# Cross-Entropy Based Ensemble Classifiers

**Giovanni Lafratta**

**Abstract** Multiple classification rules are simultaneously identified by applying the Cross-Entropy method to the maximization of accuracy measures in a supervised learning context. Optimal ensembles of rules are searched through stochastic traversals of the rule space. Each rule contributes to classify a given instance when the observed attribute values belong to specific subsets of the corresponding attribute domains. Classifications of the various rules are combined applying majority voting schemes. The performance of the proposed algorithm has been tested on some data sets from the UCI repository.

**Keywords** Big data analytics · Classification · Stochastic data mining · Supervised machine learning

## 1 Introduction

The Cross-Entropy method raises in the context of rare-event probability estimation [4]: importance sampling is exploited to define a sequence of estimators, each one based on a sample drawn from a parametric distribution under which the event hopefully becomes more and more probable than under the distribution of interest. In this way, additional estimators having a variance less than that of the previous ones are added to the sequence. New estimators are selected by minimizing the Kullback-Leibler cross-entropy between their importance sampling distributions and the theoretical one under which the estimation variance would be zero [6]. Soon the method was recognized as closely related to the optimization of continuous or discrete functions [5]. In fact, optimizing a function can be seen as the problem of estimating the

G. Lafratta(✉)
Department of Philosophical, Pedagogical and Economic-Quantitative Sciences,
"G. d'Annunzio" University, Viale Pindaro 42, 65127 Pescara, Italy
e-mail: giovanni.lafratta@unich.it

E. Buciarelli et al. (eds.), *Decision Economics, In Commemoration of the Birth Centennial of Herbert A. Simon 1916-2016*, Advances in Intelligent Systems and Computing 475,
DOI: 10.1007/978-3-319-40111-9_6

probability of the event represented by the function assuming a value over a given extreme threshold, an event that can be typically interpreted as rare.

In this paper, we propose to apply the Cross-Entropy method to the maximization of classification accuracy measures. Optimality is defined in terms of ensembles of classification rules [3], each rule applying to a given instance when the observed attribute values belong to specific subsets of the corresponding attribute domains. The method builds candidate ensembles simultaneously, in the sense that rules are not added to a candidate ensemble as a function of the rules already in the ensemble under construction. As a consequence, the algorithm scales well in both on-premises and cloud distributed computing environments.

This paper is organized as follows. In Sect. 2 we define the space of classification rules which will be searched for the optimal ensemble. Section 3 describes the steps of the proposed algorithm, whose performance is illustrated in Sect. 4 by investigating its application to the analysis of some data sets from the UCI Repository [2]. Finally, Sect. 5 contains some concluding remarks.

## 2 Ensembles of Classification Rules

Let us assume that instances from a given space $\mathcal{X}$ must be classified into a set $\mathcal{Y}$ of labels. If $m$ input attributes are taken into account, and if attribute $j \in \{1, \dots, m\}$ has finite domain $D_j$, then $\mathcal{X}$ can be represented as the Cartesian product $D_1 \times \cdots \times D_m$. A *classification rule* $R$ is a pair $(\mathcal{C}, y)$, where $\mathcal{C}$ is a proper subset of $\mathcal{X}$ and $y \in \mathcal{Y}$. Given an instance $\mathbf{x} \in \mathcal{X}$, the rule will classify $\mathbf{x}$ as $y$ if and only if $\mathbf{x} \in \mathcal{C}$. When this is the case, rule $R$ is said to *vote to classify* $\mathbf{x}$ as $y$. If an ensemble of rules $R_k = (\mathcal{C}_k, y_k)$ is available, with $k$ in a finite set $\mathcal{K}$, then unlabeled instances can be classified by performing majority vote procedures. If weight $w_k > 0$ is assigned to rule $R_k$, $k \in \mathcal{K}$, then the classification of a given instance $\mathbf{x}$, say $y(\mathbf{x})$, can be defined as satisfying

$$y(\mathbf{x}) = \arg\max_{y \in \mathcal{Y}} \sum_{k \in \mathcal{K} : \mathbf{x} \in \mathcal{C}_k \wedge y = y_k} w_k ,$$

where ties are eventually resolved by selecting one of the maximizing arguments at random.

## 3 Cross-Entropy Algorithm for Accuracy Maximization

The Cross-Entropy method can be exploited to search the rule space and create ensembles, among those having a fixed size $n$, which correspond to maximal accuracy. The algorithm iteratively executes two main steps. The first one, the *sampling step*, is responsible for generating diverse candidate ensembles of classification rules. In this step, ensembles are interpreted as points of a sample drawn from a distribution defined on the Cartesian product of $n$ replicates of the rule space, each replicate being independent from the others. Sampling a rule $R_k = (\mathcal{C}_k, y_k)$ from the $k$-th replicated space is executed as follows. Firstly, a finite discrete distribution $p_{\mathcal{Y},k}$ is defined on the label domain $\mathcal{Y}$ and sampled to obtain label $y_k$. Secondly, the rule condition $\mathcal{C}_k$ must be sampled too. This is equivalent to select at random, for each input attribute $j$, a subset $C_{j,k}$ of domain $D_j$, and hence define $\mathcal{C}_k = C_{1,k} \times \cdots \times C_{m,k}$: to obtain such result, a Bernoulli distribution, say $p_{j,i,k}$, is assigned to each value $i$ in the attribute domain $D_j$: sampling $x_{j,i,k}$ from $p_{j,i,k}$ enable us to set $i \in C_{j,k}$ if and only if $x_{j,i,k} = 1$.

At first iteration, each rule has the same probability to be included in a sample point (ensemble); this is accomplished by defining a balanced Bernoulli distribution for each value in each input attribute domain, and a uniform finite discrete distribution on the label domain $\mathcal{Y}$. The *update step* aims to modify the distribution from which the samples will be obtained in the next iteration, in order to improve the probability of sampling better performing rules. Let us represent the sampled ensembles in the current iteration as $\mathcal{E}_1, \ldots, \mathcal{E}_N$. The corresponding accuracies, say $a_1, \ldots, a_N$, are computed and sorted in increasing order, $a_{(1)} \leq \cdots \leq a_{(N)}$. In this way, the ensembles which guarantee the best performances can be identified as those whose accuracies occupy the last positions in the ordering. The Cross-Entropy method refers to such sample points as the *elite samples*. A way to define them exactly is as follows. A parameter $\rho$ is set to a value in the interval $[.01, .1]$ and all those samples occupying a position greater than or equal to $\eta = \lceil (1 - \rho)\, N \rceil$ are defined as elite ones. Given the elite ensembles, $\mathcal{E}_{(\eta)}, \ldots, \mathcal{E}_{(N)}$, the joint distribution from which the ensembles will be sampled in the next iteration is updated by estimating its parameters using the elite ensembles only. Since they correspond to better accuracies, this updating mechanism guarantees that rules having better performances will be included in subsequent samples with greater probabilities. The algorithm stops if the accuracy *level* which defines the elite samples points, i.e. $a_{(\eta)}$, remains unchanged for a predefined number of consecutive iterations.

## 4 Empirical Evidence

The proposed inducer has been tested against the *Congressional Voting Records* (CVR) and *Nursery* UCI data sets, using the training set sizes described in Table 1. Let $m$ represent one of those sizes for a given data set; a corresponding experiment has been designed to estimate classification accuracies following [1]: the data set has been randomly partitioned into two subsets, the first, referred to as the *universe*, has size $2m$ and is the one from which we draw 10 samples, having size $m$, which represent the training sets passed to the inducer; the second one, referred to as the *test* subset, is intended to evaluate the accuracy of the induced classifiers. All ensembles were formed by 100 classification rules, with uniform weights equal to 1. Further parameters of the Cross-Entropy algorithm were set as follows: at each iteration, a sample of $N = 1000$ ensembles was drawn and the elite ensembles have been

**Table 1** Data sets under study and corresponding training set sizes

| Data set | Data set size | Attributes | Classes | Training set sizes |
|---|---|---|---|---|
| CVR | 435 | 16 | 2 | 50, 100, 150, 200 |
| Nursery | 12960 | 8 | 5 | 500, 1000, 2000, 3000 |

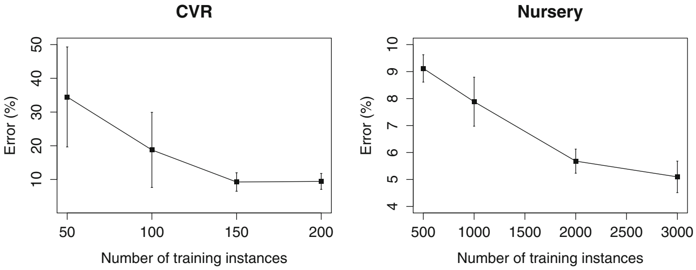

**Fig. 1** Learning curves of 100 classification rules for data sets CVR and Nursery

identified by setting $\rho = .1$; finally, the stopping criterion came into effect if the *level* did not change for 10 iterations. Results for data sets *CVR* and *Nursery* are shown in Fig. 1. The points represents the mean error rate of the 10 training runs, while the error bars show one standard deviation of the estimated error. It can be noted that, for data set *CVR*, stabilization occurs at about 150 instances, while for data set *Nursery* it occurs approximately at 2000 instances.

## 5 Conclusions

In this paper, the Cross-Entropy method has been exploited to induce ensembles of classification rules. Observed accuracies deserve further investigations in order to detect possible enhancements based on other than uniform majority voting schemes. Given that a by-product of the algorithm is the estimated distribution of the target attribute conditional to each rule, the votes can be easily modulated by taking into account the variability of such distributions: for example, one can assign greater weights to rules whose corresponding distributions are less heterogeneous.

We also expect to devote future researches to study how the accuracy relates to the size of the induced ensembles, especially when applied to high dimensional data sets.

## References

1. Kohavi, R., Wolpert, D.H.: Bias plus variance decomposition for zero-one loss functions. In: Saitta, L. (ed.) Machine Learning: Proceedings of the Thirteenth International Conference, Morgan Kaufmann, pp. 275–283 (1996)

2. Lichman, M.: UCI Machine Learning Repository. University of California, School of Information and Computer Science, Irvine, CA (2013). http://archive.ics.uci.edu/ml
3. Rokach, L.: Ensemble-based classifiers. Artificial Intelligence Review **33**, 1–39 (2010)
4. Rubinstein, R.Y.: Optimization of computer simulation models with rare events. European Journal of Operational Research **99**, 89–112 (1997)
5. Rubinstein, R.Y.: The cross-entropy method for combinatorial and continuous optimization. Methodology and Computing in Applied Probability **2**, 127–190 (1999)
6. Rubinstein, R.Y., Kroese, D.P.: The cross-entropy method: A unified approach to combinatorial optimization, Monte-Carlo simulation, and machine learning. Springer (2004)