Neurocomputing 000 (2018) 1–13

Contents lists available at ScienceDirect

Neurocomputing

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

# Pruning strategies for nearest neighbor competence preservation learners☆

Fabrizio Angiulli[a,*], Estela Narvaez[b]

[a] DIMES Department University of Calabria, 87036 Rende, CS, Italy
[b] Faculty of Enginnering, National University of Chimborazo, 060150 Riobamba, Ecuador

ARTICLE INFO

*Article history:*
Received 9 August 2017
Revised 9 December 2017
Accepted 22 April 2018
Available online xxx

Communicated by Dr Xiaofeng Zhu

*Keywords:*
Classification
Nearest neighbor rule
Training-set consistent subset
Overfitting
Pessimistic error estimate

ABSTRACT

In order to alleviate both the spatial and temporal cost of the nearest neighbor classification rule, competence preservation techniques aim at substituting the training set with a selected subset, known as consistent subset. In order to improve generalization and to prevent induction of overly complex models, in this study the application of the Pessimistic Error Estimate (PEE) principle in the context of the nearest neighbor rule is investigated. Generalization is estimated as a trade-off between training set accuracy and model complexity. As major results, it is shown that PEE-like selection strategies guarantee to preserve the accuracy of the consistent subset with a far larger reduction factor and, moreover, that sensible generalization improvements can be obtained by using a reduced subset. Moreover, comparison with state-of-the-art hybrid prototype selection methods highlight that the here introduced FCNN-PAC strategy is able to obtain a model of size comparable to that obtained by the best prototype selection methods, with far smaller time requirements, corresponding to four orders of magnitude on medium-sized datasets.

© 2018 Elsevier B.V. All rights reserved.

## 1. Introduction

The nearest neighbor decision rule [1] (nearest neighbor rule for short) assigns to an unclassified sample point the classification of the nearest of a set of previously classified points. A strong point of the nearest neighbor rule is that, for all distributions, its probability of error is bounded above by twice the Bayes probability of error [1–3]. That is, it may be said that half the classification information in an infinite size sample set is contained in the nearest neighbor.

Naive implementation of the nearest neighbor rule requires to store all the previously classified data points, and then to compare each sample point to be classified to each stored point.

In order to reduce both space and time requirements the concept of *training set consistent subset*, that is a subset of the original training set that correctly classifies all the training samples, was introduced by Hart [4] together with an algorithm, called the CNN rule (for Condensed nearest neighbor rule), to determine a consistent subset of the original sample set. Since then different techniques have been introduced [5–8], referred to as training set reduction, training set condensation, reference set thinning, and prototype selection algorithms.

Using a training set consistent subset, instead of the entire training set, to implement the nearest neighbor rule has the additional advantage that it may guarantee better classification accuracy. Indeed, [6] showed that the VC dimension of an nearest neighbor classifier is given by the number of reference points in the training set.

Thus, in order to achieve a classification rule with controlled generalization, it is better to replace the training set with a small consistent subset.

In this study, motivated by approaches used in the context of other classification algorithms in order to improve generalization and to prevent induction of overly complex models, such as in the case of decision trees, we investigate the application of the Pessimistic Error Estimate (PEE) principle [9] in the context of the nearest neighbor rule.

With this aim, we relax the notion of consistency of a subset by introducing the notion of $\alpha$-consistent subset ($\alpha \in [0, 1]$), that is a subset that correctly classifies at least the $\alpha$ fraction of the training set. Then we describe a variant of the FCNN algorithm [8], called $\alpha$-FCNN rule, that computes an $\alpha$-consistent subset.

☆ A preliminary version of this work appears under the title "Pruning Nearest Neighbor Competence Preservation Learners" in the proceedings of the 27th IEEE International Conference on Tools with Artificial Intelligence, ICTAI 2015, Vietri sul Mare, Italy, November 9–11, 2015, pp. 943–949.

* Corresponding author.
*E-mail addresses:* fabrizio.angiulli@unical.it (F. Angiulli), e.narvaez@dimes.unical.it (E. Narvaez).

https://doi.org/10.1016/j.neucom.2018.04.017
0925-2312/© 2018 Elsevier B.V. All rights reserved.

Please cite this article as: F. Angiulli, E. Narvaez, Pruning strategies for nearest neighbor competence preservation learners, Neurocomputing (2018), https://doi.org/10.1016/j.neucom.2018.04.017

We then introduce some subset selection strategies, namely PACOPT, MAXOPT, and TRNOPT, intended to select the most promising subset according to different ways of estimating expected accuracy. Among them, the PACOPT strategy is based on PEE principle and estimates generalization as a trade-off between training set accuracy and model complexity.

Moreover, a variant of the PACOPT strategy, called aPACOPT for approximate-PACOPT, is described. The technique attempts to reduce time complexity by early terminating the learning phase on the basis of the current trend of the pessimistic accuracy estimate curve.

Before going into the details, next we summarize the contributions of this research:

- As the first major result, we show that the PACOPT selection strategy guarantees to preserve the accuracy of the consistent subset with a larger reduction factor, since on the average the subset selected by PACOPT contains the 30% of the training set consistent subset objects.
- As the second major result, we show that a sensible, on the average of the 2%, generalization improvement can be obtained by using a reduced subset of intermediate size, consisting on the average of the 63% of the training set consistent subset objects.
- Moreover, the aPACOPT strategy allows to compute approximatively the same model determined by the PACOPT strategy, but with sensibly smaller time requirements (the execution time of aPACOPT corresponds to the 25% – 50% of the time required by PACOPT).
- Comparison with state-of-the-art prototype selection methods highlight that the aPACOPT strategy is able to obtain a model of size comparable to that obtained by the best prototype selection methods in terms of reduction ratio, with far smaller time requirements, corresponding to 4 orders of magnitude on medium-sized datasets.

The rest of the work is organized as follows. Section 2 discusses scenario and related works. Section 3 describes the $\alpha$-FCNN algorithm for computing a soft consistent training set subset. Section 4 describes the four selection strategies for reducing training set complexity, namely TRNOPT, PACOPT, MAXOPT, and aPACOPT. Section 5 illustrates experimental results, including accuracy, scalability, and comparison with state-of-the-art related methods. Section 6 concludes the work.

## 2. Related work

Errors committed by classification models [10] are generally divided into two types: *training errors* and *generalization errors*. The training error, also known as resubstitution error, is the number of misclassification errors committed on training records, whereas the generalization error is the expected error of the model on previously unseen records. A good model must have low training error as well as low generalization error. This is important because a model that fits the training data too well can have a poorer generalization error than a model with a higher training error, this phenomenon is known as model *overfitting*. The model overfitting problem has been investigated by several authors including [11–13].

The chance for model overfitting increases as the model becomes more complex. For this reason, might be preferred simpler models, a strategy that agrees with a well-known principle known as Occam's razor: given two models with the same generalization error, the simpler model is preferred over the more complex model [9].

Thus, overfitting happens when a model is more flexible than it needs to be and incorporates noise in the training data to the extent that it negatively impacts the performance on the model on new data [14]. There are several possible causes why overfitting happens: the presence of noise, a model too complex, a small training set, a very rich hypothesis space, a domain with many features [9].

A way to help learning algorithms to select the most appropriate model, is to be able to estimate the generalization error. The right complexity is that of a model that produces the lowest generalization error. The problem is that the learning algorithm has access only to the training set during model building. It has no knowledge of the test set, and thus, does not know how well the model will perform on records it has never seen.

Decision trees are one of the most common classification techniques used in the practice. A decision tree is a hierarchical structure consisting of nodes and directed edges. A problem common when a decision tree is built is that many of the branches will reflect anomalies in the training data due to noise or outliers. As the number of nodes in the decision tree increases, the tree will have fewer training errors. Up to a certain size also the test error will decrease. However, once the tree becomes too large, its test error rate begins to increase even though its training error rate continues to decrease, due to overfitting.

Two well-known techniques for incorporating model complexity into the evaluation of classification models are PEE and MDL, which are described next in the context of decision trees.

The *Pessimistic error estimate* (PEE) approach explicitly computes generalization error as the sum of training error and a penalty term for model complexity. The resulting generalization error can be considered its pessimistic error estimate. Let $n(t)$ be the number of training records classified by node $t$ and $e(t)$ be the number of misclassified records. The pessimistic error estimate of a decision tree $T$, $e_g(T)$, can be computed as follows:

$$e_g(T) = \frac{\sum_{i=1}^{k}[e(t_i) + \Omega(t_i)]}{\sum_{i=1}^{k} n(t_i)} = \frac{e(T) + \Omega(t)}{N(t)},$$

where $k$ is the number of leaf nodes, $e(T)$ is the overall training error of the decision tree, $N_t$ is the number of training records, and $\Omega(t_i)$ is the penalty term associated with each node $t_i$.

The *Minimum Description Length* (MDL) principle is a general method for inductive inference, based on the idea that the more we are able to compress a set of data, the more regularities we have found in it [15]. Is based on an information-theoretic approach, when two models fit the data equally well, MDL will choose the one that is the simplest in the sense that it allows for a shorter description of the data [9]. The MDL principle involves adding to the error function an extra term that is designed to penalize mappings that are not smooth [16,17]. MDL principle use encoding techniques to define the best decision tree as the one that requires the fewest number of bits to both (1) encode the tree and (2) encode the exceptions to the tree.

There are two common approaches to tree pruning that exploit the above mentioned principles, that are *pre-pruning* and *post-pruning* [18–25]. The first approach stops growing the tree earlier, before it perfectly classifies the training set. Upon halting, the node becomes a leaf. The leaf may hold the most frequent class among the subset tuples or the probability distribution of those tuples. The second approach, post-pruning, removes subtrees from a fully grown tree. A subtree at a given node is pruned by removing its branches and replacing it with a leaf. The leaf is labeled with the most frequent class among the subtree being replaced.

Reducing model complexity is transversal to many other learning scenarios, as in the case of ensemble pruning [26], instance selection for time-series prediction [27], resampling for improving classifier performance [28], and unsupervised feature selection [29].

Please cite this article as: F. Angiulli, E. Narvaez, Pruning strategies for nearest neighbor competence preservation learners, Neurocomputing (2018), https://doi.org/10.1016/j.neucom.2018.04.017

Download English Version:

https://daneshyari.com/en/article/6863691

Download Persian Version:

https://daneshyari.com/article/6863691

Daneshyari.com