Available online at www.sciencedirect.com
ScienceDirect
Information Sciences 178 (2008) 2621–2638
INFORMATION SCIENCES
AN INTERNATIONAL JOURNAL
www.elsevier.com/locate/ins

Risk-sensitive loss functions for sparse multi-category classification problems

S. Suresh *, N. Sundararajan, P. Saratchandran

School of Electrical and Electronics Engineering, Nanyang Technological University, 50 Nanyang Avenue, Singapore 638768, Singapore

Received 14 July 2007; received in revised form 1 February 2008; accepted 6 February 2008

Abstract

In this paper, we propose two risk-sensitive loss functions to solve the multi-category classification problems where the number of training samples is small and/or there is a high imbalance in the number of samples per class. Such problems are common in the bio-informatics/medical diagnosis areas. The most commonly used loss functions in the literature do not perform well in these problems as they minimize only the approximation error and neglect the estimation error due to imbalance in the training set. The proposed risk-sensitive loss functions minimize both the approximation and estimation error. We present an error analysis for the risk-sensitive loss functions along with other well known loss functions. Using a neural architecture, classifiers incorporating these risk-sensitive loss functions have been developed and their performance evaluated for two real world multi-class classification problems, viz., a satellite image classification problem and a micro-array gene expression based cancer classification problem. To study the effectiveness of the proposed loss functions, we have deliberately imbalanced the training samples in the satellite image problem and compared the performance of our neural classifiers with those developed using other well-known loss functions. The results indicate the superior performance of the neural classifier using the proposed loss functions both in terms of the overall and per class classification accuracy. Performance comparisons have also been carried out on a number of benchmark problems where the data is normal i.e., not sparse or imbalanced. Results indicate similar or better performance of the proposed loss functions compared to the well-known loss functions.
Published by Elsevier Inc.

Keywords: Multi-category classification; Neural network; Risk-sensitive loss function; Cross-entropy; Satellite imaging; Micro-array gene expression

1. Introduction

In classification problems, the goal is often to correctly identify the class label for an observed input pattern. For any classifier, its performance is dependent on the chosen loss function. For a given problem, selection of proper loss function $V(x)$ is often difficult [8,29]. Different classification techniques in machine learning

* Corresponding author. Tel.: +65 8157 4822.
E-mail address: sundaram.suresh@hotmail.com (S. Suresh).

0020-0255/$ - see front matter Published by Elsevier Inc.
doi:10.1016/j.ins.2008.02.009

employ different loss functions to get better classification accuracy. For example, classifiers developed using adaptive boosting algorithms (AdaBoost) employ exponential ($\exp^{-V(x)}$) loss function [5,9], support vector machines employ loss functions of the form $(1 - V(x), 0)$ [8,29] and neural network classifiers employ mean square error minimization [7,22] or cross-entropy [23]. Recently, much attention has been paid to solve multi-category classification problems using neural networks [7,13,22,23,25] using different loss functions.

It is well known, [7,13,22,23,25] that the classifiers using mean square error loss function produce an output that approximates the desired posterior probability. It has also been shown [7,13,22,23,25] that other loss functions such as modified mean square error, cross-entropy also approximate the posterior probability. The conditions on the above loss functions such that the outputs of the neural classifiers closely approximate the posterior probability are discussed in [13,25]. However, in most of these works only binary classification problems have been considered. The central issue in machine learning is the extension of the results presented for the binary classification problem to multi-category classification problems.

Usually, the multi-category classification problem is first converted into multiple two class problem (one-versus-all or one-versus-one) and solved independently [1,3,12,20]. The one-versus-all method is then combined with a winner-take-all strategy to get the class number for unknown pattern. Similarly, the one-versus-one method is combined with max-wins or pairwise coupling strategy [12]. In these approaches, the posterior probabilities for each class in a multi-category problem are estimated using the binary classification problem for each category. In recent years, researchers have proposed an "all-together" approach to solve multi-category classification problem in one step by considering all the classes together [15,31]. In [14], incremental class learning approach is presented. Here, the classifiers learn classes one by one. However, such approaches require more training time and training samples per class. Recently in [7,31], the posterior probability estimation results for binary classification problems are extended to multi-category classification problem. A complete overview on loss functions and their behavior on decision performance has been presented in [6].

In general, the performance of neural classifiers is measured by their expected misclassification instances. To achieve better performance, the classifier employs loss functions that minimize the misclassification for all classes. Here, the loss functions assign equal weightage to the samples in every class. For many real world classification problems, minimization of misclassification alone is not adequate. For example, in sensitive applications such as medical diagnosis, false alarm detection in radar systems, financial and critical control applications, misclassification of samples in different classes have different consequences. It is often important to measure the confidence level in the class label prediction and corresponding risk associated with the action behind every classifier decision. This risk associated with the decision should be taken into account for better performance of the classifier. One way to handle this problem is to integrate the risk factor in the loss function.

Another important issue in classification problems is the availability of small number of samples and a high imbalance in the samples per class (also known as sampling bias). The small number of samples and a high imbalance in class will incur additional difficulty to know the distribution completely. This can result in an increased error in classification. Also, the overlap between the classes adds an additional complexity to the problem and may cause in the deterioration of the classifier performance.

Detailed explanation of the above issues are provided in [11]. In [16], a new approach based on statistical learning theory with a fixed cost of misclassification is presented by assuming the existence of hyperplanes that separates the data completely. In [18,4], support vector machine based classifiers are developed using unequal cost of misclassification to handle the above issues. In all these approaches, the cost of misclassification is fixed a prior. Fixing the cost of misclassification for a given problem is often difficult. Since, the classifier formulation includes the cost of misclassification, selection of these values influences its performance significantly. For multi-category classification problems with strong overlap between classes and high imbalance in samples per class, fixing the cost a priori is nearly impossible.

In order to overcome these difficulties in multi-category classification problems, in this paper, we introduce two new risk-sensitive loss functions. In the proposed loss functions, a risk factor is integrated with the existing cross-entropy and modified hinge loss functions. During the learning process, the risk factor is adapted using the cost of misclassification and approximate posterior probability of appropriate class of the given training samples. The estimated risk factor is used further in the learning process to reduce the error in classification. Here, the risk factors are estimated such that they increase with an increase in misclassification. For example,