Expert Systems with Applications 41 (2014) 861–868

Contents lists available at ScienceDirect

Expert Systems with Applications

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

Minimizer of the Reconstruction Error for multi-class document categorization

Juan Carlos Gomez *, Marie-Francine Moens

Department of Computer Science, KU Leuven, Celestijnenlaan 200A, Heverlee, B-3001, Belgium

ARTICLE INFO

Keywords:
Document categorization
Text mining
Dimensionality reduction
Principal Component Analysis

ABSTRACT

In the present article we introduce and validate an approach for single-label multi-class document categorization based on text content features. The introduced approach uses the statistical property of Principal Component Analysis, which minimizes the reconstruction error of the training documents used to compute a low-rank category transformation matrix. Such matrix transforms the original set of training documents from a given category to a new low-rank space and then optimally reconstructs them to the original space with a minimum reconstruction error. The proposed method, called Minimizer of the Reconstruction Error (mRE) classifier, uses this property, and extends and applies it to new unseen test documents. Several experiments on four multi-class datasets for text categorization are conducted in order to test the stable and generally better performance of the proposed approach in comparison with other popular classification methods.

1. Introduction

The current information era allows users to generate large amounts of digital information. A good proportion of this information is in the form of text documents. Text is still the main way of communication among humans. Such text documents are daily exchanged among individuals, companies, organizations, etc. In this direction, it is generally recognized that document categorization plays an important role in the flow of document interchanges, since it facilitates the tasks of accessing and retrieving relevant information by users and systems. Document categorization is a key component for many practical applications such as digital library management, opinion analysis, and Web search engines. Nevertheless, document categorization is very difficult because of the high dimensionality of document representations to be classified and their content diversity.

In the current work we tackle the problem of single-label multi-class text categorization. This problem is defined as: given a training set of documents $\mathbf{T} = [\mathbf{T}_1, \mathbf{T}_2, \ldots, \mathbf{T}_c]$, organized in c categories, where $\mathbf{T}_i = [\mathbf{t}_{i,1}, \mathbf{t}_{i,2}, \ldots, \mathbf{t}_{i,m}]$ corresponds to a term-document matrix composed by m documents in the i-th category, and given a test document $\mathbf{y}$; the goal is then to train a model over the set $\mathbf{T}$ which is able to assign *one* of the c categories to the new test document $\mathbf{y}$.

In the single-label multi-class problem the categories are mutually exclusive, that means each document can belong to only one category. This assumption could not occur in some real scenarios, where some documents could belong to more than one category (multi-label problem) at the same time. Nevertheless, a method for single-label multi-class categorization could be generalized and could be transformed into a set of independent binary categorization problems (Sebastiani, 2002), which is in fact the approach of many systems for multi-label categorization (Tsoumakas, Katakis, & Vlahavas, 2010).

In this work, we present a novel text document classifier which relies on and extends the statistical properties of Principal Component Analysis (PCA). In essence the proposed model uses the framework which derives PCA from the minimization of reconstruction error of the training examples. The model then uses such property to classify new unseen documents, using the idea that new documents are better reconstructed by a transformation matrix which was computed from similar documents. We call this method the Minimizer of Reconstruction Error (mRE) classifier. During training, the mRE classifier computes a set of category transformation matrices $\mathbf{W}_i$; $i = 1, 2, \ldots, c$ of rank r by means of PCA. Inside the classifier, such rank r could be learned using a standard k-fold validation over the training set. During testing, given a new unseen document, the model projects such document using each one of the different category matrices, then reconstructs the document using again each one of the matrices and finally computes the reconstruction errors, by measuring the Frobenius norm of the difference between the set of reconstructed documents and the original one. The matrix which produces the minimum error indicates the category to be assigned to the new document.

* Corresponding author.
E-mail addresses: juancarlos.gomezcarranza@cs.kuleuven.be (J.C. Gomez), sien.moens@cs.kuleuven.be (M.-F. Moens).

0957-4174/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.eswa.2013.08.016

In order to test the validity of the mRE classifier, we perform experiments with several public datasets for text categorization: the Classic dataset, the 20Newsgroups dataset, and the WIPO-alpha (World Intellectual Property Organization) and WIPO-de datasets. We test the model using standard training/test splits of the data. The results show that the mRE classifier gives good and stable results across the different datasets and experiments. In the same direction, with the purpose of having a better overview of the performance of the mRE classifier, we present a comparison for every experiment with three other well known categorization methods: Multinomial Naive Bayes (NB), K-Nearest Neighbors (K-NN) and a linear Support Vector Machine (SVM), which have a very good behavior in text document categorization.

The contributions of our work are the following:

- The feasibility of applying the statistical property of minimizing the reconstruction error with PCA in a single-label multi-class text categorization task, exploiting the sparseness of the category matrices in order to perform a fast training of the model.
- The empirical evidence that mRE is able to properly model the categories of the documents by extracting transformation matrices which represent most of the information from the data in terms of variance and minimization of the reconstruction error of the training documents.
- The evidence that the property of minimizing the reconstruction error could be extended and applied to new unseen documents, where new documents similar to the ones used to compute a given matrix $\mathbf{W}_p$ are better reconstructed by such matrix, allowing in this way to assign the proper category for each new document.
- The evidence that a suitable rank r for the projection matrices $\mathbf{W}_i$ could be learned from the training data using a standard k-fold validation.

The rest of this paper is organized as follows: Section 2 is devoted to recall several statistical and probabilistic works in the field of dimensionality reduction for text document categorization related to mRE. In Section 3 we first give a brief introduction to PCA, from which we derive the mRE model, and secondly we present and describe the general architecture for the mRE classifier. Section 4 illustrates the experimental evaluation framework for the mRE classifier, describing the datasets and setup used during experimentation, and we present the results obtained from the experiments together with a discussion about them. Finally, in Section 5 we present the conclusions and future work.

2. Related research

Since the mRE classifier relies on PCA, it is directly related with several works which employs PCA or similar techniques that transform the original data to a new space, where the variables in this new space are linear combinations of the original features. Using this transformation, it is expected that, under given conditions, the new variables describe in a better way the original data. In the areas of natural language processing, text mining and information retrieval, one of the most known technique is Latent Semantic Analysis (LSA) (also known as Latent Semantic Indexing) (Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990). LSA uses the term-document matrix of a collection of documents and finds a low-rank approximation of such matrix using Singular Value Decomposition (SVD), producing a set of concepts related to the documents and terms. In matrix terms, LSA is similar to PCA, but without centering the term-document matrix. In line with this approach, most of the works devoted to text categorization where PCA or LSA are used, employ PCA as a first step to project the original term-document matrix to a new low-rank space, considering only the first few components with the highest variance. After this initial dimensionality reduction phase the categorization is performed using standard classification algorithms (e.g., SVM, NB, K-NN, etc.) (Gomez, Boiy, & Moens, 2012; Kim, Howland, & Park, 2005; Li & Jain, 1998; Schütze, Hull, & Pedersen, 1995; Sebastiani, 2002; Weigend, Wiener, & Pedersen, 1999). Contrary to these basic approaches, inside the mRE classifier, we do not use the original training documents projected to a low-rank space to train a model, rather we use the transformation matrices computed from such original training documents as the classification model. The computed transformation matrices are able to optimally compress and reconstruct the original documents used to create the corresponding matrix, and are useful to classify new unseen documents.

Linear Discriminant Analysis (LDA) is a categorization/dimensionality reduction technique (Fisher, 1936), which uses the category information to project the data into a new space where the ratio of between-class-variance to within-class-variance is maximized in order to obtain adequate category separability. LDA could be used as a dimensionality reduction technique similar to PCA, but including the category information to improve the separation between classes in the new space (Anderson, 2003); and could be used as well to perform categorization. LDA categorizes a new unseen document by projecting it into the new space and then its projection is compared with the mean of each projected training category. Torkkola (2001) was one of the first authors to use LDA for text categorization purposes. There, the author mentions that PCA aims at an optimal representation of the data but that it does not help for an optimal discrimination of the data, and then proposes LDA to categorize text documents. Nevertheless LDA as classifier tends to perform worse than a SVM for text classification (Kim et al., 2005). In the present work we actually exploit the discriminative properties of PCA for text categorization, by including the category information in form of a transformation matrix per category, which minimize the reconstruction error of the training documents inside the corresponding category.

Non-Negative Matrix Factorization (NMF) is another dimensionality reduction technique. NMF, similar to PCA, projects the data to a new space, but the values in the transformation matrices obtained with NMF are only positive (Barman, Igbal, & Lee, 2006; Berry, Gillis, & Glineaur, 2009). Similarly, other models such as probabilistic Latent Semantic Analysis (pLSA) (Hoffmann, 2007) and Latent Dirichlet Allocation (Blei, Ng, & Jordan, 2003) are probabilistic extensions of the LSA model, and are currently popular as topic representation models. These models reduce the dimensions of the documents by representing them as a mixture of topic distributions and topics as a mixture of words distributions. These models have the disadvantage that identifying the correct number of latent components is a difficult and computationally expensive problem (Blei et al., 2003). In the mRE classifier, we could estimate the proper number of latent components (the rank r of the matrix $\mathbf{W}$), by applying a standard k-fold cross validation over the training set.

As stated before, in this work we propose the mRE classifier, which relies on the framework derived from the property of PCA to minimize the reconstruction error of the training documents used to compute the matrix $\mathbf{W}$ and extend it to apply such property to unseen documents. The same property of PCA has already been used in the computer vision tasks of object detection (Malagón-Borja & Fuentes, 2009) and novelty detection (Hoffmann, 2007), and in a spam filtering task (Gomez & Moens, 2012). However, to the best of our knowledge there is no work devoted to minimize the reconstruction error on single-label multi-class text categorization, where the best rank r for the matrix $\mathbf{W}$ is learned from the training data.