The use of a simple metric as criteria for the selection and composition of more robust ensembles.

Jose Gomes Lopes Filho
Federal University of Rio Grande do Norte
Natal, Brazil
zefilho@ppgsc.ufrn.br

Marco Cesar Goldbarg
Federal University of Rio Grande do Norte
Natal, Brazil
marcocgold@gmail.com

Antonio Carlos Gay Thome
Federal University of Rio Grande do Norte
Natal, Brazil
thome@dimap.ufrn.br

RESUMO

Neste artigo faz-se uma análise das principais dificuldades encontradas na construção de comitês neurais aplicados na classificação de padrões. Estratégias para geração de agentes acurados, que apresentem boa margem para generalização e, ao mesmo tempo, sejam relativamente diversos para constituÃrem comitês mais eficazes e robustos são empiricamente avaliadas e comparadas. Uma métrica simples e eficaz para comparar e prever, com razoável exatidão, a capacidade de generalização de um comitê é desenvolvida e sua eficácia é comprovada experimentalmente.

PALAVRAS CHAVE. Sistemas Multi Classificadores, Comitês de Redes Neurais, Dilema Acurácia e Diversidade.

ABSTRACT

This article presents an analysis of the main difficulties found in the construction of neural network ensembles applied to pattern classification. Strategies for the generation of accurate agents, which have good margin of generalization and, at the same time, are relatively different in order to build more effective and robust ensembles, are empirically evaluated and compared. A simple yet effective metric to compare and predict with reasonable accuracy a ensemble's generalization ability is proposed and its effectiveness is experimentally demonstrated.

KEYWORDS. Multi Classifier Systems. Neural network ensembles. Accuracy and Diversity Dilemma.

1. Introduction

The use of classifiers that are built not in a monolithic form, but from the combination of multiple agents (or classifiers) combined together in different ways in light of the solution strategy adopted for the problem, has been studied for over 30 years. Dasarathy and Sheela (1979) released one of the first works in this sense.

Multiple Classifier Systems - MCS have been used extensively to solve various problems since, in general, they achieve better performances than the ones shown by the classifiers (or agents) that were used as basis to form them, Kittler, Hojatoleslami and Windeatt (1997), Xu, Krzyzak and Suen (1992). Nowadays, in the literature, there is a large amount of material proposing different ways to organize multiple agents in the composition of a MCS, Ranawana and Palade (2006), Lima (2004), Coelho *et al.* (2006), Dieguez (2012).

One way to compose a Multi Classifier System is to gather a set of distinct agents into an ensemble, which is a machine-learning paradigm. In this kind of organization, a finite collection of alternative hypothesis for the complete solution of the problem is used in order to form a unique and global proposition to reach better performance than the one offered in separate by each of the classifiers used as the basis of the structure, Dietterich (2000). For this to happen, however, it is necessary that the classifiers are both accurate and different from each other, which is known in the literature as the diversity versus accuracy dilemma.

The idea of forming classifier ensembles that have a good "knowledge" (or accuracy) on a particular problem and at the same time have "opinions" that are in some degree different from the other components of the ensemble (diversity) was initially proposed by Hansen and Salamon (1990).

According to the understanding of the scientific community, the similarity and discrepancy (diversity) rates found in the answers produced by different agents trained individually and independently make it possible to identify and select a subset to form an ensemble that is most likely to provide a better performance than the one achieved by all of the agents individually.

Intuitively, the disagreement between different agents is directly related to the potential improvement in the set's performance, but this is true only if they make mistakes on different samples (that is, they are complementary in a certain degree). The difficulty lies, even today, on the correct way to measure the diversity not between pairs of agents, but for the group as a whole - and the correct way to assess the relative importance of the diversity in light of the accuracy of these agents. Schapire *et al.* (1998), introduced a third component, called margin, in order to measure the ensemble's reliability regarding their response.

Considering these three parameters, the construction of an ensemble, regarding the determination of its size and the selection of the most suitable agents to compose it, was showed be a NP-complete and multiobjective optimization problem according to Tamon and Xiang (2000), where the search space grows exponentially with the number of candidate agents, as shown in equation 1 below.

$$SS = \sum_{i=2}^{N} \left(C_i^N\right) = 2^N - 2 \tag{1}$$

Where SS stands for the search space size, N represents the number of candidate agents and C_i^N is the combinatory operator of N agents into groups of i agents.

Despite the intense research in the last decades, there is no consensus yet about the relative importance of these three parameters. The same occurs among the different metrics proposed to measure them and about the methods used to generate a good set of candidate agents that, once gathered in an ensemble, may ensure performance improvement in the generalization of the solution to a given problem.

Considering the shortcomings and lacks of consensus that characterize the current context, we decided to carry out an empirical study in order to investigate the importance and relative relevance among accuracy, margin and diversity in the formation of good ensembles of classifiers.

In this article, organized in five section, we make a brief description of some of the many experiments performed and a summary of the main results achieved. In section 2, we describe the main steps used for the construction of an ensemble. In section 3, we present some of the many metrics to compute diversity, margin and accuracy found in the literature and we propose a new metric called Robustness. In section 4, we analyze the results obtained considering the relative relevance among diversity, margin, accuracy and robustness in the formation of ensembles. The conclusions are in the section 5.

2. Ensemble Construction

The construction of an ensemble of classifiers involves three steps performed in sequence. The first one consists on the generation of a set of accurate and diverse candidates, the second is the identification and selection of a subset of the candidates that be capable of forming the best ensemble or the one with higher probability to provide good generalization and the third consists on the selection of the most suitable synthesis strategy. In this session, we make a brief description of some of the different strategies found in the literature.

2.1. Agent Generation

The goal is to generate a number of agents with good level of accuracy and relatively diverse among them, once if all agents make the same mistakes and successes (which means diversity almost zero) it will not make sense to group them into an ensemble. Different strategies are suggested in the literature, where the most well-known techniques are Bagging and Boosting.

Bagging, which is an acronym for *booststrap aggregating*, is a machine learning meta-algorithm introduced by Breiman (1996). In it, the set of agents is generated by training each one over a different set of samples. Prior each agent be trained, a new training set is generated from the original set of samples using uniform sampling with replacement. Because of this, it is expected that the differences among the training sets are able to produce a reasonable level of diversity among the generated agents.

Boosting is another learning meta-algorithm introduced by Schapire (1990). Many other meta-algorithms that are variants to the original proposal can now be found in the literature, the most important of them is AdaBoost proposed by Freund and Shapire (1997 and 1999). In the Adaboost strategy, different training sets are generated from the original set with uniform sampling but without replacement. Another difference to Bagging is that the probability of choosing a particular sample grows in the direct proportion to its contribution to the already trained agents' error, that is, if a sample has not been classified correctly by them, the probability of its selection in-creases when compared to the others.

ReinSel is another method, which was proposed in Canuto *et al.* (2012), where diversity is not fetched from the use of different training sets. The approach proposes the identification and the selection of different subsets of the feature space to provide specialized training for the agents, i.e., with greater capacity to label different classes.

In the event of neural agents ensembles, there are alternatives suggested in the literature, Maclin and Shavlik (1995), Cherkauer (1996), Opitz and Shavlik (1996) with focus on the change of configuration and training parameters, such as number of training cycles, learning rate, stop criteria and many others.

2.2. Selection of Agents and Composition of the Ensemble

With the set of candidates to compose the ensemble already trained, the following step is to choose, among the candidates, those that will really contribute with the improvement of the ensemble's performance.

Coelho, in his master's thesis, Coelho (2006), provides an overview of selection strategies that were available so far in the literature.

- *Constructive without Exploration*: initially, all candidates are sorted based on their individual performance considering a unique and new set of data that was not used for training. After that, the candidate with the best performance is automatically inserted in the ensemble and the others, starting from the second, is only inserted if its insertion improves the performance of the ensemble.
- *Constructive with Exploration*: this strategy differs from the strategy without exploration by the fact that for each new possible inclusion step a search is done over all candidates still available, along with the inclusion of the candidate that brings the greater performance gain to the ensemble. The search process ends when none of the still available candidates is able to bring any improvement for the performance of the ensemble. The computational cost of this strategy is greater than the one in the previous strategy and, in general, it generates smaller ensembles, i.e., ensembles with a smaller number of agents.
- *With no Selection*: This technique consists simply in using all of the trained candidates to compose the ensemble.

The first two strategies are characterized as greedy, where each inclusion is only accepted and carried out if it brings an increase to the ensemble's accuracy. In both strategies, the resulting ensemble is generally composed of a much smaller number than the number of available candidates. The controversy over the formation of an ensemble based on a subset or with the use of all the available candidates can be better assessed by reading the articles published by Schapire *et al.* (1998) and Zhou, Wu and Tang (2002).

2.3. Construction of the Decision Making Module

The choice of the decision module is another important step in the construction of ensembles and consists in choosing the strategy to be adopted to integrate the response of each component in a single response. Several strategies are suggested in the literature, Coelho (2006), as for example:

1. *Simple Average*: The ensemble response is given by the simple average of the responses of all of the agents that are part of it.
2. *Majority Vote*: The ensemble response is the same as the response that is provided by the majority of the staff.
3. *Winner-takes-all*: The ensemble response is the same as the response of the agent that is the most convincing, i.e., the one that presents the largest absolute value in relation to all others.

3. Accuracy, Diversity and Margin

Despite the intense investigation held in the recent years, there is still no consensus among researchers about the metrics that can identify, in advance, a good group of candidate that can form the best or, at least, the more robust ensemble with regard to the generalization ability.

Kuncheva and Whitaker (2001; 2003), Tang, Suganthan and Yao (2006), Brown (2004), Brown and Kuncheva (2010) and Schapire *et al.* (1998), among others, are the researchers that are most dedicated to the study and discussion of the concepts and metrics to measure accuracy, diversity and margin.

3.1. Accuracy

Accuracy or performance measures the level (in percentage) of success achieved by an agent or by an ensemble when labeling a set of samples. The simple division between the number of correct labeling and the global number of samples computes accuracy.

Depending on the type of problem addressed, additional measures such as false positive and false negative are also used, Fawcett (2006) .

3.2. Diversity

The concept of this measure, when applied between two independent agents, is simple and intuitive once it represents the degree of non-similarity of their mistakes. As more similar are the errors, closer to zero is the diversity measure and vice-versa. On the other hand, when applied to a set of independent agents, the metric becomes less intuitive and its computation more complex and less precise. Below, we present some of the main measures as mentioned in the literature.

1. *Disagreement*, Kuncheva and Whitaker (2001):

$$D_{i,j} = \frac{N^{01} + N^{10}}{N^{11} + N^{00} + N^{01} + N^{10}} \tag{2}$$

$$D_L = \frac{2}{L(L-1)} \sum_{i=1}^{L-1} \sum_{j=i+1}^{L} D_{i,j} \tag{3}$$

Where $D_{i,j}$ represents the diversity between the classifiers *i* e *j* and D_L represents the average diversity of a set of L classifiers. N^{11} is the total number of samples correctly classified by the two classifiers; N^{00} is the number of times in which both were wrong; N^{01} and N^{10} are, respectively, the number of samples in which one classifier was wrong while the other was right. The value of D_L varies between $[0, 1]$, zero when there is no diversity and one when the disagreement between the agents is at a maximum level.

2. *Difficulty Index* - $diff$, Hansen and Salamon (1990):

Suggested by Hansen and Salomon, it defines a discrete random variable V_i, calculated by Equation 4 for the x_i sample that was randomly extracted from the training set; L is the number of agents in the ensemble; and l_i is the number of agents that classified x_i incorrectly.

$$V_i = \frac{L - l_i}{L} \tag{4}$$

$$diff = var[V_i], i = 1, 2, ...N \tag{5}$$

The diversity increases as the value of the diff increases.

3. *Good (Gd) and Bad (Bd) diversity*, Brow and Kuncheva (2010):

$$Gd = \frac{1}{|P^+|\,C} \sum_{i=1}^{|P^+|} v_i^- \tag{6}$$

$$Bd = \frac{1}{|P^-|\,C} \sum_{i=1}^{|P^-|} v_i^+ \tag{7}$$

Where C is the amount of classifiers in the ensemble, v_i^+ is the amount of classifiers that were right about sample i with the set of examples P^-, which the ensemble classified incorrectly; and v_i^- is the amount of classifiers that were wrong in the classification of sample i in the set of instances P^+, which the ensemble correctly classified.

3.3. Margin

Proposed by Schapire *et al.* (1998) with the purpose of explaining the success of the strategy introduced by the Boosting algorithm as opposed to the Bagging algorithm. According to the researchers, the ensemble's error on the agents' training set is not sufficient to predict the performance on the test set (generalization). As an alternative, they created the concept of margin in order to measure the ensemble's reliability regarding their response, i.e., in the case of a decision made by majority vote, this margin was going to be given by the difference between the number of votes correct and wrong for each of labeled samples. In equation 8, T_i is the expected response for the x_i sample and y_j is the output of the j-th agent that makes up the ensemble of L agents.

$$m_i = \frac{1}{L} \sum_{j \in C / y_j = T_i} 1 \; - \sum_{j \in C / y_j \neq T_i} 1 \tag{8}$$

Since, in this work, instead of majority vote we use a simple average to compute the ensemble's response, it was necessary to adapt the formula proposed by Schapire. Equations 9 to 11 provide, respectively, the new way to compute the margin value for i-th sample and the average margin for the entire set of samples.

$$m_i = \lambda(max(\widehat{Y}_i) - max2(\widehat{Y}_i)) - (1 - \lambda)(max(\widehat{Y}_i) - \widehat{y}_i) \tag{9}$$

$$\lambda = \begin{cases} 1, se\, \widehat{Y}_i = T_i \\ 0, se\, \widehat{Y}_i \neq T_i \end{cases} \tag{10}$$

$$M_L = \frac{1}{N} \sum_{i=1}^{N} m_i \tag{11}$$

Where $\widehat{Y}_i$ represents the ensemble's response vector for the i-th sample; T_i is the expected response vector (target) for the same sample. For $\lambda = 0$, $\widehat{y}_i$ represents the j-th output value of the ensemble vector that matches with the expected response given by T_i. In case of $\lambda = 1$, $max()$ and $max2()$ functions return, respectively, the value of the highest and second highest output of $\widehat{Y}_i$. This measure also varies between $[-1, +1]$ and and has the same interpretation as the one proposed by Schapire et al. (1998).

Here, following the same line of reasoning of Schapire *et al.* (1998), we suggest another metric to estimate the effectiveness of the ensemble regarding the generalization of the acquired knowledge, called Robustness (ROB). The computation of the metric is given by equations 12 and 13.

$$N_l = \textstyle\sum_{i=1}^{N} l_i \quad \therefore \quad l_i = \begin{cases} 1, se\; v_i^+ \leq \frac{L}{2} \\ 0, se\; v_i^+ > \frac{L}{2} \end{cases} \tag{12}$$

$$Rob = \frac{1}{N} \left(\frac{N - N_l}{max(1, N_l)} \right) \tag{13}$$

Where N is the number of samples in the training set, L is the number of agents in the ensemble and v_i^+ is the number of agents that got success on labeling the i-th sample.

4. Accuracy, Margin and Diversity Analysis

In this section, we make an empirical and detailed analysis on the importance and the correlation of accuracy, margin and diversity in the construction of an ensemble that is most likely to present the best performance in the generalization phase.

4.1. Agent Generation

Although agents built from different paradigms and organized as a homogeneous or heterogeneous group might form an ensemble, it was decided to consider, in this work, only the homogeneous type formed only by neural networks of MLP type. More information about this class of ensembles can be found in the work of Shapire *et al.* (1998), Breiman (1996), Freund and Schapire (1999), Canuto *et al.* (2012). According to Cherkauer (1996), multiple neural networks systems seek to increase the accuracy of the classification or, at least, reduce the inherent variance on the training process, selecting those agents that might be complementary in some way.

The problem used as case study for the development of this research consists on the reading task of the three first characters extracted from the automotive vehicles' license plate.

The database contains 14,992 letters samples (from A to Z) encoded as vectors with 51 numerical attributes. The original set of samples was randomly divided into three subsets: 1) a subset "A" with 7,566 letters used to train the agents; 2) a subset "B" with 2,992 letters used to build the ensemble; and 3) a subset "C" with 4,434 letters used to test the ensemble generalization capability. All subsets, although randomly picked, maintain the balance of the classes according to the original set. Both sets "A" and "B" were yet subdivided in a ratio of 70% and 30%, for training and validation purposes on the two phases: agents' training and ensemble construction.

With the objective to generate agents with reasonable performance and diversity levels, it was decided to use the following described approach. In first place, adopt the Bagging strategy to generate the different sets of samples to train the agents (one set for each agent). Secondly, randomly select a different configuration scheme (1 or 2 hidden layers) for each agent. Third, select a different starting point for each training section and, finally, randomly select a maximum number of cycles for each training section (between 50 and 200).

4.2. Ensemble Composition Strategies

Once the search for the best ensemble becomes exponentially costly as the number of candidates increases, the use of heuristic methods becomes almost imperative. The selection of the candidates can be done, basically, in two ways: static or dynamic. In the static form, the ensemble is formed in advance and remains fixed (the same) for all posterior usage. In the dynamic form, the ensemble is built on line (on demand) and a different ensemble can be selected for each new sample presented. In this study, we explored only the static selection process.

A total of 673 different ensembles were systematically generated with the purpose of enabling further analysis of the influence of the accuracy, margin and diversity measures in their generalization capacity. The ensembles were created varying in size from 2 to 50 agents, chosen from a set of 50 candidates with individual accuracy varying from 0.8987 and 0.9332. For each ensemble size (from 2 to 49) and each metric (accuracy, margin and diversity), 02 ensembles were selected, one that maximizes and other that minimizes the respective metric.

4.3. Accuracy, Margin and Diversity Analysis

Finding a mechanism that might guide the process of choosing the most adequate set of candidates and also, may allow the estimation, in advance, of the accuracy and the generalization capability of the being formed ensemble, is an old aspiration of the scientific community that is still an open challenge.

For this analysis, we select 07 different performance metrics as suggested in the literature. One metric (DMM), representing the average accuracy of the agents that compose the ensemble. Two metrics that estimate the accuracy of the ensemble: Margin - MAR (eq. 9, 10 and 11) and Robustness - ROB (eq. 12 and 13), and 04 metrics that estimate the diversity: Disagreement - DIV (eq. 2 and 3), Difficulty Index - DIFF (eq. 4 and 5), Good Diversity - Gd (eq. 6) and Bad Diversity - Bd (eq. 7).

Observing the correlation index between each metric and the performance provided by all formed ensembles over the test sample set (Table 1), it is noticed that the accuracy metrics (DMM,

MAR and ROB) are positively correlated with the generalization capability of the ensemble; the diversity metrics (DIV and DIFF) are negatively correlated, while (GD and BD) are positively correlated. The DIFF metric is the one that shows the highest absolute correlation index.

Table 1: Correlation between measures and the ensemble's performance on the test sample set.

DMM	MAR	ROB	DIV	DIFF	GD	BD
0.1726	0.4262	0.6863	-0,029	-0.923	0.6959	0.3173

By analyzing the behavior of each of the measures in relation to the performance achieved by the different ensembles, as shown in Figure 1, it is noticed that the best and the worst performances were not restricted to the extreme values (maximum or minimum) of any of the 07 measures.

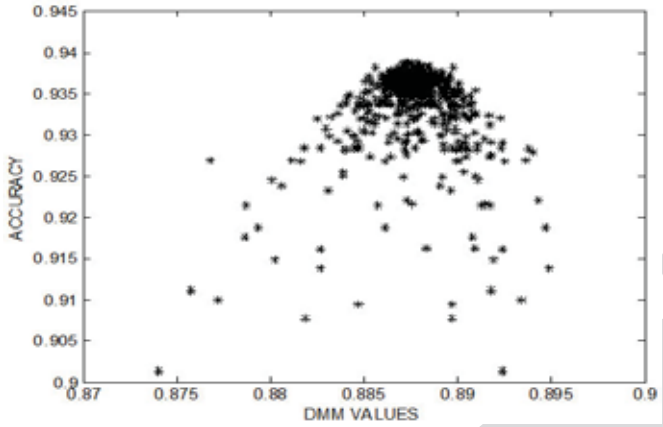

Figure 1: Average perfomace of the 673 ensembles set in order by their respective DMM accuracy measure.

Based on the same analysis, when comparing all of the 07 graphics that were generated, one for each metric, it is seen that the diversity measure given by DIFF - difficulty index - is that one that shows greater coherence in relation to the ensemble's performance, that is, the lower is DIFF, better is the ensemble's performance and vice versa (Fig. 2).

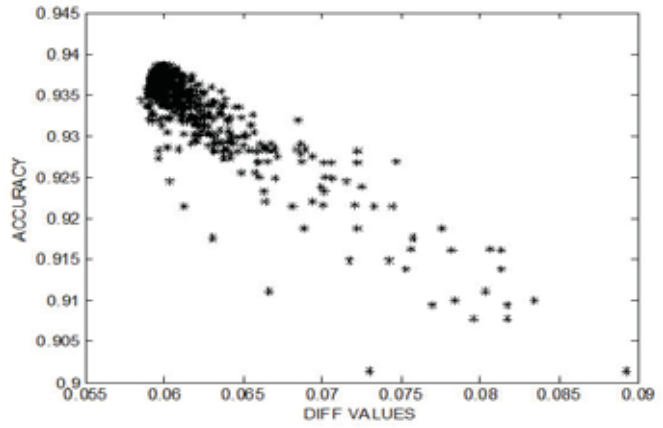

Figure 2: Average perfomace of the 673 ensembles set in order by their respective DIFF - difficulty index.

In the search of a more consistent metric combining measure of accuracy and diversity, we found that the better results were provided by following ratios: ROB/DIFF, MAR/DIFF and DMM/DIFF as shown in Figure 3. The figure shows the average accuracy of the ensembles on the

vertical axis and, on the horizontal axis, eleven windows of fixed size used to join ordered values of the cited metrics.

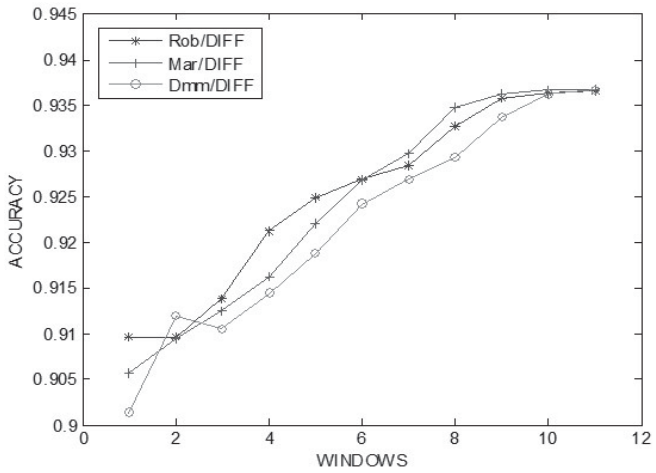

Figure 3: Average performance curve of the 673 ensembles (vertical axis) by the ratio of the measures ROB/DIFF, MAR/DIFF and DMM/DIFF.

The results show that the three metrics present almost the same behavior, indicating clearly an estimate that the higher the value of the metric, higher might be the expectation of a good performance of the ensemble on the generalization phase. Weaker results are observed when the GD diversity measure is used in substitution to DIFF (Fig. 4).

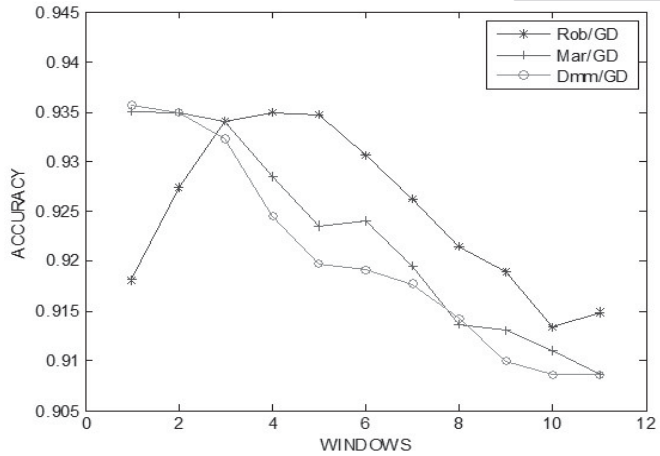

Figure 4: Average performance of the 673 ensembles (vertical axis) set in order by their GD diversity measure.

Figure 5 shows the result of the comparison of the three better rated metrics (DIFF, DMM/DIFF, and ROB/DIFF) when used as the selection criteria to identify the possible best ensemble for the generalization phase. In the Figure, from left to right, we see the performance of the five best ensembles selected based on each of the criterions.

The image clearly shows that the use of the ensemble's simple accuracy, given by DMM, is weaker than all other metrics, which are nearly equal, with a very small advantage for the metric presented by ROB/DIFF.

5. Conclusion

The focus of the study was on the construction of homogeneous ensembles based on MLP neural networks and the identification of metrics that could serve as mechanism to guide the

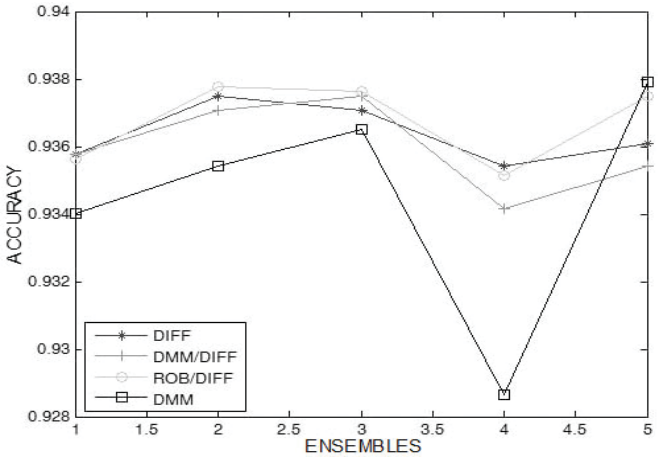

Figure 5: Performance of the top five ensembles on the test set.

process of choosing the more adequate candidate agents in order to get the best performance on the generalization phase.

The research was empirical and applied to the pattern recognition problem, using as a case study the recognition of the letters of the Brazilian alphabet drawn from motor vehicle license plates whose images were obtained in a real environment through video cameras.

The work included the generation of a set of 50 neural agents of MLP type to be used as candidates for composing the different formed ensembles. Bagging was the strategy used for training and generating a relatively accurate and, at the same time, diverse set of agents.

A total of 673 ensembles were generated from the base of candidates and, through a detailed analysis, it was possible to evaluate the significance and correlation of the 07 considered metrics to measure accuracy, margin and diversity.

It was found, leastwise in the problem used as case study, that the difficulty index - DIFF alone is a good metric to be used. Other metrics that also proved to be effective were (ROB/DIFF), the ratio between robustness (ROB) and difficulty index (DIFF), (DMM/DIFF), the ratio between average performance of the agents (DMM) and difficulty index (DIFF).

References

Breiman, L., Bagging predictors. *Machine learning*, 24(2):123-140, 1996.

Brown, G. and Kuncheva, L. I., Good and bad diversity in majority vote ensembles. *Multiple classifier systems*, pages 124-133. Springer, 2010

Brown, G., Diversity in neural network ensembles. 2004.

Canuto, A. M. P., Vale, K. M. O., Feitos, A. and Signoretti, A., Reinsel: A class-based mechanism for feature selection in ensemble of classifiers. Applied Soft Computing, 12(8):2517-2529, 2012.

Cherkauer, K. J., Human expert-level performance on a scientific image analysis task by a system using combined artificial neural networks. *Working notes of the AAAI workshop on integrating multiple learned models*, pages 15-21. Citeseer, 1996.

Coelho, G. P., Geração, seleção e combinação de componentes para ensembles de redes neurais aplicadas a problemas de classificação. Universidade Estadual de Campinas, 2006.

Dasarathy, B. V. and Sheela, B. V., Composite classifier system design: Concepts and methodology. *Proceedings of the IEEE*, 67(5):708-713, 1979.

Dieguez, A. A., Geração de um comite de agentes classificadores usando otimização. Universidade Federal do Rio de Janeiro, Rio de Janeiro, RJ, Brasil, 2012.

Dietterich, T. G, Ensemble methods in machine learning. *Multiple classifier systems*, pages 1-15. Springer, 2000.

Fawcett, T., An introduction to roc analysis. *Pattern recognition letters*, 27(8):861-874, 2006.

Freund, Y. and 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.

Freund, Y. and Schapire, R. E., A short introduction to boosting. *Journal-Japanese Society For Artificial Intelligence*, 14(771-780):1612, 1999.

Hansen, L. K. and Salamon, P., Neural network ensembles. *IEEE transactions on pattern analysis and machine intelligence*, 12(10):993-1001, 1990.

Kittler, J., Hojatoleslami, A. and Windeatt, T., Strategies for combining classifiers employing shared and distinct pattern representations. *Pattern Recognition Letters*, 18:1373-1377, 1997.

Kuncheva, L. I. and Whitaker, C. J., Ten measures of diversity in classifier ensembles: limits for two classifiers. *Colloquium Digest-IEE*, pages 16-25. IEE; 1999, 2001.

Kuncheva, L. I. and Whitaker, C. J., Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine learning*, 51(2):181-207, 2003.

Lima, C. A. M., Comitê de Máquinas: uma abordagem unificada empregando maquinas de vetores-suporte. PhD thesis, Universidade Estadual de Campinas, 2004.

Maclin, R. and Shavlik, J. W., Combining the predictions of multiple classifiers: Using competitive learning to initialize neural networks. *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, pages 524-531, 1995.

Opitz, D. W. and Shavlik, J. W., Generating accurate and diverse members of a neural network ensemble.*Advances in neural information processing systems*, pages 535-541, 1996.

Ranawana, R. and Palade, V., Multi-classifier systems: Review and a roadmap for developers. *International Journal of Hybrid Intelligent Systems*, 3(1):35-61, January 2006.

Schapire, R. E., Freund, Y., Bartlett, P. and Lee, W. S., Boosting the margin: A new explanation for the effectiveness of voting methods. *Annals of statistics*, pages 1651-1686, 1998.

Schapire, R. E., The strength of weak learnability. *Machine learning*, 5(2):197-227, 1990.

Tamon, C. and Xiang, J. On the boosting pruning problem. *Proceedings of the 11th European Conference on Machine Learning*, pages 404 - 412, 2000.

Tang, E. K., Suganthan, P. N. and Yao, X., An analysis of diversity measures. *Machine Learning*, 65(1):247-271, 2006.

Xu, L., Krzyzak, A. and Suen, C. Y., Methods of combining multiple classifiers and their applications to handwriting recognition. *IEEE Transactions on Systems, Man and Cybernetics*, 22(3), 1992.

Zhou, Z., Wu, J. and Tang, W., Ensembling neural networks: many could be better than all. *Artificial intelligence*, 137(1):239-263, 2002.