Available online at www.sciencedirect.com

ScienceDirect

Applied Soft Computing 7 (2007) 908–914

Applied Soft Computing

www.elsevier.com/locate/asoc

Text classification: A least square support vector machine approach

Vikramjit Mitra [a], Chia-Jiu Wang [b,*], Satarupa Banerjee [c]

[a] ECE Department, University of Maryland, College Park, MD, United States
[b] ECE Department, University of Colorado Colorado Springs, CO, United States
[c] CS Department Villanova University Villanova, PA, United States

Received 12 April 2004; received in revised form 1 March 2006; accepted 2 April 2006
Available online 5 June 2006

Abstract

This paper presents a least square support vector machine (LS-SVM) that performs text classification of noisy document titles according to different predetermined categories. The system's potential is demonstrated with a corpus of 91,229 words from University of Denver's Penrose Library catalogue. The classification accuracy of the proposed LS-SVM based system is found to be over 99.9%. The final classifier is an LS-SVM array with Gaussian radial basis function (GRBF) kernel, which uses the coefficients generated by the latent semantic indexing algorithm for classification of the text titles. These coefficients are also used to generate the confidence factors for the inference engine that present the final decision of the entire classifier. The system is also compared with a K-nearest neighbor (KNN) and Naïve Bayes (NB) classifier and the comparison clearly claims that the proposed LS-SVM based architecture outperforms the KNN and NB based system. The comparison between the conventional linear SVM based classifiers and neural network based classifying agents shows that the LS-SVM with LSI based classifying agents improves text categorization performance significantly and holds a lot of potential for developing robust learning based agents for text classification.

© 2006 Elsevier B.V. All rights reserved.

Keywords: Least square support vector machines; Latent semantic indexing; Text classification; Kernel based learning algorithms

1. Introduction

Rapid advancement in technology has motivated text documents to be available in electronic form. The World Wide Web itself contains a huge amount of documents, conference materials, publications, journals, editorials, news and information etc., available in electronic form. These materials along with others result in enormous amount of easily available information, which lack organization. The lack of organization of materials in the World Wide Web necessitates a growing interest in assisting people to manage the huge amount of information. Organized search, browsing, information routing, filtering, objectionable material identification, junk mail, topic identification etc., are the central issues in current information management efforts. This requires implementation of sophisticated learning agents that are capable of classifying relevant information and hence increases text organization. Previous research in the field of Internet agents has used manual or simple encoding techniques [1], linear SVMs and neural network [2] based intelligent agents for information retrieval.

Text classification (TC) is a text content-based classification technique that assigns texts to some predefined categories [3–5,19,27]. The key issues in TC are feature encoding and classifier design, tuning and implementation. Feature extraction is a method of document encoding; that automatically construct internal representations of documents. This paper aims to organize the materials available in a library, which has both electronic materials as well as physical documents. A library cataloging system usually stores the entire information regarding a material, which results in higher storage space requirement.

This paper presents an analysis of learning agents based on support vector machines (SVM). In particular least square support vector machine (LS-SVM) [13] with latent semantic indexing (LSI) for feature extraction will be explored. Furthermore the details of the internal structure of the classifying agent along with the results obtained from our research are presented. The results obtained from the LS-SVM based classifier will be compared with K-Nearest Neighbor (KNN) and Naïve Bayesian (NB) based classifier, which claims the potential and pertinence of using LS-SVMs as the intelligent classifying and search agents for semantic text classification.

* Corresponding author. Tel.: +1 719 262 3495; fax: +1 719 262 3589.
E-mail addresses: vmitra@umd.edu (V. Mitra), cwang@eas.uccs.edu (C.-J. Wang), satarupa.banerjee@villanova.edu (S. Banerjee).

1568-4946/$ – see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.asoc.2006.04.002

2. Latent semantic indexing

Latent semantic indexing commonly known as LSI [3–6] is a text classification and document indexing technique that generates a vector model of semantics based upon word co-occurrences. Using the estimate of the most significant statistical factors in the weighted word space, LSI [5,6] extracts the underlying semantic structure of a word corpus. LSI considers documents that have many words in common to be semantically close and those with few words in common to be semantically distant. This method emulates human knowledge to classify and categorize a document collection based on its content. LSI search agents look at similarity values it has calculated for every content word, and returns the documents that it thinks best fit the query [4]. This makes LSI successful where a plain keyword search will fail if there is no exact match. At the initial stage, LSI preprocesses the text corpus by purging all the extraneous words from a document [4,5], leaving only content words that have some semantic meaning. This way it eliminates those words that introduces noise to the decision making task.

LSI algorithm generates a matrix representation of the corpus, with rows corresponding to words in the vocabulary and columns to the documents. Each value in this matrix is a weighted frequency of the corresponding term in the corresponding document, which reduces the influence of frequently occurring term [7]. The matrix thus generated is large sparse one, which is then reduced to a compressed matrix based on singular value decomposition (SVD) technique, given in (1):

$$R = UPV \tag{1}$$

where the matrix R is decomposed into a matrix of reduced rank U, a diagonal matrix of singular values P and a document matrix V. The row vector of matrix U and the column vector of matrix V are the projections of word vectors and document vectors into singular value space.

3. Support vector machines (SVM)

The general form of support vector machine is used to separate two classes by a function, which is induced from available examples [8,15]. The main goal of this classifier is to find an optimal separating hyperplane that maximizes the separation margin or the distance between it and the nearest data point of each class. For a set of training vectors belonging to two separate classes, shown in (2):

$$\{(x^1, y^1), \ldots, (x^m, y^m)\}, \qquad x \in R^n, y \in \{1, -1\} \tag{2}$$

A hyperplane as shown in Eq. (3) can be found to separate these two classes:

$$\langle w, x \rangle + b = 0 \tag{3}$$

The above set of vectors is said to be optimally separated by the hyperplane if it is separated without error and the distance between the closest vector to the hyper plane is maximal. Vapnik [15,8] introduced a canonical hyperplane, where the parameters w, b are constrained by Eq. (4):

$$\min_i |\langle w, x^i \rangle + b| = 1 \tag{4}$$

From the above set of equations, it can be derived that the optimal separating hyperplane given by:

$$w^* = \sum_{i=1}^{l} \alpha_i y_i x_i \tag{5}$$

$$b^* = -0.5\langle w^*, x_r + x_s \rangle \tag{6}$$

where α_i is the Lagrange multiplier. x_r and x_s are any support vectors from each class satisfying $\alpha_r > 0$, $y_r = -1$; $\alpha_s > 0$, $y_s = 1$.

3.1. Kernel functions

In the case where a linear boundary is inappropriate the SVM can map the input vector, x, into a high dimensional feature space, z [8]. By selecting the non-linear mapping as a priori, the SVM constructs an optimal separating hyperplane in this higher dimensional space. This idea exploits the method of Aizerman et al. (1964), which enables the curse of dimensionality (Bellman, 1961) to be addressed. The idea of kernel function is to enable operations to be performed in the input space rather than the potentially high dimensional feature space. Due to this the inner product does not need to be evaluated in the feature space, which provides a way of addressing the curse of dimensionality.

The theory of Reproducing Kernel Hilbert Space (RKHS) (Wahba, 1990; Aronszajn, 1950; Girosi, 1997; Heckman, 1997), claims that an inner product in feature space has an equivalent Kernel in input space:

$$K(x, x') = \langle \phi(x), \phi(x') \rangle \tag{7}$$

provided certain conditions hold. The Gaussian radial basis function (GRBF) has received significant attention and its form is given by:

$$K(x, x') = \mathrm{e}^{-\|x-x'\|^2/2\sigma^2} \tag{8}$$

Classical techniques utilizing RBFs employ some method of determining a subset of centers; typically a method of clustering is employed to select a subset of centers [8]. The most attractive feature of SVM is its implicit selection process, with each support vectors contributing on local Gaussian functions, centered at that data point. By further consideration it is possible to select the global basis function width using the SRM principle (Vapnik, 1995).

3.2. Least square-SVM

SVM is a powerful technique for solving problems in non-linear classification, function estimation and density estimation, which had led to many recent developments in kernel based learning methods [9,11,12,16,15]. Least square support vector machines (LS-SVM) are reformulations to

Download English Version:

https://daneshyari.com/en/article/497094

Download Persian Version:

https://daneshyari.com/article/497094

Daneshyari.com