that at the beginning, the generated *linear classifiers* are highly diverse but as the learning progresses, the diversity gradually returns to its starting level. This suggests that it could be beneficial to stop *AdaBoost* before diversity drops. The authors confirm that there are a consistent patterns of diversity with many measures using *linear classifiers*. However, they report that the pattern might change if other classifier models are used. In the paper, we will prove that this pattern is the same with *nominal classifiers*.

Many measures can be used to determine the diversity between classifiers [11]. In this section, we present three(3) of them: $\mathcal{Q}$ *Statistic*, *Correlation Coefficient* ($\mathcal{CC}$) and *Pairwise Interrater Agreement* (*kp*). These pairwise measures have the same diversity value (0) when the classifiers are statistically independent. They are called pairwise because they consider the output classifiers, two at a time and then they average the calculated pairwise diversity. These measures are computed based on the agreement and the disagreement between each 2 classifiers (see Table 1).

Table 1. The agreement and disagreement between two classifiers

	h_k correct (1)	h_k incorrect (0)
h_j correct (1)	N^{11}	N^{10}
h_j incorrect (0)	N^{01}	N^{00}

$$N=N^{00}+N^{01}+N^{10}+N^{11}$$

N^{vw}(v,w=1,0) is the number of instances correctly or incorrectly classified by the two classifiers: h_j and h_k ($j,k = 1..T$).

The Q Statistic: Using table 1, this measure is calculated as follows:

$$\mathcal{Q}_{j,k} = \frac{N^{11}N^{00} - N^{10}N^{01}}{N^{11}N^{00} + N^{10}N^{01}} \tag{11}$$

$\mathcal{Q}$ *Statistic* varies between -1 and 1. Classifiers that tend to recognize correctly the same instances, have positive $\mathcal{Q}$ values, and classifiers that commit errors on different instances have negative $\mathcal{Q}$ values. In [11], the authors showed that the negative dependency of linear classifiers can offer a dramatic improvement in *Boosting*.

The Correlation Coefficient: The correlation between 2 classifiers is given by:

$$\rho_{j,k} = \frac{N^{11}N^{00} - N^{10}N^{01}}{\sqrt{(N^{11}+N^{10})(N^{01}+N^{00})(N^{11}+N^{01})(N^{10}+N^{00})}} \quad (12)$$

The Pairwise Interrater agreement: For this measure, the agreement between each pair of classifiers is calculated as:

$$kp_{j,k} = 2\frac{N^{11}N^{00} - N^{10}N^{01}}{(N^{11}+N^{10})(N^{01}+N^{00})(N^{11}+N^{01})(N^{10}+N^{00})} \quad (13)$$

For any pair of classifiers, $\mathcal{Q}$ and ρ have the same sign. The maximum value of ρ and kp is 1 but the minimum value depends on the individual performance of the classifiers.

In [11], it is reported that there is not unique choice of diversity measure. But, for linear classifiers, the authors recommended the use of $\mathcal{Q}$ *Statistic* for it's simplicity and it's significant results. Then, it's interesting to compare the previous measures in *Boosting Nominal Concept*.

4 Experimental Study

The goal of this section is to study the relationship between the *nominal classifiers* diversity and *AdaBoost* performance for 2-class problems.

Table 2. Characteristics of used data sets

Data Sets	*Instances*	*Attributes*	*Data diversity*
Credit German	1000	20	**98.59**%
Diabetes	768	8	22.83%
Ionosphere	351	4	**90.2**%
Tic Tac Toe	958	9	**100**%
Transfusion	748	4	1.07%

The experiments are performed on 5 real data sets extracted from *UCI Machine Learning Repository*[1] [2] and the algorithms are implemented in `WEKA`[2], a widely used toolkit.

The characteristics of these data sets are reported in Table 2. For each data set, we respectively give the number of instances and the number of attributes. Also, we present the *data diversity* rate that indicates the samples which are different (including the class label) in the data [9].

[1] `http://archive.ics.uci.edu/ml/`
[2] `http://www.cs.waikato.ac.nz/ml/weka/`

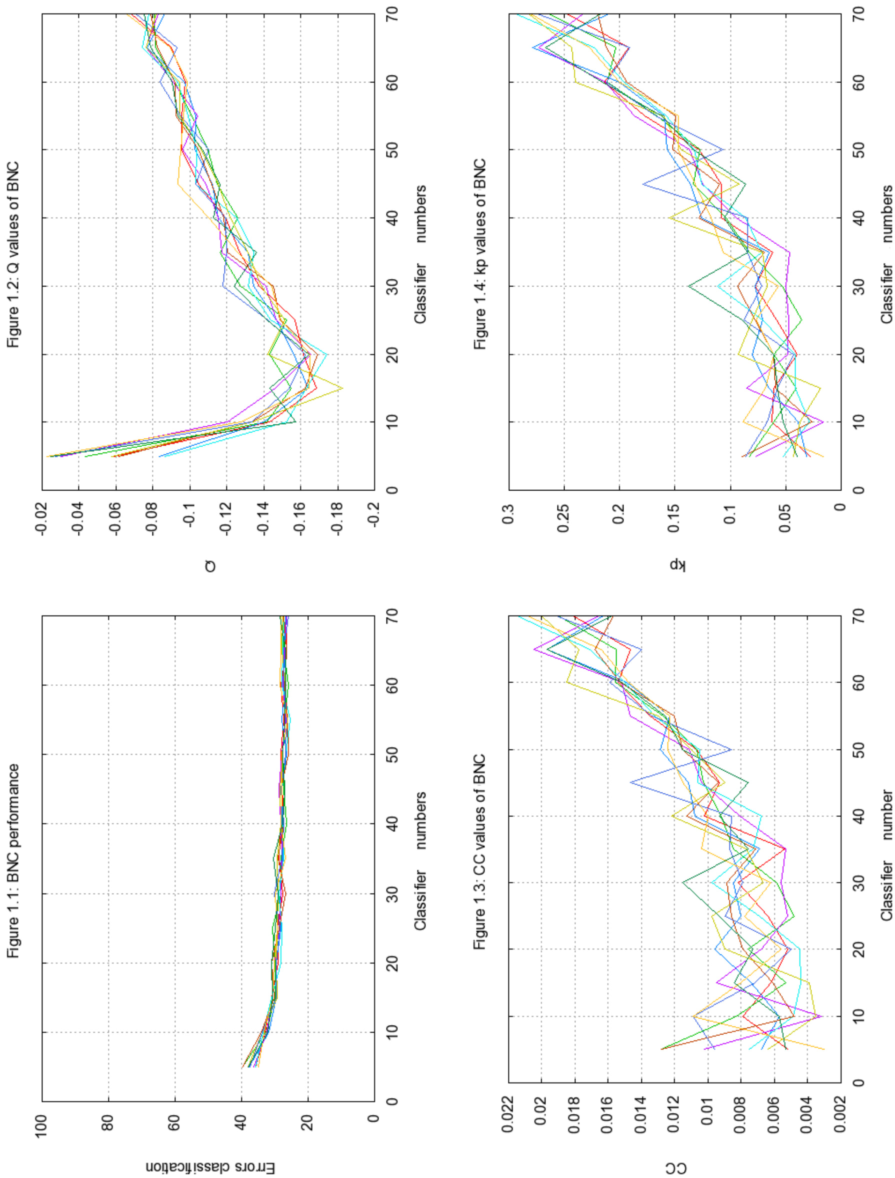

Fig. 1. Diversity and error rates of BNC on Credit German

The performance of *BNC* is evaluated in terms of error rates. To calculate this performance, we report the average of 10 experimentations. Each experiment was performed using *10 cross-validations*, that is the most used method in the literature for validation [10].

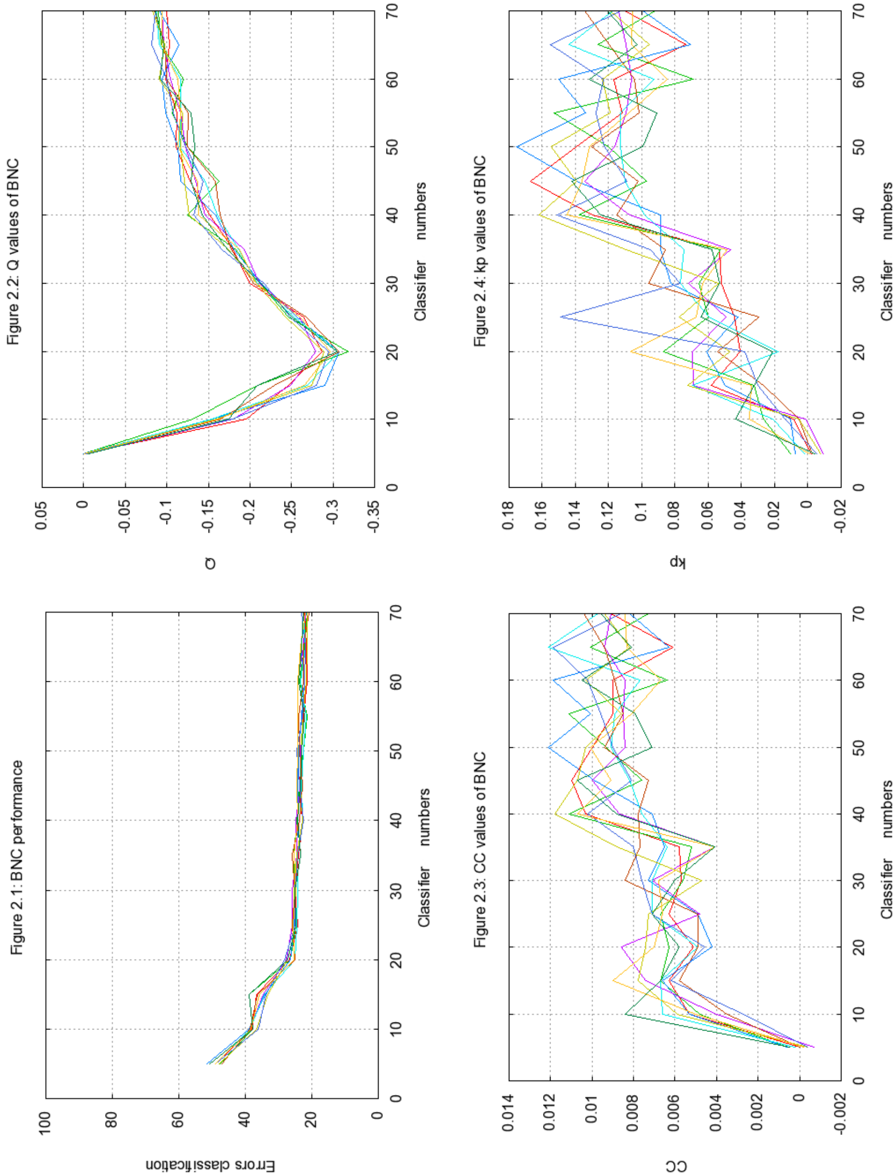

Fig. 2. Diversity and error rates of BNC on Diabetes

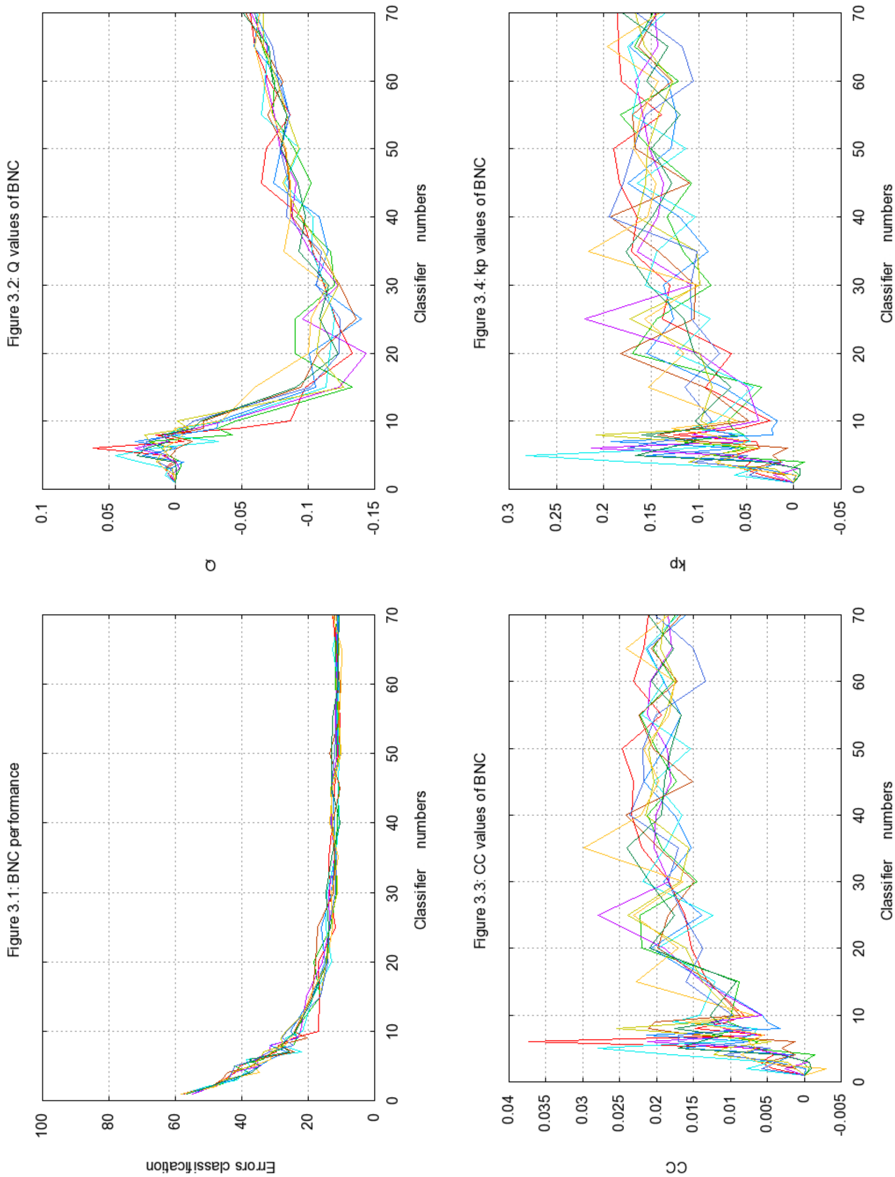

Fig. 3. Diversity and error rates of BNC on Ionosphere

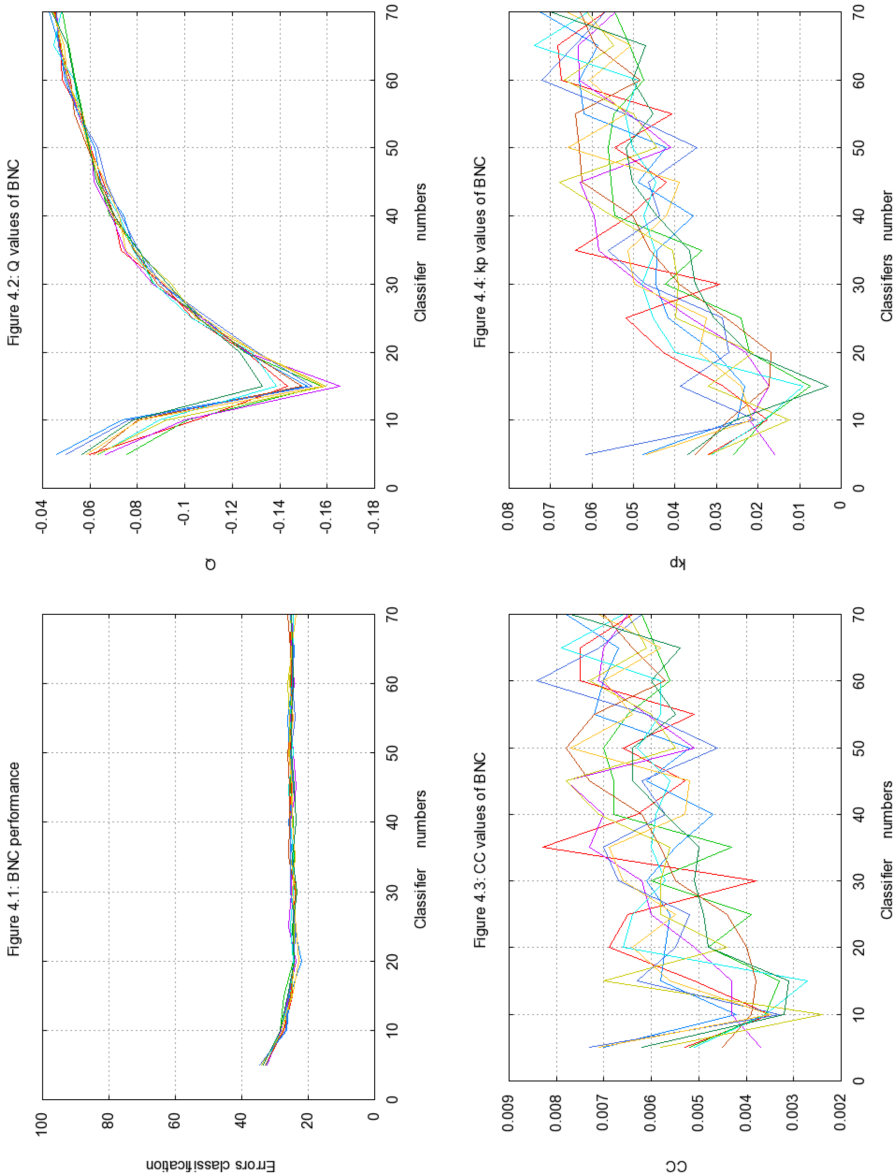

Fig. 4. Diversity and error rates of BNC on Tic Tac Toe

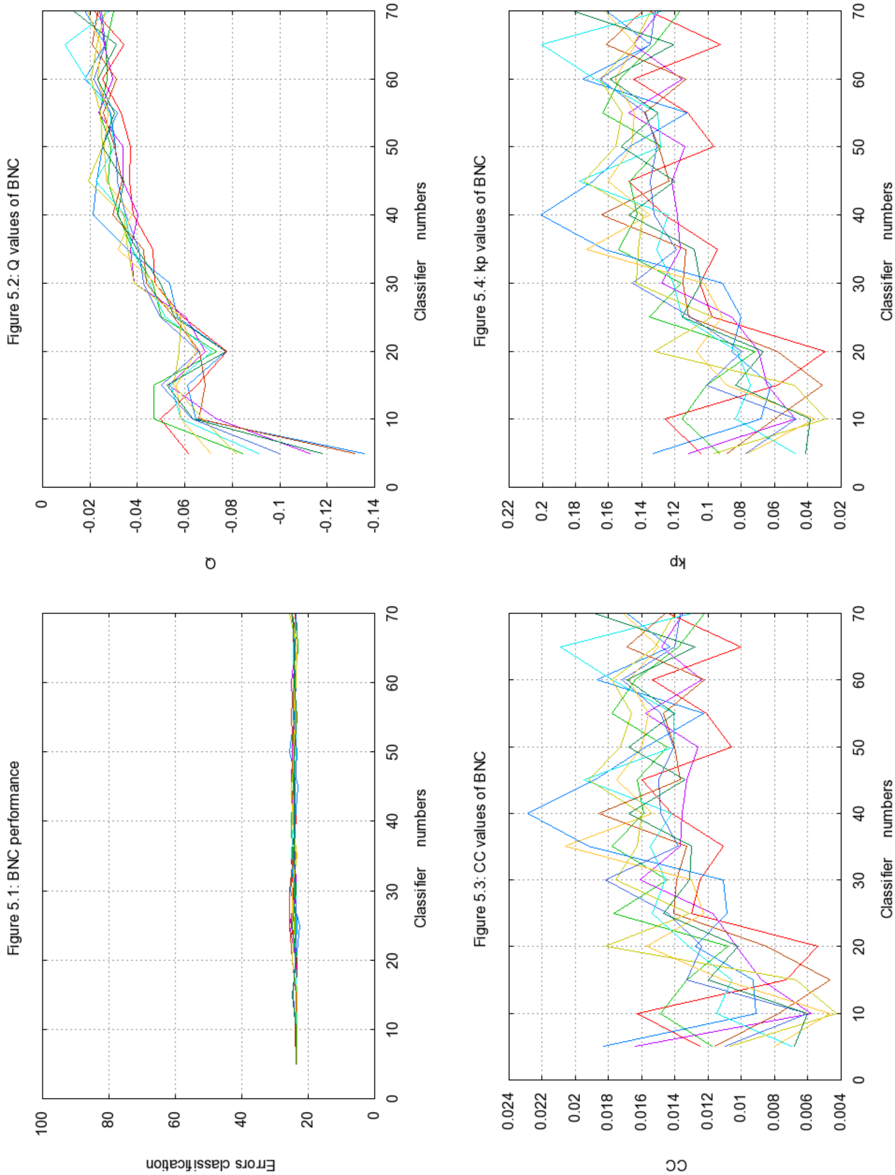

Fig. 5. Diversity and error rates of BNC on Transfusion

It consists on dividing the data sample into 10 subsets. In turn, each subset will be used for testing and the rest are assembled together for learning. Finally, the average of these 10 runs is reported.

Figures 1, 2, 3, 4 and 5 present the performance of *BNC* and the values of the 3 diversity measures on the *Credit German*, *Diabetes*, *Ionosphere*, *Tic Tac Toc* and *Transfusion* data sets respectively.

In figure 1.1, we remark that the performance of *BNC* starts to stabilize when using ensembles of 20 classifiers, for high diversity data (*DD*=98.59%). The classifiers generated are negatively depend ($\mathcal{Q} \leq$ -0.02). The minimum values of $\mathcal{Q}$ *Statistic* are obtained with classifier numbers varying between 10 and 20. From figure 1.3 and figure 1.4, the values of the 2 measures *CC* and *Kp* are very divers and the variation curves are ascending, while the curve of the values of Q is upward then downward.

In figure 2.1, the best performance of *BNC* is obtained with divers classifier ensembles (with $\mathcal{Q}$=-0.3 as minimum average values). In figure 2.2, the minimum values of $\mathcal{Q}$ *Statistic* are obtained with 20 classifiers. For *Diabetes* data (*DD*=22.83%), there is a relation between $\mathcal{Q}$ *Statistic* and the *BNC* performance.

With high divers data (figure 3.2), the first generated classifiers are independent but the rest are negatively depend ($\mathcal{Q} \leq$ -0.15). The minimum values of $\mathcal{Q}$ *Statistic* are obtained with classifier numbers varying between 15 and 30. With less than 20 classifiers, the error rate decreases about 40% (figure 3.1).

From figure 4.1, the difference between the error rates of the first classifier and the generated thereafter, is not important. This show that *Boosting* can converge to the best performance with few classifiers. For this case, Q Statistic is informative. In Figure 4.3 and 4.4, the values of Kp and CC vary an a very arbitrary way.

For the *Transfusion* data set (*DD*=1.07%), the classifier generation does not help to increase *BNC* performance. We conclude that it is not recommanded to use *AdaBoost* for this type of data.

Concerning diversity measures, we can note that for 2-class problems, the values of ρ and k_p are not correlated with *AdaBoost* performance using *nominal classifiers*. The $\mathcal{Q}$ *Statistic* seems like a good measure of model diversity that has a relationship with the performance of *AdaBoost* and then for can be used to stop classifier learning .

5 Conclusions

In this paper, we have study the diversity of *nominal classifiers* in *AdaBoost.M2*. We have compared 3 diversity measures for 2-class problems. We have found that the $\mathcal{Q}$ *Statistic* is significantly correlated with the *AdaBoost* performance, especially for very divers data sets. Then, it's possible to use this measure as a stopping criteria for ensemble learning . But for very correlated data sets, no measure is useful. This results should be confirmed with more correlated data. The diversity of data sets should then be taken into account in *AdaBoost* learning process. It's also interesting to study $\mathcal{Q}$ *Statistic* diversity to see if it can be used in *AdaBoost* for many class problems.

References

1. Aksela, M., Laaksonen, J.: Using diversity of errors for selecting members of a committee classifier. Pattern Recognition 39(4), 608–623 (2006)
2. Asuncion, A., Newman, D.: Machine Learning Repository (2007)
3. Gavin, B., Jeremy, W., Rachel, H., Xin, Y.: Diversity creation methods: A survey and categorisation. Journal of Information Fusion 6, 5–20 (2005)
4. Brown, G., Kuncheva, L.I.: "Good" and "Bad" Diversity in Majority Vote Ensembles. In: El Gayar, N., Kittler, J., Roli, F. (eds.) MCS 2010. LNCS, vol. 5997, pp. 124–133. Springer, Heidelberg (2010)
5. Freund, Y.: Boosting a weak learning algorithm by majority. Information and Computation 121, 256–285 (1995)
6. Freund, Y., Schapire, R.E.: Experiments with a new boosting algorithm. In: 13th International Conference on Machine Learning, Bari, Italy (1996)
7. Freund, Y., Schapire, R.E.: A decision-theoretic generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences 55(1), 119–139 (1997)
8. Gacquer, D., Delcroix, V., Delmotte, F., Piechowiak, S.: On the Effectiveness of Diversity When Training Multiple Classifier Systems. In: Sossai, C., Chemello, G. (eds.) ECSQARU 2009. LNCS, vol. 5590, pp. 493–504. Springer, Heidelberg (2009)
9. Ko, A.H.R., Sabourin, R., Soares de Oliveira, L.E., de Souza Britto, A.: The implication of data diversity for a classifier-free ensemble selection in random subspaces. In: International Conference on Pattern Recognition, pp. 1–5 (2008)
10. Kohavi, R.: A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. In: Actes d'International Joint Conference on Artificial Intelligence, pp. 1137–1143 (1995)
11. Kuncheva, L.I., Skurichina, M., Duin, R.P.W.: An experimental study on diversity for bagging and boosting with linear classifiers. Information Fusion 3(4), 245–258 (2002)
12. Kuncheva, L.I., Rodriguez, J.J.: Classifier Ensembles for fMRI Data Analysis: An Experiment. Magnetic Resonance Imaging 28(4), 583–593 (2010)
13. Meddouri, N., Maddouri, M.: Adaptive Learning of Nominal Concepts for Supervised Classification. In: Setchi, R., Jordanov, I., Howlett, R.J., Jain, L.C. (eds.) KES 2010. LNCS, vol. 6276, pp. 121–130. Springer, Heidelberg (2010)
14. Shipp, C.A., Kuncheva, L.I.: An investigation into how adaboost affects classifier diversity. In: Proc. of Information Processing and Management of Uncertainty in Knowledge-Based Systems, pp. 203–208 (2002)