Information Processing and Management 45 (2009) 631–642

Contents lists available at ScienceDirect

Information Processing and Management

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

A discrete mixture-based kernel for SVMs: Application to spam and image categorization

Nizar Bouguila *, Ola Amayri

Concordia Institute for Information Systems Engineering, Faculty of Engineering and Computer Science, Concordia University, Montreal, Qc, Canada H3G 2W1

ARTICLE INFO

Article history:
Received 14 May 2008
Received in revised form 1 December 2008
Accepted 7 May 2009
Available online 5 June 2009

Keywords:
SVM
Kernels
Multinomial dirichlet
Finite mixture models
Maximum likelihood
EM
CEMM
Deterministic annealing
MDL
Spam
Image database

ABSTRACT

In this paper, we investigate the problem of training support vector machines (SVMs) on count data. Multinomial Dirichlet mixture models allow us to model efficiently count data. On the other hand, SVMs permit good discrimination. We propose, then, a hybrid model that appropriately combines their advantages. Finite mixture models are introduced, as an SVM kernel, to incorporate prior knowledge about the nature of data involved in the problem at hand. For the learning of our mixture model, we propose a deterministic annealing component-wise EM algorithm mixed with a minimum description length type criterion. In the context of this model, we compare different kernels. Through some applications involving spam and image database categorization, we find that our data-driven kernel performs better.

1. Introduction

The technological developments of the last few decades have increased the volume of information (text, images and videos) on the Internet and Intranets. Different approaches have been proposed to manage, filter and retrieve these information. Two main categories of approaches are: model-based approaches and discriminative classifiers. Model-based approaches are based on generative probabilistic models and discriminative classifiers allow the construction of flexible decision boundaries. SVM is a well-known example of discriminative classifiers (Boser, Guyon, & Vapnik, 1992; Vapnik, 1999). As a theoretically rich method and because of their advantages such as their use of over-fitting protection independently from the number of features and their effectiveness in the case of sparse data, SVMs have been widely used in many applications. Finite mixture models, on the other hand, provide a principled and effective way for clustering (McLachlan & Peel, 2000). The majority of the work done with both techniques has focused on continuous data. This paper, however, concerns the modeling and classification of count data which are an important component in many applications and information management tasks (Bouguila, 2008; Bouguila & Ziou, 2007). To reach this goal, we use both mixture models and SVMs approaches in a way that combines their respective advantages. Indeed, combining model-based and discriminative approaches has been shown to be effective in different applications (Jaakkola & Haussler, 1999). We propose, then, a mixture model-based kernel for SVMs to

* Corresponding author.
E-mail addresses: bouguila@ciise.concordia.ca (N. Bouguila), o_amayri@encs.concordia.ca (O. Amayri).

0306-4573/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ipm.2009.05.005

classify count data. The proposed mixture model-based kernel assume that the data follow a multinomial Dirichlet mixture (MDM) distribution (Bouguila & Ziou, 2007) letting the data tell us as much as possible about its structure. The MDM model is learned using a modified deterministic annealing expectation maximization (DAEM) algorithm and a minimum description length (MDL) type criterion. The advantages of using a data-driven kernel, instead of "blindly" choosing classic kernels, are shown through some experiments involving spam filtering and image database categorization.

The organization of the remainder of this paper is as follows: In Section 2, we propose our data-driven kernel. Section 3, outlines the proposed mixture model estimation and selection. Experimental results are presented in Section 4. Finally, we conclude the paper in Section 5.

2. A mixture model-based kernel for count data

2.1. SVMs

SVM is basically a learning machine for two-group[1] classification problems (Boser et al., 1992). Assume that we have a data set of N V-dimensional vectors $\mathscr{X} = (\vec{X}_1, \ldots, \vec{X}_N)$ with labels $y_i \in \{-1, 1\}$ belonging to either of two linearly separable classes C_1 and C_2. Let each $\vec{X}_i = (X_{i1}, \ldots, X_{iV}),\ i = 1, \ldots, N$ be a vector representing a document or an image i, and let X_{iv}, $v = 1, \ldots, V$, the number of times word or visual feature v appears in the document or in the image, respectively. v ranges from 1 to the vocabulary (or the visual corpus in the case of image) size V. Using these training vectors, the SVM algorithm (see Boser et al. (1992) and Vapnik (1999), for instance, for more details) determines the parameters of a decision function $f(\vec{X})$ during a learning phase ($\vec{X} \in C_1$ if $f(\vec{X}) > 0$, otherwise, $\vec{X} \in C_2$). In the case of non-separable data, however, the vectors are mapped to another high-dimensional feature space where a linear decision surface is constructed. Indeed, the classification of an unknown vector $\vec{X}$ is done by taking the sign of the following function:

$$f(\vec{X}) = \sum_{i=1}^{n} y_i \lambda_i \Phi^{tr}(\vec{X}) \Phi(\vec{X}_i) + b = \sum_{i=1}^{n} y_i \lambda_i K(\vec{X}, \vec{X}_i) + b \tag{1}$$

where n is the number of support vectors containing the relevant information about the classification problem, $\{\lambda_i\}$ are the weights of the support vectors determined by solving a constrained quadratic programming problem which aims to maximize the margin between the classes (Vapnik, 1999), b is a bias term, $\Phi(\vec{X})$ is a nonlinear vector function that maps the V-dimensional input vector $\vec{X}$ into another V'-dimensional feature space ($\Phi(\vec{X}) = (\phi_1(\vec{X}), \ldots, \phi_{V'}(\vec{X}))$ and $K(\vec{X}, \vec{X}_i) = \Phi^{tr}(\vec{X})\Phi(\vec{X}_i)$), where tr denotes the transpose, is a symmetric positive definite kernel function.

A challenging problem in the case of SVMs is the choice of the kernel function which is actually a measure of similarity between two vectors. Different choices of kernel functions have been proposed and extensively used in the past and the most popular are the gaussian RBF, polynomial of a given degree, and multi layer perceptron (Vapnik, 1999). These kernels are in general used, independently of the problem, for both discrete and continuous data. In the case of count data, for instance, these kernels have been applied especially for text categorization (Dumais, 1998; Dumais, Platt, Heckerman, & Sahami, 1998; Joachims, 1999; Joachims, 1998). However, in most of the applications, the intrinsic structure of the data has been ignored by these standard kernels. Moreover, it was shown that the kernel function should be generated directly from data which gives better results (Jaakkola & Haussler, 1999). One of the most successful approaches is the Fisher kernel proposed in Jaakkola and Haussler (1999). Thus, we will focus on Fisher kernels derived from generative probability models. An excellent choice as a generative model for count data is the multinomial Dirichlet mixture (MDM) (Bouguila & Ziou, 2007) that we will discuss in the following section.

2.2. The MDM: a generative model for count data

Using mixture models, each vector of counts $\vec{X}$ can be assumed to be generated by a linear combination of component density functions resulting in the following:

$$p(\vec{X}|\Theta) = \sum_{j=1}^{M} p_j p(\vec{X}|\xi_j) \tag{2}$$

where M is the number of components in the model, $\Theta = \{\theta_j = (\xi_j, p_j)\}$, denotes the set of parameters for the overall model, $\{p_j\}$ are the mixing parameters and represent the weights of each cluster j, $0 \leqslant p_j \leqslant 1$, $\sum_{j=1}^{M} p_j = 1$ and $p(\vec{X}|\xi_j)$ is the component density with parameters ξ_j. An important problem in the case of finite mixture models is the choice of the component density. A common choice to model count data is the multinomial distribution. However, recent developments have shown that this choice is inappropriate in many cases and applications (see Bouguila & Ziou (2007) and Elkan (2006), for instance, for

[1] Although designed for binary classification, SVMs can be used for multiclass problems (see Hsu & Lin (2002) where different methods are proposed and compared).