# EMnGA: Entropy Measure and Genetic Algorithms Based Method for Heterogeneous Ensembles Selection

Souad Taleb Zouggar[1] and Abdelkader Adla[2(⊠)]

[1] Department of Economics, University of Oran 2, Bir El Djir, Algeria
souad.taleb@gmail.com
[2] Department of Computer Science, University of Oran 1, Oran, Algeria
adla.abdelkader@univ-oran.dz

**Abstract.** Generating ensembles of classifiers increase the performances in classification and prediction but on the other hand it increases the storage space and the prediction time. Selection or simplification methods have been proposed to reduce space and time while maintaining or improving the performance of initial ensemble. In this paper we propose a method called EMnGA that uses a diversity-based entropy measure and a genetic algorithm-based search strategy to simplify a heterogeneous ensemble of classifiers. The proposed method is evaluated against its prediction performance and is compared to the initial ensemble as well as to the selection methods of heterogeneous ensembles in the literature using a sequential way.

**Keywords:** Classification · Diversity · Homogeneous ensemble
Heterogeneous ensemble · Ensemble selection · Genetic algorithms
Fitness function

## 1 Introduction

The purpose of classification is to learn a target function that links a set of data to a set of predefined categories (classes). The constructed target functions are also called classification models.

To address the problem of instability of some classifiers [1, 2], it is asserted that the use of an ensemble classifiers generally gives better results than the use of a singular classification model [3–6]. These ensembles can be either homogeneous or heterogeneous depending on the nature of the models used. The homogeneous ensembles are obtained by performing different executions of the same learning algorithm; for example these ensemble can be obtained by varying the parameters of the learning algorithm or by manipulating the data (input or output) [2, 7–9]. As for heterogeneous models, they are obtained by executing different learning algorithms on the same set of data. These models have different visions of the data which allows obtaining diverse sets.

Generating a large number of predictors allows exploring the solution space largely, and by aggregating all the predictions, we recover a predictor that accounts for all this

exploration. The increase in the number of models can be done without risk of over-learning [10]. However, a very large number of models require a large storage space and an important time for the prediction due to the interrogation of all the predictors.

The simplification of ensembles of classifiers, called ensemble selection, reduces the size of ensembles prior to their integration or combination [11–16]. In this paper, we use a diversity-based entropy measure combined to a genetic algorithm-based search strategy to simplify an ensemble of heterogeneous classifiers.

The rest of the paper is organized as follows: In Sect. 2, we present the recent works in the ensemble selection field and especially those related to heterogeneous ensembles. Section 3 details the proposed method describing the entropy measure and search strategy used. In Sect. 4, we give the experimentation results and a comparative study with heterogeneous ensemble selection method in literature. Finally in Sect. 5, we conclude and give some future work.

## 2   Literature Review

Ensemble methods have attracted a growing interest since their appearance. These methods have been applied in several areas namely, statistics, pattern recognition, and machine learning. These methods consist of two main phases: a model production phase and an aggregation or combination phase of these models. In the first phase, variable performance models are added arbitrarily. The reduced performance models negatively affect the ensemble performance. An ensemble may contain several similar models which reduces its diversity.

Several recent works on ensemble selection [3, 17–19] have been developed in the literature. Dai [20] proposes an improvement of the ensemble selection method of the same authors. This method uses backtracking in depth, which is perfectly adapted to systematically seek solutions to combinatorial problems of great magnitude. This improvement concerns the response time of this method, which has been considerably improved in this study.

Bhatnagar et al. [21] perform ensemble selection using a performance-based and diversity-based function that considers the individual performance of classifiers as well as the diversity between pairs of classifiers. A bottom-up search is performed to generate the sub ensembles by adding various pairs of classifiers with high performance.

Simplifying a set of classifiers usually involves reducing the number of trees while maximizing performance. Based on the approximate ensembles, Guo et al. [22] propose a new framework for ensemble selection. In this context, the relationship between attributes in an approximate space is considered a priori as well as their degree of maximum dependence. This effectively reduces the search space and increases the diversity of selected sub-ensembles. Finally, to choose the appropriate sub-ensemble, an evaluation function that balances diversity and precision is used. The proposed method allows repetitively changing the search space of the relevant sub-ensembles and selecting the next ensemble from a new search space.

The selection methods were applied for either homogeneous or heterogeneous cases [16, 23]. The heterogeneous ensembles are developed for sensitive application

domains. Haque et al. [24] propose a search based on genetic algorithms to find the optimal combination to form a heterogeneous ensemble from an ensemble of classifiers. For this, they develop an algorithm that uses decimal cross-validation on the learning ensemble to evaluate the quality of each candidate ensemble. The proposed method uses a random resampling approach to balance the class distribution and is particularly used for class imbalance cases.

Pölsterl et al. [25] use a heterogeneous ensemble including survival models with the purpose of predicting the survival of patients with stage 3 prostate cancer. The results demonstrate that the constructed ensemble can predict the date of death of patients with this disease.

Partalas et al. [26] propose a diversity measure and a forward selection search strategy that considers all the possible cases that may exist when adding a certain model $h_t$ to an ensemble. The measure called FES (Focused Ensemble Selection) considers 4 cases when adding.

$$FES_{Eval}(h_k, SUB) = \sum_{i=1}^{|Eval|} (\alpha * I(y_i = h_k(x_i)ETy_i \neq Sub(x_i)) - \beta * I(y_i \neq h_k(x_i)ETy_i = Sub(x_i))$$
$$+ \beta * I(y_i = h_k(x_i)ETy_i = Sub(x_i)) - \alpha * I(y_i \neq h_k(x_i)ETy_i \neq Sub(x_i))).$$

Eval: the sample of evaluation or selection;

$I(y_i = h_k(x_i)ETy_i \neq Sub(x_i)) = 1$ if the instance $x_i$ is well classified by the model $h_k$ and is not properly classified by the current ensemble Sub and 0 otherwise.

The factors $\alpha$, $\beta$ represent respectively the number of models in the *Sub* ensemble correctly classifying the instance $(x_i, y_i)$ and the number of models not properly classifying the same instance.

## 3 The EMnGA Method

### 3.1 Entropy Measure

The key idea in this approach is to generate only trees that have maximum diversity (they are less correlated with each other). This is based on the principle that the generalization error of the random forest will be on the wane while diversity among the trees increases.

Let $\Omega_V$ be a sample of individuals with their labels (classes), $|\Omega_V| = n$, $\Omega_V = \{v_1, \ldots, v_n\}$, and $\Omega_V = \{v_1, \ldots, v_n\}$. Each individual $v_j$ is described by m variables denoted $x_{1j}, \ldots, x_{mj}$. Let $Ci$ be a classifier of the classifiers ensemble $\{C_1, \ldots, C_i, \ldots, C_T\}$ represented by a n-dimensional binary vector $y_i = (y_{1i}, \ldots, y_{ni})^T$ such that $y_{ji}=1$ if the classifier $C_i$ classifies the individual $v_j$ and 0 otherwise. The entropy function $f_E$ measures the diversity within an ensemble [27]. Given an individual $x_j \in \Omega_V$, if half of the classifiers T/2 don't misclassify $x_j$ then the other half T-T/2 misclassifies it necessarily and vice versa. We speak in this case of maximum diversity.

We note $nc(x_j)$ the number of classifiers of T which correctly classify $x_j$, $nc(x_j) = \sum_{i=1}^{T} y_{ij}$. The entropy measure $f_E$ is written as:

$$f_E = \frac{1}{n}\sum_{j=1}^{n}\frac{1}{T-\frac{T}{2}}\min\{nc(x_j), T - nc(x_j)\}$$

$f_E \in [0, 1]$ where 1 indicates a very large diversity and 0 a lack of diversity. Thus, the goal is to maximize the $f_E$ function.

## 3.2   Genetic Algorithms

Genetic algorithms are a preferred technique for selection because they are inspired by natural selection. They generate individuals that optimize an evaluation function also called fitness function.

A genetic algorithm is defined by [28]:

- Individual also called chromosome or sequence represents a potential solution of the problem. In our case, a solution of the problem corresponds to a binary string of size T (corresponds to the number of trees composing the forest). A chromosome is noted $ch = (val_1 val_2 \ldots val_T)$ *where* $val_i = 1$ if the tree is present in the selected chromosome and 0 otherwise;
- Population corresponds to all the chromosomes representing all possibilities of 1 and 0 in a binary chain of size T;
- Environment represents the search space $|ER| = 2^T$.
- Fitness function corresponds to the function $f\!f_E = f_E$ ($f_E$ is the diversity function defined above). The goal is to maximize the value of $f\!f_E$.

Calculate the $f\!f_E$ fitness function for chromosome $ch_1$ is equivalent to calculate the function $f_E$:

$$f_E = \frac{1}{n}\sum_{j=1}^{n}\frac{1}{T-\frac{T}{2}}\min\{nc(x_j), T - nc(x_j)\}$$

For instance, if $n = 2 = |\Omega_v|$, T = 2 (the classifiers to which correspond the value 1 $T_1$ and $T_3$), $x_1$ and $x_2$ are the individuals of $\Omega_v$ classified respectively $(1\ 0)^t$ and $(0\ 1)^t$ by $T_1$ et $T_3$.

$nc(x_1) = 1$ (the number of trees that correctly classify instances $x_1$).

$nc(x_2) = 1$ (the number of trees that correctly classify instances $x_2$).

$f_E = \frac{1}{2}\left(\frac{1}{2-\frac{2}{2}} * \min(1, 2, -1) + \frac{1}{2-\frac{2}{2}} * \min(1, 2, -1)\right) = 1$ and $f\!f_E = f_E = 1$, $f\!f_E$ takes its minimum equal to 0 when the trees are diverse and its maximum 1 when they disagree; hence $f\!f_E \in [0\ 1]$.

We give hereafter the EMnGA algorithm that uses a genetic algorithm-based search strategy and an entropy-based fitness function. It is described by the following pseudo code:

*Algorithme $AG_{fE}$ ;*

*Input :*
*$C=\{C_1,...,C_i,...,C_T\}$: a heterogeneous ensemble of classifiers ;*
*$y_i=(y_{1i},...,y_{ni})^T$: a classification vector associated with $C_i$ on $\Omega_V$ ;*
*$\Omega_V$: a validation or selection sample;*
*$ch_i$: chromosome i of the search space;*
*$ff_E(ch,\Omega_V)$ : fitness function ;*
*Output :*
*Ch_sol : solution chromosome;*
*Begin*
*Generate a population of bits of size T;*
*Evolve the population where the fitness of a chromosome $ch_i$ is calculate by $ff_E(ch_i,\Omega_V)$ ;*
*$ch\_sol :=argmax_{chi}(ff_E(ch_i,\Omega_V))$*
*End.*

## 4   Experiments and Results

In this section, we describe information about the datasets used to carry out our experiments. We experienced 8 benchmarks datasets downloaded from the UCI Repository [29] as depicted in Table 1.

**Table 1.**  Description of datasets

| Dataset | Size | Variables | Classes |
|---|---|---|---|
| cmc | 1473 | 9 | 3 |
| kr-vs-kp | 3196 | 36 | 2 |
| credit-g | 1000 | 21 | 2 |
| tic-tac-toe | 958 | 9 | 2 |
| vehicle | 946 | 18 | 4 |
| vowel | 990 | 13 | 11 |
| Hypothyroïde | 3772 | 30 | 4 |
| segment | 2310 | 20 | 7 |

All data sets contain enough data to split them into three samples: a learning sample $\Omega_L$ (40% of the initial sample size), a validation or selection sample (20% of the size of the initial sample) and the remaining 20% are used for the test.

We adopt a similar approach to that proposed in [26]. We generate, for each dataset, a heterogeneous ensemble of 200 models (classifiers) containing:

- 60 Kppv: with 22 values for K ranging from 1 to the size of the training sample. Three weighting methods are applied: no weighting, inverse weighting and similarity weighting;
- 110 SVM: we use 10 values for the complexity parameter $\{10^{-7}, 10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 0.1, 1, 10, 100\}$, and 10 different kernels (2 polynomials of degree 2 and 3 and, 8 radial with gamma $\in \{0.001, 0.005, 0.01, 0.05, 0.1, 0.5, 1, 2\}$;
- 2 naive bayes: one model is built with the default parameters and another with a kernel estimate;
- 4 decision trees: two values are used for the confidence factor $\{0.25, 0.5\}$;
- 24 MLPs with a single hidden layer containing 6 neurons.

For each dataset, the generation of the ensemble (composed of 200 models) is repeated 10 times and the majority vote is used for the combination of the different models.

In Table 2, we give the performance results or success rates obtained (calculated on an average of 10 iterations for each dataset) on the 8 datasets. Comparisons are made between the Initial ensemble EI, the method EMnGA, and FES [26].

**Table 2.** Success rates obtained respectively by EI, EMnGA and FES for the 8 datasets

|             | EI     | EMnGA     | FES    |
|-------------|--------|-----------|--------|
| cmc         | 47.6%  | **53.2%** | 52.7%  |
| kr-vs-kp    | 95.6%  | 98.55%    | **99%**|
| credit-g    | 70.8%  | **85.32%**| 74.4%  |
| tic-tac-toe | 63.9%  | 97.43%    | **98.7%**|
| vehicle     | 75.3%  | **83.02%**| 81.1%  |
| vowel       | 90.7%  | **98.5%** | 90.3%  |
| Hypothyroïde| 91.9%  | 98.6%     | **99.3%**|
| segment     | 97.8%  | **98.66%**| 96.9%  |
| Average     | 74%    | 86%       | 83%    |

The EMnGA method gives better performance compared to FES on 5 of the 8 datasets, with improvements ranging from 0.5% to 10.92%. FES is doing better than the proposed method on 3 dataset with improvements of 0.45% for kr-vs-kp and segment, and 1.27% for tick-to-toe. On average, across all datasets, EMnGA improves the performance of FES by 3% and the initial ensemble EI by 12% (Table 3).

We note that FES method generate ensembles with reduced sizes in 6 cases among the 8 with reductions ranging from 0.2 to 2.3. EMnGA reduces the size of the credit-g and tick-to-toe datasets by 0.1 and 1.8, respectively. On average, on the 8 datasets, FES allows a reduction of 0.5 compared to EMnGA.

**Table 3.** Sizes of ensembles obtained by EMnGA and FES on the 8 datasets

|              | EMnGA | FES  |
|--------------|-------|------|
| cmc          | 20.6  | **18.3** |
| kr-vs-kp     | 14    | **13.4** |
| credit-g     | **15.6** | 15.7 |
| tic-tac-toe  | **26.1** | 27.9 |
| vehicle      | 10.3  | **8.6** |
| vowel        | 8.7   | **8.5** |
| hypothyroïde | 12.5  | **11.7** |
| segment      | 21.3  | **20.5** |
| Average      | 15.9  | 15.4 |

## 5   Conclusion and Future Work

In this paper we proposed an ensemble selection method using diversity based entropy measurement combined to a genetic algorithm based strategy search. The proposed method was compared with the FES method [27] which uses a forward selection course for heterogeneous ensembles. Based on the comparison made against the criteria of performance and size of the ensemble generated, we have observed that the EMnGA method generate larger ensembles in most cases compared to FES due mainly to the strategy search method used to explores more possibilities but gives better performance.

In future work, we will compare the measure with other methods in the literature using other search strategies and will use more datasets in order to formally validate our results with appropriate statistical tests. We will also experience a measure based on diversity and performance in order to see the impact of jointly using both criteria. Finally, we intend to apply the EMnGA method in random forest ensemble simplification.

## References

1. Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J.: Classification and Regression Trees. Chapman & Hall, New York (1984)
2. Geurts, P.: Contributions to decision tree induction: bias/variance tradeoff and time series classification. Ph.D. thesis, Department of Electronical Engineering and Computer Science, University of Liège, May 2002
3. Margineantu, D., Dietterich, T.: Pruning adaptive boosting. In: Proceedings of the 14th International Conference on Machine Learning, pp. 211–218 (1997)
4. Dietterich, T.G.: Ensemble methods in machine learning. In: Kittler, J., Roli, F. (eds.) MCS 2000. LNCS, vol. 1857, pp. 1–15. Springer, Heidelberg (2000). https://doi.org/10.1007/3-540-45014-9_1
5. Banfield, R.E., Hall, L.O., Bowyer, K.W., Kegelmeyer, W.P.: A comparison of decision tree ensemble creation techniques. IEEE Trans. Pat. Recog. Mach. Int. **29**, 173–180 (2007)

6. Bian, S., Wang, W.: On diversity and accuracy of homogeneous and heterogeneous ensembles. Intl. J. Hybrid Intell. Syst. **4**, 103–128 (2007)

7. Breiman, L.: Bagging predictors. Mach. Learn. **26**(2), 123–140 (1996)

8. Schapire, R.E.: The strength of weak learnability. Mach. Learn. **5**(2), 197–227 (1990)

9. Tang, E.K., Suganthan, P.N., Yao, X.: An analysis of diversity measures. Mach. Learn. **65**(1), 247–271 (2006)

10. Breiman, L.: Random forests. Mach. Learn. **45**, 5–32 (2001)

11. Banfield, R.E., Hall, L.O., Bowyer, K.W., Kegelmeyer, W.P.: Ensemble diversity measures and their application to thinning. Inf. Fusion **6**(1), 49–62 (2005)

12. Martinez-Munoz, G., Suarez, A.: Aggregation ordering in bagging. In: International Conference on Artificial Intelligence and Applications (IASTED), pp. 258–263. Acta Press (2004)

13. Martinez-Munoz, G., Suarez, A.: Pruning in ordered bagging ensembles. In: 23rd International Conference in Machine Learning (ICML-2006), pp. 609–616. ACM Press (2006)

14. Caruana, R., Niculescu-Mizil, A., Crew, G., Ksikes, A.: Ensemble selection from libraries of models. In: Proceedings of the 21st International Conference on Machine Learning (2004)

15. Partalas, I., Tsoumakas, G., Katakis, I., Vlahavas, I.: Ensemble pruning using reinforcement learning. In: Antoniou, G., Potamias, G., Spyropoulos, C., Plexousakis, D. (eds.) SETN 2006. LNCS (LNAI), vol. 3955, pp. 301–310. Springer, Heidelberg (2006). https://doi.org/10.1007/11752912_31

16. Partalas, I., Tsoumakas, G., Vlahavas, I.: An ensemble uncertainty aware measure for directed hill climbing ensemble pruning. Mach. Learn. **81**, 257–282 (2010)

17. Fan, W., Chu, F., Wang, H., Yu, P.S.: Pruning and dynamic scheduling of cost- sensitive ensembles. In: Eighteenth National Conference on Artificial Intelligence, American Association for Artificial Intelligence, pp. 146–151 (2002)

18. Zhang, Y., Burer, S., Street, W.N.: Ensemblepruning via semi-definite programming. J. Mach. Learn. Res. **7**, 1315–1338 (2006)

19. Hernández-Lobato, D., Martínez-Muñoz, G.: A statistical instance-based pruning in ensembles of independant classifiers. IEEE Trans. Pattern Anal. Mach. Intell. **31**(2), 364–369 (2009)

20. Dai, Q.: An efficient ensemble pruning algorithm using one-path and two-trips searching approach. Knowl.-Based Syst. **51**, 85–92 (2013). ISSN 0950-7051

21. Bhatnagar, V., Bhardwaj, M., Sharma, S., Haroon, S.: Accuracy-diversity based pruning of classifier ensembles. Prog. Artif. Intell. **2**(2–3), 97–111 (2014)

22. Guo, Y., et al.: A novel dynamic rough subspace based selective ensemble. Pattern Recognit. **48**(5), 1638–1652 (2015). ISSN 0031-3203

23. Partalas, I., Tsoumakas, G., Vlahavas, I.: A study on greedy algorithms for ensemble pruning. Technical report TR-LPIS-360-12, Department of Informatics Aristotle University of Thessaloniki, Greece (2012)

24. Haque, M.N., Noman, N., Berretta, R., Moscato, P.: Heterogeneous ensemble combination search using genetic algorithm for class imbalanced dataclassification. PLoS ONE **11**(1), e0146116 (2016). https://doi.org/10.1371/journal.pone.0146116

25. Pölsterl, S., et al.: Heterogeneous ensembles for predicting survival of metastatic, castrate-resistant prostate cancer patients. F1000Research **5**(2676) (2016). *PMC*.Web 13 October 2017

26. Partalas, I., Tsoumakas, G., Vlahavas, I.: Focused ensemble selection: a diversity- based method for greedy ensemble selection. In: Ghallab, M., Spyropoulos, C.D., Fakotakis, N., Avouris, N.M. (eds.) ECAI 2008 – Proceedings of the 18th European Conference on Artificial Intelligence. Frontiers in Artificial Intelligence and Applications, Patras, Greece, vol. 178, pp. 117–121 (2008)
27. Kuncheva, L.I., Whitaker, C.J.: Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. Mach. Learn. **51**, 181–207 (2003)
28. Lerman, I., Ngouenet, F.: Algorithmes génétiques séquentiels et parallèles pour une représentation affine des proximités, Rapport de Recherche de l'INRIA Rennes - Projet REPCO 2570, INRIA (1995)
29. Asuncion, A., Newman, D.: UCI machine learning repository (2007). http://www.ics.uci.edu/»mlearn/MLRepository.html