Neurocomputing 73 (2010) 1272–1280

Contents lists available at ScienceDirect

Neurocomputing

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

AUC maximization linear classifier based on active learning and its application

Guang Han *, Chunxia Zhao

College of Computer Science and Technology, Nanjing University of Science and Technology, Nanjing 210094, China

ARTICLE INFO

Article history:
Received 24 June 2009
Received in revised form
28 December 2009
Accepted 1 January 2010
Communicated by T. Heskes
Available online 14 January 2010

Keywords:
Obstacle detection
Active learning
AUC maximization
Linear classifier
Dynamic clustering
Gradient descent method

ABSTRACT

Aiming at labeling and ranking difficulties caused by a large number of samples, as well as uneven distribution of samples in outdoor obstacle detection of the autonomous mobile robot, an AUC maximization linear classifier method based on active learning is proposed in this paper. This method firstly uses dynamic clustering algorithm to select the representative samples and labels these samples, then these labeled samples are put in the training set. Next, a linear classifier is trained using the AUC maximization method on the training set. The above process will be repeated until the AUC converges. The experiments are performed on real outdoor environment image database. The experiment results show that the very good detection results are obtained using the method proposed in this paper with only 120 samples. More importantly, using the proposed method can significantly reduce the workload of labeling the samples and size of the sample set, and AUC maximization proposed also excels the existing methods.

1. Introduction

Receiver operating characteristic (ROC) curve is a useful performance evaluation tool. This is because the produced curve is independent of class distribution and underlying misclassification costs [32]. In order to measure the quality of different classification algorithm directly, a common method is proposed—that is to calculate the area under the ROC curve (area under the ROC, abbreviated as AUC). In [24] a detailed statistical analysis of the relationship between the AUC and the error rate has been given, while the studies in [1,28] suggested that AUC is a more effective classifier performance evaluation method than the classification error rate. The majority of the available classification systems focus on the minimization of the classification error rate. This is not always suitable, especially when two-class problems with skewed classes and cost distributions are solved [30,28,31]. Recently, several new techniques of training the classifier had been developed, which directly optimized the AUC to train the classifier. Rank optimizing support vector machines (SVM) have come into focus. In [26,33,34,27] the linear programming and quadratic programming algorithm for maximizing the AUC under the framework of the SVM have been proposed. In [2] rank optimizing kernels have been investigated and [3] introduces a similar kernel formulation, which leads to a better ranking performance compared to the previous work. In [4] a linear weighting of features has been successfully applied to the detection of the interstitial lung disease. In [35] methods for optimizing the AUC value locally have been developed in the context of decision trees while in [13,14,36] some algorithms also have been proposed to obtain approximations of the global AUC value, but in general these algorithms did not obtain AUC values significantly better than those obtained by an algorithm designed to minimize the error rate. The combinations of rules have been analyzed to maximize AUC of the combiner directly and a method to evaluate the weight of the linear combination of two or more classifiers has been proposed in [25,5] while in [6] a nonparametric linear classifier based on the maximization of AUC has been proposed, which lies on an iterative pairwise coupling of the features for the optimization of the ranking of the combined feature. Also, a method of estimating the AUC with censored data has been proposed in [29].

The main problem of the above algorithms is that computing AUC is a costly operation: it requires sorting the database, a cost of order $n\log(n)$ for a database of size n. Therefore, if the size n is very great, the corresponding cost of computation is also very great. One method of solving this problem is to use polynomial approximations in the AUC optimization, such as the method in [7]. The polynomial approximation has the advantage that it can be computed in only one scan over the database, and hence it does

* Corresponding author. Tel.: +86 25 6855 8427.
E-mail addresses: hanguang8848@163.com (G. Han), zhaochx@mail.njust.edu.cn (C. Zhao).

0925-2312/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.neucom.2010.01.001

not require resorting the database every time the AUC for a new or updated classification function is needed. Furthermore, when the classification function is only slightly changed, it is even possible to find the new AUC without a database scan based on a small summary of the database. Another method is to reduce the size of sample set as far as possible without losing most of the information in the sample set, if the size of the sample set becomes small, even if computation complexity of the algorithm is very high, the total computation amount is also acceptable. Now the main point of this method is how to effectively reduce the size of the sample set without losing the information. Therefore, active learning algorithm is introduced in this paper and an AUC maximization linear classifier based on active learning is proposed. Active Learning is primarily used to solve a large number of the unlabeled samples; it can automatically select which samples should be labeled so that using some few samples can achieve good learning results [9]. More properties about active learning can be found in Section 2.1. The proposed method firstly uses dynamic clustering to select the representative samples and labels these samples, and then these labeled samples are put in the training set (this is active learning algorithm used in this paper). Next, a linear classifier is trained using a new AUC maximization method on the training set. The above process will be repeated until the AUC converges. Another advantage of using the proposed method (which is also the purpose of active learning algorithm) is that the sample set need not be labeled in advance. This is particularly important for the sample set of containing a large number of samples, because it can significantly reduce the workload of manual labeling.

Autonomous mobile robot navigation in outdoor, off-road environments requires solving complex classification problems. Obstacle detection is one of the tasks that has been successfully approached using supervised machine learning techniques for classification. Obstacle detection can be seen as the problem of using sensory data to classify regions around a robot as traversable or not [8]. Large amounts of training data are usually necessary in order to obtain satisfactory generalization and are easily achieved for obstacle detection. However, manually labeling these samples and sorting their output of classifier is an expensive and tedious task. In addition, the obstacle/non-obstacle samples are serious uneven in distribution, and the sample size of the former is often far less than the latter. Aiming at the characters of the obstacle samples, the method proposed in this paper is appropriate and effective. The experiments are performed on outdoor scene image database and the results show that the use of the proposed method, first of all, can greatly reduce the workload of labeling the samples; secondly, the size of sample set can be reduced as far as possible without losing most of the information in the sample set; finally, the AUC maximization method proposed can achieve better detection results than other existing methods.

An AUC maximization linear classifier method based on active learning is proposed in this paper. The originality of our method is mainly on the following four points:

1. AUC maximization linear classifier method based on active learning. The training of the available classifiers is mostly based on the minimization of the classification error rate. The classifier is trained based on effective fusion of active learning and AUC maximization in this paper.
2. Active learning with the association of three conditions. If redundancy between samples is very high, that is to say, similar samples are too many near the decision boundary. General active learning method cannot quickly find the representative and informative samples. To solve this problem, we propose an active learning based on dynamic clustering, and for the samples that are selected, the following three conditions should be met: (1) The distance between samples and decision boundary is small; (2) the similarity degree among the samples is high; (3) the similarity degree between samples and samples having been labeled is low.
3. The improvement of AUC maximization method. When the AUC is optimized by using gradient descent algorithm, a large number of training samples as well as the high dimension feature vector easily makes algorithm a local optimum [14,5]. Therefore, this paper proposes the following method to solve this problem. For the new indicator function, the gradient descent method of continuous function and the combining tree technology are used in combination. This method not only reflects the ordinal comparison between the classification outputs of two sets of features but also reflects the amount of difference of the classification output of two sets of features.
4. Our method is firstly introduced to outdoor obstacle detection of the autonomous mobile robot.

2. Preliminary material

2.1. Active learning

Active learning is an area of machine learning that addresses the situation in which a large amount of unlabeled data is available, and only a limited subset of it can be labeled. Active learning can automatically select which samples should be labeled so that using some few samples can achieve good learning results. Most of the active learning methods are based on certain criteria to select the samples and these samples are labeled by the experts. Then all the labeled samples are used to train the classifier. The above process is repeated until the algorithm meets some conditions [9]. The representative algorithms on active learning have Seung and Freund's QBC (Query-By-Committee) algorithm [48,49], Abe and Mamitsukat's QBBAG (Query-by-Bagging) algorithm [44] and Query-by-Boosting algorithm [50], Lewis and Gale's Uncertainty Sampling algorithm [43], Muslea and Minton's Co-Testing algorithm [45], Tong and Koller's active learning algorithm based on support vector machines [9], etc. Active learning algorithm also has been applied successfully to text and web page classification, image retrieval and other fields [43–46,9–11]. Therefore, we believe that active learning can solve the problems of labeling the samples and reducing the size of samples set for the autonomous mobile robot. The general active learning methods cannot be applied to obstacle detection directly [47]; they need to be improved according to the characters of obstacle detection. Aiming at the problem of similar samples near the decision boundary being too many, this paper uses a dynamic clustering method to find the representative samples.

2.2. ROC curve

The ROC curve depicts the performance of a classifier by plotting the true-positive rate against the false-positive rate. Assuming that a classifier produces a continuous output (e.g., class posterior probabilities), then the output must be thresholded to label each sample as positive or negative. Thus, for each setting of the decision threshold, a true-positive rate and a false-positive rate are obtained. By varying the decision threshold over a range from 0 to 1, the ROC curve is produced, as shown in Fig. 1. More properties about ROC curve can be found in [12].