Journal of Computational Science 20 (2017) 154–161

Contents lists available at ScienceDirect

Journal of Computational Science

journal homepage: www.elsevier.com/locate/jocs

Using Anticipative Hybrid Extreme Rotation Forest to predict emergency service readmission risk

Arkaitz Artetxe [a], Borja Ayerdi [b,c], Manuel Graña [b,c,*], Sebastian Rios [d]

[a] Vicomtech-IK4 Research Centre, Mikeletegi Pasealekua 57, 20009 San Sebastian, Spain
[b] Computer Intelligence Group, UPV/EHU, Dept. CCIA, Paseo Manuel Lardizabal, 20018 San Sebastian, Spain
[c] ACPySS, Paseo Manuel Lardizabal, 20018 San Sebastian, Spain
[d] Business Intelligence Research Center (CEINE), Industrial Engineering Department, University of Chile, Beauche 851, Santiago 8370456, Chile

ARTICLE INFO

Article history:
Received 24 September 2016
Received in revised form
30 November 2016
Accepted 23 December 2016
Available online 27 February 2017

Keywords:
Ensemble classifiers
Adaptive ensembles
Emergency readmission prediction

ABSTRACT

This paper provides a real life application of the recently published Anticipative Hybrid Extreme Rotation Forest (AHERF), which is an heterogeneous ensemble classifier that anticipates the correct fraction of instances from each basic classifier architecture to be included in the ensemble. Heterogeneous classifier ensembles aim to profit from the diverse problem domain specificities of each classifier architecture in order to achieve improved generalization over a larger spectrum of problem domains. Given a problem dataset, anticipative determination of the desired ensemble composition is carried out as follows: First, we estimate the performance of each classifier architecture by independent pilot cross-validation experiments on a small subsample of the data. Next, classifier architectures are ranked according to their accuracy results. The likelihood of each classifier architecture instance appearing in the ensemble is computed from this ranking. Finally, while building the ensemble, the architecture of each individual classifier is decided by sampling this likelihood probability distribution. In this paper we provide an application of AHERF to a real life problem. Readmission of patients short time (i.e. 72 h) after being released poses a great economical and social challenge, so that many efforts are being addressed to predict and avoid readmission events. We present the results of the application of AHERF over a real life dataset composed of 156,120 admission cases recorded between January 2013 and August 2015. AHERF archives results over or close to 70% sensitivity in the prediction of readmissions for adults and pediatric cases, suggesting that it can be used to build institution specific prediction systems.

1. Introduction

A well known principle in machine learning is that we can not expect a classifier architecture to outperform all others over all problem domains, which has been stated as the *no free lunch* theorem [1,2]. Ensembles of classifiers aim to improve classification results by the combination of weak and diversified classifiers [3,4]. Early ensemble of classifiers were homogeneous, such as the Random Forest, which combines the outputs of a collection of Decision Trees (DT) by majority voting built from bootstrapped training data over random variable selections [5,6], the ensembles of Support Vector Machines (SVM) [7], or the ensembles of bootstrapped dendritic classifiers [8]. In ensemble design, there is an emphasis on diversification expecting that classifiers with quite different domains of expertise may have a synergistic effect on the whole. Ensemble classifiers assume that individual classifier error follows a symmetric (if not Gaussian) distribution around zero (unbiased), so that the joint effect of the ensemble is to cancel the individual classifier errors, assuming that their combination is additive. Classifier diversification procedures include the preprocessing of the data by randomized rotation, as in Rotation Forests [9], and to build heterogeneous ensembles, where individual classifiers have different architectures, so that heterogenous sensitivity of the learning methods to the data distribution may be exploited.

In this line of work, the Hybrid Extreme Rotation Forest (HERF) [10] was introduced as an heterogeneous ensemble of Extreme Learning Machines (ELM) [11–14] and DTs trained over random rotations of bootstrapped subsampled data. HERF has been shown to be successful in remote sensing and medical image segmentation [15]. Recently, an anticipative HERF (AHERF) for the selection of the ensemble components has been proposed [16], which has an extension of the pool of elementary classifier architectures as well. The procedure is based on the estimation of the likelihood of including

* Corresponding author at: Computer Intelligence Group, UPV/EHU, Dept. CCIA, Paseo Manuel Lardizabal, 20018 San Sebastian, Spain.
E-mail address: manuel.grana@ehu.eus (M. Graña).

http://dx.doi.org/10.1016/j.jocs.2016.12.008
1877-7503/

an instance of a classification architecture in the ensemble on the basis of the performance achieved on a subsample of the dataset. Hence, building the ensemble involves sampling of this probability distribution to decide the architecture of each individual classifier added to the ensemble. Individual classifier specialization is enhanced by this anticipative modeling and sampling [4]. That is, the AHERF may have quite different architecture composition depending on the problem. Note that some ensembles, such as Random Forests and AdaBoost, are used as elementary classifiers in AHERF, so that the realization of AHERF in this paper is indeed an ensemble of ensembles. The approach has already been successfully applied to hyperspectral image classification [17].

This paper aims to consolidate the practical value of AHERF by an application to a real life problem for which we have a large dataset. Hospital readmissions in a short period of time after a previous discharge, are indicative of either a bad quality of healthcare service, or structural problems in the healthcare systems, such as chronic patients being attended in the emergency department (ED) for lack of a proper planning of their care. There is a growing need for sensitive predictive tools, some specifically related with the patients treated at ED [18], others covering general healthcare services [19], or institution specific prediction models [20] some focused on specific fragile populations [21–23]. The anonymized dataset used in the computational experiments on readmission risk covers more than three years of the activity of the ED in a university hospital of Santiago in Chile, containing over 150,000 records. Previous prediction works on readmission risk were focused on a small sample of older patients [18], while the dataset in this paper includes adult and pediatric patients, which have quite different patterns of attention and readmission. The goal set in this paper is to achieve an acceptable prediction performance on this highly imbalanced dataset. We compare the AHERF with other state of the art classifiers in order to assess the improvement achieved by AHERF.

The paper is structured as follows: Section 2 reviews the issue of readmission risk prediction. Section 3 summarizes the description of individual classifier architectures used in the AHERF. Section 4 gives the description of the AHERF. Section 5 provides a discussion of the causes for AHERF improved performance. Section 6 discusses the experimental design details. Section 7 describes the emergency department dataset. Section 8 gives the experimental results and a discussion of their significance. Section 9 gives our conclusions and future work.

2. Readmission risk prediction

Prediction of readmission risk has been approached from a variety of angles, a recent review is [24], many of them taking into account specific subpopulations. The approaches often combine several administrative, demographic, biochemical and other test, such as psychological tests, to compute a risk index. Differences between centers in electronic medical records and recorded patient information lead to institution specific models, i.e. trained with institution specific data [20], or specific healthcare networks [19] Often indices are limited to specific subpopulations, such as people suffering from Chronic Obstructive Pulmonary Disease [21] or Acute Myocardial Infarction [25].

The highest rates of ED readmission, the longest stays, and greatest resources invested in ancillary tests correspond to adults above 75 years of age [22,26]. Despite this intense use of resources, these patients often leave the ED unsatisfied, with poorer clinical outcomes, and higher rates of misdiagnosis and medication errors compared to younger patients. Additionally, they have a higher risk of ED readmission, hospitalization, death and institutionalization [27]. Readmission risk prediction is therefore critical for this kind of patients [28]. Specific traits, such as medication regime are less predictive than expected, but clustering patients have been found to improve prediction [23].

The LACE readmission index [29] is based on logistic regression analysis on a set of 48 variables collected from 4812 patients from several Canadian hospitals. A variant called LACE + [30] makes use of variables drawn from administrative data. Closely related to LACE, HOMR (Hospital patient One-year Mortality Risk) [31] is a model for predicting death within one year after hospital admission. According to the authors the goal is to predict long-term survival after admission to hospital. The variables used are included in the following categories: Demographics (age, sex, etc.) Health status (Charlson comorbidity index, number of visits to hospital emergency, etc.) Acuity disease (emergency admissions, direct ICU admissions, etc.) The dataset used for the development and validation of this model consists of more than three million instances obtained from several hospitals in the areas of Ontario, Alberta and Boston.

3. Elementary classifiers

Elementary classifiers implementation in the experiments reported in this paper are extracted from SciKit[1] Python package [32], often using default parameter settings. Classifier definitions are well known in the literature, so we will provide a summary overview of them. The Python implementation of AHERF is available at [2] .

3.1. Decision trees and random forests

Decision Trees (DT) [33,34] are built by recursive partitioning of the data space using a quantitative criterion (e.g., mutual information, gain-ratio, gini index), maybe followed by a pruning process to reduce overfitting. Tree leaves correspond to the probabilistic assignment of data samples to classes. Ensembles of DT classifiers such as Bagging [5] and Random Forests [6] were early proposals of ensemble structures. Random Forests are ensembles of DT, where each individual DT is built on a bootstrapped training data subset over a random subset of the input variables. The majority voting rule applied to the ensemble of outputs decides the input data class assignment.

3.2. ELM

Extreme Learning Machines (ELM) [11,35,13] was proposed as a very fast training algorithm for single-layer feedforward neural networks (SLFN). The ELM avoids gradient descent of the hidden layer weights by performing a random sampling, equivalent to a random subspace projection. The training problem reduces to the estimation of the output weights by linear least squares resolution of the network response minimizing the classification error, often solved by the Moore-Penrose generalized pseudo-inverse. Randomization of hidden layer weights introduce training instability which has been tackled in many ways. Ensembles of ELM, such as the Voting ELM [36–38], and the HERF [10,15], help improve the training stability. The sought effect is that the individual classifier errors compensate in the limit, when the ensemble size grows, assuming that the probability distribution of the individual classifier error is symmetric around zero.

[1] http:// scikit- learn. org/ stable/ supervised_ learning. html#supervised-learning.

[2] http:// www. ehu. es/ ccwintco/ index. php/ GIC- source- code- free- libre.

3.3. Support Vector Machines

Support Vector Machines (SVM) [39,40] look for the set of support vectors that allow to build the optimal discriminating surface in the sense of providing the greatest margin between the classes. When no linear separation of the training data is possible, the kernel trick transforms the hyperplane defining the SVMs into a non-linear decision boundary in the feature space [40]. Model selection in SVM involves the selection of the appropriate kernel function as well as the fine tuning of its parameters, which not trivial task.

3.4. k-Nearest Neighbors

k-Nearest Neighbors (k-NN) is the simplest formulation of the supervised training, where the training samples are used as class prototypes. The class assigned to the test input pattern is the result of majority voting on the K closest training patterns according to some distance in pattern space, often the Euclidean distance.

3.5. Adaboost

Adaptive Boosting (AdaBoost) [41,42] iterates the following process. First, a collection of weak classifiers is trained. Second, the best classifier is selected and added to the ensemble. A weight of its importance in the final decision is computed. Finally, the data samples are weighted according to their misclassification rate by the ensemble classifiers for the next iteration. The result is a pipeline of classifiers of increasing specialization in the most difficult instances.

3.6. Gaussian Naive Bayes

Naive Bayes methods follow from the "naive" assumption that the components of the pattern vectors are statistically independent, so that the posterior probability of the class can be approximated by a product decomposition of into the likelihood of the individual features. The Gaussian Naive Bayes assumes that the likelihood follows a Gaussian distribution, where the mean and standard deviation of each feature is estimated from the sample.

4. Adaptive Hybrid Extreme Rotation Forest

Let $\mathbf{x} = [x_1, \ldots, x_n]^T$ be a data sample described by n feature variables, where F is the feature set and X is the validation dataset containing N samples, which can be stored in a matrix of size $n \times N$. Let Y be a vector containing the class labels of the data samples, $Y = [y_1, \ldots, y_N]^T$. The number of classes is denoted Ω. Denote by D_1, $\ldots$, D_L the classifiers in the ensemble that can be trained in parallel.

The Anticipative Hybrid Extreme Rotation Forest (AHERF) algorithm training and testing phases are summarized in Algorithm 1. We specify the training and test phases of each cross-validation fold. For training, first, a model selection phase is performed, where 30% of the training data is used. This size of model selection data is a balance between an appropriate sampling of the data distribution and allowing data for ensuing ensemble training and testing, because model selection data can not be reused for ensemble cross-validation. For each classifier type described in the previous section, a 5-fold cross-validation is carried out on the model selection data (line M3). The individual model selection cross-validation average accuracies are ranked, so that r_k is the ranking value of the kth classifier type (line M4). Then (line M5), each classifier is assigned a selection probability according to the expression $p_k = \frac{Fib((C+1)-r_k)}{\sum_{i=1}^{C} Fib(i)}$, where $Fib(i)$ is the ith value of the Fibonacci series. Fig. 1 shows the plot of the probability distribution computed from the ranking of the classifier types.

The ensemble strategy cross-validation is carried out on the remaining 70% data, involving a 10-fold cross-validation process. Notice that the test data size at each fold is reduced to a 7% of the available data, hence larger model selection data can not be afforded because of the risk of test data misrepresenting the actual data distribution. The following steps are carried out at each fold: for each classifier D_i in the ensemble the first step is the construction of the randomized rotation matrix (line 3) which requires the random partition of the set of features into a K subsets (line 4). For each subset of features $F_{i,j}$, the algorithm extracts the corresponding sample values in a matrix $X_{i,j}$ (line 6), used to build a component $C_{i,j}$ rotation matrix (line 7). The randomized rotation matrix R_i^{α} is built by composing the component rotation matrices reordering the columns in order to match the original variable ordering, as detailed in [16]. Next (line 9) there is a random decision on the type of the classifier, using the selection probabilities $\{p_k\}$ (built in line M5). Finally, the D_i classifier is trained on the rotated data. In the test phase, a new vector $\mathbf{x}^{test}$ is first applied each classifier in the ensemble, obtaining a class hypothesis d_i, (line C2). Majority voting is implemented as follows: the counter c_ω has the number of classifiers that have casted their vote for class ω, (line 3, where $\delta_{i,j}$ is the Kronecker's delta function). Finally, the class with the maximum votes is selected (line C4) and returned as the classification result.

Algorithm 1. Anticipative tuning of Hybrid Extreme Rotation Forest (AHERF)

Training Phase
Given
X: z-scores of input dataset ($n \times N$ matrix).
Y: the labels of the dataset ($1 \times N$ matrix)
L: the number of classifiers in the ensemble
K: the number of feature subsets
Begin
Anticipative Model selection
M1 Select 30% of the dataset for model selection
M2 For each classifier type $k = 1, \ldots, M$
M3 Perform 5-fold cross-validation, obtain accuracy A_k
M4 Rank A_k, assigning r_k to the k-th classifier
M5 Assign selection probability $p_k = \frac{Fib((C+1)-r_k)}{\sum_{i=1}^{C} Fib(i)}$, $k = 1, \ldots, M$
On the 70% unused data, perform 10-fold cv, at each fold:
Ensemble construction on each training fold
2 For each individual classifier D_i, $i = 1 \ldots L$
3 Computation of rotation matrix R_i^{α}:
4 Partition F into K random subsets: $F_{i,j}; j = 1 \ldots K$
5 For each F_{ij}, $j = 1 \ldots K$
6 – Let $X_{i,j}$ be the subset of X corresponding to features in $F_{i,j}$.
7 – $C_{i,j}$ obtained from PCA on $X_{i,j}$
8 Compose R_i^{α} using matrices $C_{i,j}$.
9 Decide the model of D_i sampling $\{p_k; k = 1, \ldots, M\}$
10 Train classifier D_i on training set $(R_i^{\alpha} X, Y)$ or (X, Y)
End ensemble construction
Test on each testing fold
Let Ω be number of classes
C1 For each unknown $\mathbf{x}^{test}$ z-scores.
C2 $d_i = D_i(R_i^{\alpha} \mathbf{x}^{test})$; $i = 1, \ldots, L$
C3 $c_\omega = \sum_{i=1}^{L} \delta_{d_i,\omega}$; $i = 1, \ldots, L$
C4 $c^{test} = \underset{\omega}{argmax}\{c_\omega, \omega = 1, \ldots, \Omega\}$

5. Rationale for AHERF

The work on the design of heterogeneous ensembles of classifiers is motivated by the well known no-free lunch theorems [2,1], which state that no single machine learning approach is optimal for all instances of classification and regression problems. Therefore, the strategy of AHERF is to predict which kind of classifier architecture is better suited for the problem domain at hand. In fact, we rank the classifier architectures according to their estimated performance on the dataset, so that this ranking guides the selection of the architecture for the individual classifiers.

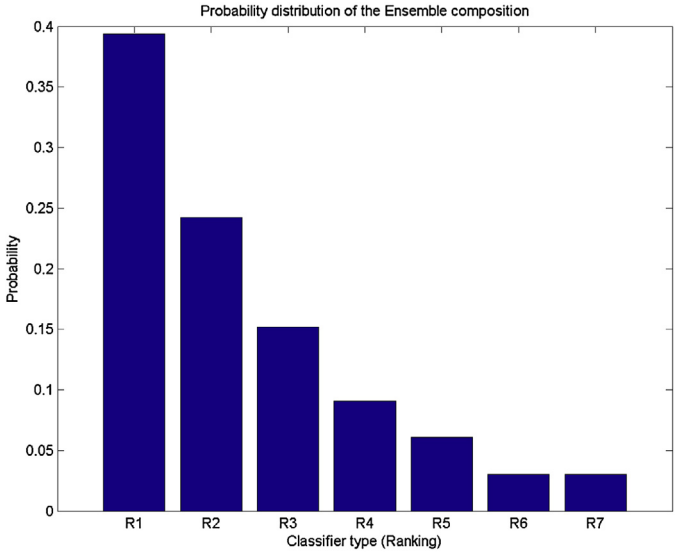

Fig. 1. The architecture selection probability distribution of classifiers ranking. R1 to R7 denote the rank resulting from the model selection.

The idea in AHERF is to build an ensemble where the best fitted classifier types are more frequent. Suppose we have a problem domain characterized by a ground truth classification mapping $\mathcal{C} : \mathcal{X} \rightarrow \Omega$, that gives the true class $\omega \in \Omega$ corresponding to each input feature vector $\mathbf{x} \in \mathcal{X}$. Any classifier C^t that we may built from a collection of input/output patterns $X = \{(\mathbf{x}_i, \omega_i)\}_{i=1}^{N}$, where $t \in T$ denotes the classifier type from a collection of methods T, i.e. the classification building type, provides us with its best estimation of the true class $\hat{\omega} = C^t(\mathbf{x})$. We can safely say that this estimation is given somehow as a maximum *a posteriori* estimation, i.e.

$$\hat{\omega} = \max_{\omega} \hat{P}^t(\omega|\mathbf{x}), \tag{1}$$

where t denotes the classifier architecture, and $\{\hat{P}^t(\omega|\mathbf{x})\}_{\omega \in \Omega}$ denotes the data driven estimation of the *a posteriori* probabilities by classifier C^t. The accuracy of a classifier can be computed as the expectation of the distance between the a posteriori distribution and the ground truth classification:

$$A^t = E_{\mathcal{X}}\left[\left|\left|\left[\hat{P}^t(\omega|\mathbf{x}) - \mathcal{C}(\omega, \mathbf{x})\right]_{\omega}\right|\right|\right], \tag{2}$$

where $E_{\mathcal{X}}[\,\cdot\,]$ denotes the expectation over the input space, i.e. over all possible sampling processes providing the training dataset X, and $\mathcal{C}(\omega, \mathbf{x})$ is 1 for the true class, and 0 for the others. The cross-validation experiments are a minimum variance way to provide estimates of the accuracy.

If we have an ensemble of classifiers $\{C_k^t\}_{k=1}^{M}$, then we will have as many *a posteriori* distribution estimations

$$\left\{\left\{\hat{P}_k^t(\omega|\mathbf{x})\right\}_{\omega}\right\}_{k=1}^{M} \tag{3}$$

as classifiers. If the ensemble decision is by majority voting, such as in AHERF, then the ensemble class estimation is given by

$$\hat{\omega} = arg\max_{\omega} |\{k|\omega = \hat{\omega}_k\}|, \tag{4}$$

where $\hat{\omega}_k = \max_{\omega} \hat{P}^t{}_k(\omega|\mathbf{x})$. In a broad sense, we can say that the accuracy of the ensemble can be modeled by

$$A_M \propto E_{\mathcal{X}}\left[\sum_{k} \left|\left|\left[\hat{P}^t{}_k(\omega|\mathbf{x}) - \mathcal{C}(\omega, \mathbf{x})\right]_{\omega}\right|\right|\right] \tag{5}$$

in the sense that the increase in closeness of the *a posteriori* distributions of the ensemble constituents to the ground truth will always reflect in an increase in accuracy. It is immediate that

$$A_M \propto \sum_{k=1}^{M} (A_k^t). \tag{6}$$

Let us assume that there is some accuracy ranking of the classifier types, so that $A^{t_1} > A^{t_2} > A^{t_3} > \cdots$. Let us denote by n_t the number of ensemble constituents of type t, so that an ensemble is characterized by the vector $\mathbf{n} = [n_t | t \in T^*]$, where T^* denotes the identifiers of the classifiers types ordered by accuracy ranking. Then it is immediate that for two ensembles such that $\mathbf{n}' > \mathbf{n}''$ according the lexicographic ordering, i.e. the classifier with the best ranking is more frequent, then the first ensemble will very likely have accuracy greater than the second. The strategy of AHERF is to estimate via cross-validation on a small dataset the classifier type ranking $A^{\hat{t}_1} > A^{\hat{t}_2} > A^{\hat{t}_3} > \cdots$, using this information to drive the selection of the classifier type of each individual constituent. In order to have ensembles whose characteristic vector $\mathbf{n}$ is of the form $n_{t_1} >> n_{t_2} >> n_{t_3} > \cdots$ we sample an integer random variable whose distribution of probability is an approximation of the exponential distribution built using the Fibbonacci series on the ranking. The anticipatory character of AHERF relies in the prediction of the appropriate distribution of classifier architectures before building and training the ensemble. This stochastic sampling of the classifier type (with replacement) allows for some ensemble design flexibility. When the difference in performance is not so great we would like to have a more even mixture of classifier architectures, and vice-versa, when an architecture is much better adapted we may want to have a majority of classifiers of this type.

6. Experimental design

6.1. Validation methodology

All of the reported experimental results are computed as the average of 50 repetitions of a 10-fold cross-validation approach, where all feature extraction and classification parameters are estimated from the training datasets and applied to the testing datasets as such.

6.2. z-score data normalization

We perform a data normalization by the independent computation of the z-score of each input variable given by the expression $z = \frac{x-\mu}{\sigma}$, where x is the input variable, μ is the variable mean estimation, and σ the variable standard deviation estimation. This normalization removes scale effects reducing all variables to the same order of magnitude, and linear shifts. In cross-validation approaches, the μ and σ are estimated on the training data and used as such on the testing data, resulting in some minor inconsistencies if there is any sampling bias.

6.3. Model parameter selection

The following parameters remain to be specified or selected for each combination of data rotation and ensemble of classifiers. All of them are set in the same way for all the cases, because we want to avoid any effect from them in the experimental results.

- L: The number of individual classifiers is set to $L = 35$ for all experiments.
- Classifier intrinsic parameters: The DT depth is set to 10 in all cases, except for some defaults in SciKit. The number of hidden nodes in the ELM is set to $\min\{N/3, 1000\}$. The SFLN architecture

Table 1
Statistics of ED admissions from 2013 to 2016. Age mean and standard deviation. Remaining rows give the number of records and the percentage relative to the total Columns correspond to no readmission, readmission, and total number of records. By rows, we give the total number and percentage of the occurrence of each kind of gender, class of pathology, and triage assigned upon arrival.

	Number of records	72 h readmission		Total
		No (n = 148,617)	Yes (n = 5674)	n = 154,291
	Age (years)	33.3 (24.8)	22.2 (24.6)	32.9 (24.9)
Gender	Male	69,106 (46.5%)	2832 (49.9%)	71,938 (46.6%)
	Female	79,511 (53.5%)	2842 (50.1%)	82,353 (53.4%)
Pathology	General Medicine	91,566 (61.6%)	2375 (41.9%)	93,941 (60.9%)
	Traumatology	16,651 (11.2%)	325 (5.7%)	16,976 (11%)
	Pediatric	39,999 (26.9%)	2964 (52.2%)	42,963 (27.8%)
	Gynaeco-obstetrics	401 (0.3%)	10 (0.2%)	411 (0.3%)
Triage	I	649 (0.4%)	8 (0.1%)	657 (0.4%)
	II	17,280 (11.6%)	501 (8.8%)	17,781 (11.5%)
	III	111,310 (74.9%)	4309 (75.9%)	115,619 (74.9%)
	IV	19,057 (12.8%)	848 (14.9%)	19,905 (12.9%)
	V	321 (0.2%)	8 (0.1%)	329 (0.2%)

trained by ELM has a single output unit encoding the output of the classifier as an integer value, both for two-class and many-classes datasets.

- K: The number of partitions of the set of features has been set to $K = \lfloor n/4 \rfloor$. As the effective partitions are random, it is very likely that some of them will be composed of only one vector.

6.4. Performance measures

In binary classification, accuracy is defined as the proportion of true results among the total population: Accuracy = (TN + TP)/(TN + TP + FN + FP) where TN are true negatives, TP true positives, FN false negatives, and FP false positives. Sensitivity is the proportion of correctly classified positives: Sensitivity = (TP)/(TP + FN) Specificity is defined as the proportion of negatives that are correctly identified as such: Specificity = (TN)/(TN+FP).

7. Materials

Our raw dataset is composed of ED admission events of 102,534 patients divided into 2 groups, namely adults and pediatrics, which amounts to 156,120 admission cases recorded between January 1st, 2013 and August 31, 2015 from the electronic medical records of the Hospital José Joaquín Aguirre de la Universidad de Chile. At admission time a set of 17 variables were collected. The variables or features are divided into three main groups: (i) Sociodemographic data and baseline status, (ii) Personal history and (iii) Reasons for consultation or diagnoses made at admission. The dataset contains missing values. Table 1 shows the distribution of admissions and readmission records according to gender, broad pathology class (general medicine, traumatology, pediatric, and gynaeco-obtretics), and the assigned triage.

In order to formulate the readmission prediction following a binary classification approach, the target binary variable takes values readmitted/not readmitted. Patients returning to the ED within 72 h after being discharged are considered readmitted, otherwise are considered as new admissions. It is noteworthy that one patient returning the first day and another returning the 72nd hour are both considered as readmitted. On the other hand, a patient returning at the 73rd hour from discharge is considered as not readmitted, while in practice it is very likely that the patient underwent a readmission.

The variables describing each patient include a categorical variable encoding the admission motivation, this encoding into more than 500 topics is given by the electronic medical record implementation. Table 2 contains the more frequent causes of admission and readmission, those accounting for 1.5% of the cases or more. The non informative category "OTHERS" is the most frequent, and the most frequent causes for admission appear also as causes of readmission. In our current implementation, this variable has been encoded with a vector of binary valued features, one per each admission motivation category. Additional features correspond to the encoding of the triage, demographics variables such as age, sex, adult or pediatric patient, and physiological variables such as blood pressure, temperature, heart rate, respiratory rate, glucose levels, and others. Hence, feature vector dimension is greater than 500, which is an already very high dimension.

Table 2
Distribution of admission and readmission cases. GAP general abdominal pain, 1/3DF, up to three days fever; 24HF, 24 h; fever; HA, headache; D, diarrhea; T, throwing up; EP, epigastric pain; LuP, lumbar pain; GD, general discomfort; LegP, leg pain; AD, acute disnea.

Admission		Readmission	
Motive	%	Motive	%
OTHER	14.22	OTHER	30.13
GAP	8.21	GAP	8.20
24HF	5.53	1/3DF	5.40
COUGH	5.47	COUGH	4.28
HA	4.93	24HF	4.10
1/3DF	3.65	HA	3.04
GD	2.33	D	2.59
EP	2.22	T	2.43
T	2.21	EP	1.86
D	2.16	LuP	1.51
LegP	2.11		
LuP	2.06		
AD	1.57		
FP	1.55		
NAUSEA	1.44		

8. Experimental results

In order to build predictive models of patient readmission within 72 h of discharge we have used three different classifiers, namely: AHERF, SVM and Random Forest. All the tests were conducted using 10-fold cross-validation, performing 10 independent executions and averaging the results obtained. To avoid random number generation bias, we have conducted each execution using a different random number generation seed. In order to mitigate the adverse effects caused from malformed data we have removed those instances containing one or more missing values on the features shown in Table 1. Missing values in numerical valued variables (such as glucose level or oxygen saturation) are filled with the

Table 3
Accuracy, sensitivity and specificity results (average ± standard deviation) of the classifiers for the different datasets.

		Acc	Sens.	Spec.
Pediatrics	AHERF	78.57 ± 0.47	70.6 ± 0.34	86.55 ± 0.69
	SVM	59 ± 0.17	54.01 ± 0.36	64.04 ± 0.25
	RF	72.72 ± 0.26	64.93 ± 0.52	80.52 ± 0.52
Adults	AHERF	78.17 ± 0.34	72.57 ± 0.33	83.82 ± 0.54
	SVM	65.54 ± 0.6	54.87 ± 0.54	75.23 ± 0.94
	RF	65.54 ± 0.46	55.52 ± 0.63	75.57 ± 0.43
All	AHERF	78.14 ± 0.33	68.02 ± 0.35	88.26 ± 0.6
	SVM	67.78 ± 0.10	44.46 ± 0.10	89.86 ± 0.10
	RF	71.28 ± 0.26	59.56 ± 0.22	82.37 ± 0.46

arithmetic mean of the variable across the population. The original dataset is very imbalanced, i.e. the target readmission class samples number is much less than a 0.5% of the dataset. As it is well known, imbalance makes accuracy an unreliable performance measure [43]. For instance, a 10-fold cross-validation of the RF classifier upon the entire dataset achieves over 96.2% accuracy, however its average sensitivity is down to 0.4% while specificity reaches 99.8%. The interpretation of these results is that these RF classifiers are guided by the *a priori* class probability distribution. In essence, RF classification is not very different from assigning all data instances the majority class. The goal in this paper is to shown the comparative performance of AHERF, therefore we build balanced datasets for the computational experiments. The majority class is subsampled to the size of the minority class for each repetition of the cross-validation training process. In our experiment we will consider three different datasets, namely: (i) full dataset, (ii) pediatric patients and (iii) adult patients.

Table 3 shows the average accuracy, sensitivity and specificity along with its respective standard deviation, obtained from the cross-validation experiments. In this table it can be appreciated that sensitivity is much higher than in the reference experiment with the raw unbalanced data, approaching the value of specificity for all the classifier training algorithms, due to the balance of the training dataset. Also, it can be appreciated that AHERF reports results that are significantly better than those of those of SVM and RF ($p < 10^{-6}$ in one-sided t-tests using all results of cross-validation folders). Focusing on the sensitivity results, which are more relevant than accuracy and specificity to compare classifier architectures over imbalanced datasets when we are specially concerned by the minority class, we find that AHERF reaches results over or close to 70%, hence it approaches the required performance for real life application. Taking into account that the adult and pediatric populations have quite different statistics, we have performed separate experiments for them, as well as on the entire dataset. It can be appreciated that results on the separate populations are better than on the entire dataset, which confirms that there are specific discriminant features for these subpopulations. Sensitivity is lower in the pediatric than in the adult population, because the class imbalance is greater in the pediatric dataset than in the adults dataset. Most emergency admissions of children are related to traumatic events that once healed do not relapse. Cronic conditions that are a major cause for readmissions, such as respiratory diseases, are less frequent than in the adult population. More precisely, carrying two-sided t-test in the pediatrics population between sensitivity classifier results, we find that AHERF is significantly ($p < 0.00001$) better than SVM and RF, with a performance increase of 22% and 8% respectively. Not surprisingly, RF performance is 15% greater than that of SVM. These differences are bigger if we consider the specificity results measuring success detecting the majority class. If we consider the adult population, we find again that AHERF is significantly better than RF and SVM (two-sided t-test, $p < 0.00001$), with a sensitivity performance increase of 23%, while the difference between RF and SVM is not significative. The greater performance increase from AHERF to RF and SVM in the adults population than in the pediatrics population is due to the greater sensitivity of the RF and SVM classifiers to the class imbalance ratio. If we consider the effect on the AHERF we find that there is an increase in sensitivity of 2% from the pediatrics to the adults population, which is barely significative (t-test, p=0.013). Pulling together pediatrics and adult population, there is a decrease in sensitivity of AHERF of 5% and 3% relative to the adult and pediatrics results, respectively, due to the fact that discriminant variables are different for each population, so that building a monolithic classifier lose predicitive power. The results of AHERF suggest that the approach is promising for a practical implementation of institution specific readmission risk prediction systems.

9. Conclusions and future work

The AHERF is hybrid ensemble classifier including the anticipative selection of the classifier architecture according to the estimation their classification performance on the specific dataset. AHERF builds a selection probability distribution derived from the accuracy ranking computed in model selection cross-validation experiments. There is no circularity effect in the approach, because model selection, training and testing data are strictly disjoint. The positive results in previous publications are confirmed in this paper, where we apply AHERF to data extracted from electronic medical records in order to predict readmission risk in an emergency department. AHERF achieves big improvements over other state of the art approaches. The size of the dataset, which becomes close to real life industrial problems, encourages the application to larger problems. Moreover, the increase of sensitivity performance of the AHERF over the RF and SVM classifiers makes it recommendable for institution specific emergency department readmission risk assessment. Readmission risk prediction is an imbalanced problem, where conventional monolithic classifier approaches fail because they are biased towards the majority class. AHERF shows a better distribution adaptation than conventional state of the art approaches. Performance evaluation of AHERF in other domains has also been positive. This paper adds to this positive evaluation in a specific non-trivial data domain.

Regarding the population, it is important to notice the differences in sensitivity performance achieved if we consider all the date pulled together or if we build specific predictors for each population strata. It seems that the separate treatment can be extended to aggregated of motives of admission, or other demographic variables, building specific predictors for each segment of the population defined by them. This approach of a collection of readmission predictors has the advantage of better adaptation to changing population conditions due to aging or increasing prevalence of some disease.

Future work will be directed to feature selection in order to provide the healthcare providers with specific rules and causes for

the readmission event. On other line of research, the intrinsic massive parallelism of AHERF can be beneficial for its application to big data problems.

Acknowledgment

The Basque Government Grant IT874-13 for the Computational Intelligence research group.

References

[1] D.H. Wolpert, The lack of a priori distinctions between learning algorithms, Neural Comput. 8 (7) (1996) 1341–1390.
[2] D. Wolpert, W. Macready, No free lunch theorems for optimization, IEEE Trans. Evol. Comput. 1 (1) (1997) 67–82.
[3] N.C. Oza, K. Tumer, Classifier ensembles: select real-world applications, Inf. Fusion 9 (1) (2008) 4–20.
[4] M. Wozniak, M. Graña, E. Corchado, A survey of multiple classifier system as hybrid systems, Inf. Fusion 16 (2014) 3–17.
[5] L. Breiman, Bagging predictors, Mach. Learn. 24 (2) (1996) 123–140.
[6] L. Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32.
[7] H.-C. Kim, S. Pang, H.-M. Je, D. Kim, S.-Y. Bang, Constructing support vector machine ensemble, Pattern Recognit. 36 (2003) 2757–2767.
[8] D. Chyzhyk, B. Ayerdi, J. Maiora, Active learning with bootstrapped dendritic classifier applied to medical image segmentation, Pattern Recognit. Lett. 34 (2013) 1602–1608.
[9] J. Rodriguez, L. Kuncheva, C. Alonso, Rotation forest: a new classifier ensemble method, IEEE Trans. Pattern Anal. Mach. Intell. 28 (10) (2006) 1619–1630, http://dx.doi.org/10.1109/TPAMI.2006.211.
[10] B. Ayerdi, M. Graña, Hybrid extreme rotation forest, Neural Netw. 52 (2014) 33–42.
[11] G.-B. Huang, Q.-Y. Zhu, C.-K. Siew, Extreme learning machine: theory and applications, Neurocomputing 70 (1–3) (2006) 489–501.
[12] G.-B. Huang, L. Chen, C.-K. Siew, Universal approximation using incremental constructive feedforward networks with random hidden nodes, IEEE Trans. Neural Netw. 17 (4) (2006) 879–892.
[13] G. Huang, G.-B. Huang, S. Song, K. You, Trends in extreme learning machines: a review, Neural Netw. 61 (2015) 32–48, http://dx.doi.org/10.1016/j.neunet.2014.10.001.
[14] G.-B. Huang, H. Zhou, X. Ding, R. Zhang, Extreme learning machine for regression and multiclass classification, IEEE Trans. Syst. Man Cybern. Part B: Cybern. 42 (2) (2012) 513–529.
[15] B. Ayerdi, J. Maiora, A. d'Anjou, M. Graña, Applications of hybrid extreme rotation forests for image segmentation, Int. J. Hybrid Intell. Syst. 11 (1) (2014) 13–24.
[16] B. Ayerdi, M. Grana, Anticipative hybrid extreme rotation forest, Procedia Comput. Sci. 80 (2016) 1671–1681.
[17] B. Ayerdi, M. Graña Romay, Hyperspectral image analysis by spectral-spatial processing and anticipative hybrid extreme rotation forest classification, IEEE Trans. Geosci. Remote Sens. 54 (5) (2016) 2627–2639, http://dx.doi.org/10.1109/TGRS.2015.2503886.
[18] A. Besga, B. Ayerdi, G. Alcalde, A. Manzano, P. Lopetegui, M.G. Na, A. Gonzalez-Pinto, Risk factors for emergency department short time readmission in stratified population, BioMed Res. Int. 2015 (2015) 685067.
[19] S. Hao, Y. Wang, B. Jin, A.Y. Shin, C. Zhu, M. Huang, L. Zheng, J. Luo, Z. Hu, C. Fu, D. Dai, Y. Wang, D.S. Culver, S.T. Alfreds, T. Rogow, F. Stearns, K.G. Sylvester, E. Widen, X.B. Ling, Development, validation and deployment of a real time 30 day hospital readmission risk assessment tool in the Maine healthcare information exchange, PLoS ONE 10 (10) (2015) 1–15, http://dx.doi.org/10.1371/journal.pone.0140271.
[20] S. Yu, F. Farooq, A. van Esbroeck, G. Fung, V. Anand, B. Krishnapuram, Predicting readmission risk with institution-specific prediction models, Artif. Intell. Med. 65 (2) (2015) 89–96, http://dx.doi.org/10.1016/j.artmed.2015.08.005.
[21] H.Q. Nguyen, L. Chu, I.-L. Amy Liu, J.S. Lee, D. Suh, B. Korotzer, G. Yuen, S. Desai, K.J. Coleman, A.H. Xiang, M.K. Gould, Associations between physical activity and 30-day readmission risk in chronic obstructive pulmonary disease, Ann. ATS 11 (5) (2014) 695–705, http://dx.doi.org/10.1513/AnnalsATS.201401-017OC.
[22] L. Pereira, C. Choquet, A. Perozziello, M. Wargon, G. Juillien, L. Colosi, R. Hellmann, M. Ranaivoson, E. Casalino, Unscheduled-return-visits after an emergency department (ED) attendance and clinical link between both visits in patients aged 75 years and over: a prospective observational study, PLOS ONE 10 (4) (2015) e0123803.
[23] C.H. Olson, S. Dey, V. Kumar, K.A. Monsen, B.L. Westra, Clustering of elderly patient subgroups to identify medication-related readmission risks, Int. J. Med. Inform. 85 (1) (2016) 43–52, http://dx.doi.org/10.1016/j.ijmedinf.2015.10.004.
[24] D. Kansagara, H. Englander, A. Salanitro, D. Kagen, C. Theobald, M. Freeman, S. Kripalani, Risk prediction models for hospital readmission: a systematic review, JAM 306 (15) (2011) 1688–1698, http://dx.doi.org/10.1001/jama.2011.1515.
[25] A.H. Zai, J.G. Ronquillo, R. Nieves, H.C. Chueh, J.C. Kvedar, K. Jethwani, Assessing hospital readmission risk factors in heart failure patients enrolled in a telemonitoring program, Int. J. Telemed. Appl. 2013 (2013), http://dx.doi.org/10.1155/2013/305819, 1:1–1:1.
[26] M.D. Silverstein, H. Qin, S.Q. Mercer, J. Fong, Z. Haydar, Risk factors for 30-day hospital readmission in patients >65 years of age, in: Baylor University Medical Center. Proceedings, vol. 21, 2008, p. 363.
[27] C. Carpenter, K. Heard, S. Wilber, A. Ginde, K. Stiffler, L. Gerson, et al., Research priorities for high-quality geriatric emergency care: medication management, screening, and prevention and functional assessment, Acad. Emerg. Med. 18 (6) (2011) 644–645.
[28] M. Deschodt, E. Devriendt, M. Sabbe, D. Knockaert, P. Deboutte, S. Boonen, J. Flamaing, K. Milisen, Characteristics of older adults admitted to the emergency department (ED) and their risk factors for ed readmission based on comprehensive geriatric assessment: a prospective cohort study, BMC Geriatr. 15 (1) (2015) 1.
[29] C. van Walraven, I.A. Dhalla, C. Bell, E. Etchells, I.G. Stiell, K. Zarnke, P.C. Austin, A.J. Forster, Derivation and validation of an index to predict early death or unplanned readmission after discharge from hospital to the community, Can. Med. Assoc. J. 182 (6) (2010) 551–557.
[30] C. Van Walraven, J. Wong, A. Forster, Lace+ index: extension of a validated index to predict early death or urgent readmission after hospital discharge using administrative data, Open Med. 6 (2012) 80–89.
[31] C. van Walraven, F.A. McAlister, J.A. Bakal, S. Hawken, J. Donzé, External validation of the hospital-patient one-year mortality risk (HOMR) model for predicting death within 1 year after hospital admission, Can. Med. Assoc. J. 187 (10) (2015) 725–733.
[32] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: machine learning in Python, J. Mach. Learn. Res. 12 (2011) 2825–2830.
[33] L. Breiman, J. Friedman, R. Olshen, C. Stone, Classification and Regression Trees, Wadsworth and Brooks, Monterey, CA, 1984.
[34] J.R. Quinlan, C4.5: Programs for Machine Learning, Morgan Kaufmann, 1993.
[35] G.B. Huang, D.H. Wang, Y. Lan, Extreme learning machines: a survey, Int. J. Mach. Learn. Cybern. 2 (2011) 107–122.
[36] B. Ayerdi, I. Marques, M. Graña, Spatially regularized semisupervised ensembles of extreme learning machines for hyperspectral image segmentation, Neurocomputing 149 (Part A) (2015) 373–386.
[37] B. Ayerdi, M. Graña, Hyperspectral image nonlinear unmixing and reconstruction by ELM regression ensembles, Neurocomputing 174 (2016) 299–309.
[38] D. Chyzhyk, A. Savio, M. Graña, Computer aided diagnosis of schizophrenia on resting state FMRI data by ensembles of {ELM}, Neural Netw. 68 (2015) 23–33, http://dx.doi.org/10.1016/j.neunet.2015.04.002.
[39] C. Burges, A tutorial on support vector machines for pattern recognition, Data Min. Knowl. Discov. 2 (2) (1998) 121–167 http://citeseer.ist.psu.edu/525500.html.
[40] V. Vapnik, Statistical Learning Theory, Wiley-Interscience, 1998.
[41] R. Schapire, Y. Singer, Improved boosting algorithms using confidence-rated predictions, Mach. Learn. 37 (3) (1999) 297–336, http://dx.doi.org/10.1023/A:1007614523901.
[42] Y. Freund, R. Schapire, A decision-theoretic generalization of on-line learning and an application to boosting, in: European Conference on Computational Learning Theory, 1995, pp. 37, 23. http://citeseer.ist.psu.edu/89601.html.
[43] V. López, A. Fernández, S. García, V. Palade, F. Herrera, An insight into classification with imbalanced data: empirical results and current trends on using data intrinsic characteristics, Inf. Sci. 250 (2013) 113–141.

Arkaitz Artetxe studied Computer Engineering in the University of the Basque Country (UPV/EHU) in San Sebastian, 2011. Between 2009 and 2010 he studied in the University of Aberdeen, Scotland (UK). Since 2011 he works in Vicomtech-IK4 (http://www.vicomtech.org) as a researcher in the area of Biomedical Applications. In 2014 he received the M.Sc. in Computational Engineering and Intelligent Systems from the University of the Basque Country (UPV/EHU). Currently, he is a doctoral student advised by Prof. M. Graña. His research activities focus on the application of knowledge engineering and data mining to the medical domain.

Borja Ayerdi received the M.Sc. from the Basque Country University (UPV/EHU) in Donostia, Spain, 2011, in computer science. Currently, he is a doctoral student advised by Prof. M. Graña. His current interests cover different areas of hyperspectral image processing with bioinspired tools, such as Extreme Learning Machines. He is author of over 10 papers in international journals.

Manuel Graña received the M.Sc. and Ph.D. degrees from Universidad del Pais Vasco (UPV/EHU), Donostia, Spain, in 1982 and 1989, respectively, both in computer science. His current position is a Full Professor (Catedrático de Universidad) with the Computer Science and Artificial Intelligence Department of the Universidad del Pais Vasco (UPV/EHU). He is the head of the Computational Intelligence Group (Grupo de Inteligencia Computational). His current research interests are in applications of computational intelligence to linked multicomponent robotic systems, medical image in the neurosciences, multimodal human computer interaction, remote sensing image processing, content based image retrieval, lattice computing, semantic modelling, data processing, classification, and data mining.

Sebastian Rios is Assistant Professor at the Industrial Engineering Department of the University of Chile since (2008). He received the B.E on Industrial Engineering on 2001, the B.E on Computer Science, P.E. on Industrial Engineering on 2003 from the University of Chile, Chile; and the Ph.D. on Knowledge Engineering from the University of Tokyo, Japan on 2007. He is the Founder and Director of the Business Intelligence Research Center (CEINE) at the University of Chile since 2012, a collaborative applied research effort between private companies and the University. His research interests include data mining algorithms in big dataset and its applications to different industry domains (medicine, marketing, management, etc.); he also is interested in generative topic models for text mining in social networks and knowledge representation using semantic web technologies.