Available online at www.sciencedirect.com

ScienceDirect

Pattern Recognition Letters 29 (2008) 1072–1081

Pattern Recognition Letters

www.elsevier.com/locate/patrec

On the use of different loss functions in statistical pattern recognition applied to machine translation ☆

J. Andrés-Ferrer [b,,1], D. Ortiz-Martínez [b,1], I. García-Varea [a], F. Casacuberta [b]

[a] *Departamento de Sistemas Informáticos, Universidad de Castilla-La Mancha, Spain*

[b] *Departament de Sistemes Informàtics i Computació, Universitat Politècnica de València, Spain*

Available online 14 July 2007

Abstract

In pattern recognition, an elegant and powerful way to deal with classification problems is based on the minimisation of the classification risk. The risk function is defined in terms of loss functions that measure the penalty for wrong decisions. However, in practice a trivial loss function is usually adopted (the so-called 0–1 loss function) that do no make the most of this framework. This work is focused on the study of different loss functions, and specially on those loss functions that do not depend on the class proposed by the system. Loss functions of this kind have allowed us to theoretically explain heuristics that are successfully used with very complex pattern recognition problem, such as (statistical) machine translation. A comparative experimental work has also been carried out to compare different proposals of loss functions in the practical scenario of machine translation.

Keywords: Loss function; Classification rules; Bayes' risk; Decision theory; Statistical machine translation; Direct translation rule

1. Introduction

Statistical pattern recognition is a well-founded discipline that solved many practical classification problems. A classification problem is stated as the problem of choosing which class a given object belongs to. Let $\mathcal{X}$ be the domain of the objects that a classification system might observe; and $\mathcal{Y}$ the set of classes ($\{y_1, y_2, \ldots, y_C\}$). Then, each classification system is characterised by a function that maps each object to one class, the so-called *classification function* ($c : \mathcal{X} \rightarrow \mathcal{Y}$) (Duda et al., 2000).

The performance of a classification function is usually measured as a function of the classification error. However, there are problems in which all the classification errors do not have the same repercussions. Therefore, a function that ranks these mistakes should be provided. The *loss function*, $\mathrm{l}(y|\mathbf{x}, y')$, evaluates the *loss* that is incurred by the classification function when classifying the object $\mathbf{x}$ in to the class y, knowing that the correct class is y' (Duda et al., 2000).

Taking this framework into account, *the global risk* (Duda et al., 2000) that characterises the contribution of all objects in the performance of classifiers is formally defined as follows:

$$R(c) = E_x[R(c(\mathbf{x})|\mathbf{x})] = \int_{\mathcal{X}} R(c(\mathbf{x})|\mathbf{x}) \cdot p(\mathbf{x}) \cdot \mathrm{d}\mathbf{x} \tag{1}$$

where $R(\mathrm{c}(\mathbf{x})|\mathbf{x})$ is the *conditional risk given* $\mathbf{x}$, i.e. the expected loss of classifying $\mathbf{x}$ in the class determined by the

☆ Work partially supported by the "Agència Valenciana de Ciència i Tecnologia" under the Spanish project ITEFTE (TIC2003-08681-C02-02) and by the CICYT "Centro de Investigación Científica y Técnologíca" under the Spanish Project iDoc (TIC2003-08681-C02).

Corresponding author. Tel.: +34 963877350x83533; fax: +34 96 387 73 59.

E-mail addresses: jandres@dsic.upv.es (J. Andrés-Ferrer), dortiz@iti.upv.es (D. Ortiz-Martínez), ivarea@info-ab.uclm.es (I. García-Varea), fcn@dsic.upv.es (F. Casacuberta).

[1] Work supported by the "Conselleria d'Empresa Universitat i Ciència" of the "Generalitat Valenciana" under the Grants CTBPRA/2005/004 and CTBPRB/2003/324.

0167-8655/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.patrec.2007.06.015

decision function c. This conditional risk is expressed as follows:

$$R(y|\boldsymbol{x}) = \sum_{y' \in \mathscr{Y}} l(y|\boldsymbol{x}, y') \cdot p(y'|\boldsymbol{x}) \tag{2}$$

Minimising the conditional risk for each object $\boldsymbol{x}$ is a sufficient condition to minimise the global risk. Without loss of a generality, we can say that the optimal classification rule, namely *minimum Bayes' risk*, is the one that minimises the conditional risk, i.e.:

$$\hat{c}(\boldsymbol{x}) = \arg\min_{y \in \mathscr{Y}} R(y|\boldsymbol{x}) \tag{3}$$

In practice, however, calculating the global risk when comparing systems, requires the classification of all possible objects. Therefore, an *empirical risk* on a test set T is used then:

$$\overline{R}_T(c) = \frac{1}{|T|} \sum_{x \in T} R(c(\boldsymbol{x})|\boldsymbol{x}) \tag{4}$$

A common approach is to consider that each misclassification has the same importance according to the *0–1 loss function* which distinguishes two sorts of actions: wrong classification (loss of 1) and correctly classification (zero loss):

$$l(y|\boldsymbol{x}, y') = \begin{cases} 0 & y = y' \\ 1 & \text{otherwise} \end{cases} \tag{5}$$

The goal of a classification system is to minimise the global risk, which is understood to be the classification error rate since Eq. (5) is used. When Eq. (5) is used, the minimum Bayes' risk in Eq. (3) can be simplified providing the well-known optimal Bayes' classification rule (Duda et al., 2000):

$$c(\boldsymbol{x}) = \arg\max_{y \in \mathscr{Y}} p(y|\boldsymbol{x}) \tag{6}$$

where $\boldsymbol{x}$ is the object to be classified, and y denotes one class from $\mathscr{Y}$. Depending on which loss function the system design is based on, there exists a different optimal classification rule.

However, while the 0–1 loss function is adequate for many problems with a small set of classes, there are problems where a more appropriate loss function should be defined. For example, if the system classifies diseases, it may be worse to classify an ill person as a healthy one than vice-versa. Another important example is the case in which the set of classes is large, or even infinite (but still enumerable). In such cases, as the set of all possible classes is huge, it is not appropriate to penalise all wrong classes with the same weight. In other words, since it is impossible to define a uniform distribution when the number of classes is infinite, it does not make sense to define a uniform loss function in the infinite domain because there are objects that are more probable than others, and the error will be increased if the system fails in probable objects. Instead of using the 0–1 loss function, it would be better to penalise the domain zones where the probability is high. This way, the system will avoid mistakes on probable objects at the expense of making mistakes on unlikely objects, and the error will be decreased since unlikely objects occur fewer times in comparison with probable objects. Note that we are dealing with infinite enumerable sets in this example, and, therefore, this is a classification problem and not a linear regression problem.

This work focuses on the sort of loss functions that can improve system performance while keeping the simplicity of 0–1 optimal classification rule. In (Schlüter et al., 2005) complex classification rules were analysed using *metric loss function*. There are other works that analyse general loss functions, for instance (Ueffing and Ney, 2004). We focus on loss functions of other type which are not restricted by the metric requirements at the expense of ignoring the class proposed by the system.

In this work, we propose and analyse different loss functions which are eligible for substituting the 0–1 loss function in pattern recognition problems. This substitution is specially appealing when the set of classes is infinite. In order to empirically analyse the proposed loss functions, a real scenario has been used as a case study: the pattern recognition approach to machine translation, more broadly known as *statistical machine translation* (SMT) (Brown et al., 1993).

The remainder of the paper is organised as follows. In Section 2, we introduce the case study. In Section 3 different loss functions are analysed from the viewpoint of Bayes' decision theory. In Section 4, phrase-based statistical translation models, which are broadly used in current state-of-art SMT systems, are described in detail. Section 5 is devoted to SMT experiments in which different loss functions were used; the different corpora used in the translation experiments are briefly described, as well. Finally, concluding remarks are summarised in Section 6.

2. A case study: Statistical machine translation

Until now, our discussion has been focused on general classification problems. A case of a classification problem in which the set of classes is infinite is statistical machine translation (SMT). In the purely statistical approach, the MT is tackled as a classification problem where the set of classes is the set of all sentences of the target language ($\boldsymbol{\mathscr{Y}}^*$), i.e. every target string ($\boldsymbol{y} \in \boldsymbol{\mathscr{Y}}^*$) is regarded as a possible translation for a source language string ($\boldsymbol{x} \in \mathscr{X}^*$). Hence, the system searches the target string ($\hat{\boldsymbol{y}}$) with maximum a-posteriori probability $p(\boldsymbol{y}|\boldsymbol{x})$:

$$\hat{\boldsymbol{y}} = \arg\max_{y \in \mathscr{Y}^*} \{p(\boldsymbol{y}|\boldsymbol{x})\} \tag{7}$$

where $p(\boldsymbol{y}|\boldsymbol{x})$ can be approached through a direct[2] statistical translation model (Brown et al., 1990).

[2] We will refer to $p(\boldsymbol{y}|\boldsymbol{x})$ as a direct statistical translation model and to $p(\boldsymbol{x}|\boldsymbol{y})$ as an inverse statistical translation model.