http://ansinet.com/itj

ITJ

ISSN 1812-5638

INFORMATION TECHNOLOGY JOURNAL

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Information Technology Journal 11 (3): 319-323, 2012
ISSN 1812-5638 / DOI: 10.3923/itj.2012.319.323
© 2012 Asian Network for Scientific Information

Kernel Sparse Feature Selection Based on Semantics in Text Classification

Zhantao Deng, Guyu Hu, Zhisong Pan and Yanyan Zhang
Institute of Command Automation, PLA University of Science and Technology,
Nanjing 210007, China

Abstract: Sparse representation originating from signal compressed sensing theory has attracted increasing interest in computer vision research community. In this paper, we present a novel non-parametric feature selection method based on sparse representation in text classification. In order to solve the problem of polysems and synonyms in VSM, we construct semantic structure to represent document with PLSA. Motivated by the fact that kernel trick can capture the nonlinear similarity of features, which may reduce the feature quantization error, we propose Empirical Kernel Sparse Representation (EKSR). We apply EKSR to reconstruct weight vector between samples, then design evaluating mechanism CKernel Sparsity Score (KSS) to select excellent feature subset. As the natural discriminative power of EKSR, KSS can find Agood@ feature which preserves the original structure with less information loss. The results of experiment both on English and Chinese dataset demonstrate the effectiveness of the proposed method.

Key words: Text classification, semantic structure, feature selection, kernel sparse representation

INTRODUCTION

With the rapid development of Internet applications, there will be a huge number of content-rich news release page and forum posts every day. However, the data in network is usually provided in high-dimensional form, which brings on the so-called "curse of dimensionality". Feature selection is an effective approach to deal with such problem, because it preserves original structure and selects excellent feature subsets according as evaluate mechanism. Researchers have developed a variety of feature selection methods such as Document Frequency (DF), Information Gain (IG), Mutual Information (MI), Expected Cross Entropy (ECE) (CHI) and Term Strength (TS).

Studies (Wright *et al*., 2009) have shown that no feature selection method has excellent classification effect in any datasets. Therefore, we focus on designing generalized feature selection method which selects characters of a global and makes the best use of various features of the information to reflect the inner structure of the data reasonably. In this paper, motivated by the recent development of Sparse Representation (SR) (Huang and Aviyente, 2007) which succeed in statistics and pattern recognition (Mairal *et al*., 2008a, b). We propose a simple feature selection method called kernel sparsity score based on semantic (Se-KSS). Specifically, semantic structure of documents is firstly constructed based on probabilistic latent semantic analysis (PLSA) (Hofmann, 1999). Secondly, an "Aadjacent" weight matrix of data set is constructed based on modified sparse representation framework Cempirical kernel Sparse Representation (EKSR), and then the low-dimensional embedding of the data is evaluated to best preserve such weight matrix. At last, evaluating mechanism kernel sparsity score (KSS) is designed to select excellent feature subset. Although supervised information is not needed, KSS tends to find the "good" feature which preserves the original structure.

DOCUMENT REPRESENTATION

Probabilistic Latent Semantic Analysis (PLSA) simulates the occurrence process of word in document by the probabilistic model, and shows the relationship with "document-semantic-word". Suppose we have given a collection of text documents $D = \{d_1, ..., d_M\}$ with terms from a vocabulary $W = (\omega_1, ..., \omega_M)$, $z \in Z = \{z_1, ..., z_K\}$ is an unobserved class variable. Then a joint probability model over $D \times W$ is defined by the mixture:

$$p(d_i, \omega_j) = p(d_i)p(\omega_j \mid d_i);$$
$$p(\omega_j \mid d_i) = \sum_{k=1}^{K} p(\omega_j \mid z_k)p(z_k \mid d_i) \quad (1)$$

Corresponding Author: Zhantao Deng, Institute of Command Automation, PLA University of Science and Technology, Nanjing 210007, China

According as the principle of maximum likelihood estimation, the objective function of PLSA is described as Log-Likelihood function as follows:

$$L(P(D)) = \log(\prod_{d\in D} P(d)) = \log(\prod_{d\in D}\prod_{\omega\in W} P(w,d)^{n(\omega,d)}) = \sum_{d\in D}\sum_{\omega\in W} n(\omega,d)\log P(\omega,d) \quad (2)$$

The standard procedure for maximum likelihood estimation in latent variable models is the Expectation Maximization (EM) algorithm (Hofmann, 2001).Then, K-dimensional vector is considered as term probability representation of document (Sivic *et al*., 2005).

KERNEL SPARSE FEATURE SELECTION

Motivated by kernel trick, nonlinear data is mapped into kernel space in which the nonlinear similarity of the features can be captured and reconstruct by sparse representation.

Empirical kernel sparse representation: Sparse representation is initially proposed as an extension to traditional signal representation (Mallat and Zhifeng, 1993). Suppose we have given a signal $\chi \in R^m$ and a matrix $X = [x1, x2,...., x3] \in R^{m\times n}$ containing the elements of an over-complete dictionary (Murray and Kreutz-Delgado, 2007.) in its columns, the goal of SR is to represent x using as few entries of X as possible. In many practical problems, the signal x is generally noisy, thus an error tolerance ϵ was used to handle this problem. This can be formally expressed as follows:

$$\min_s \|s\|_l \quad \text{s.t.} \quad \|x - sX\| < \varepsilon \quad (3)$$

In general, the minimization problem can be solved by standard ℓ_1 -magic software[1].

In order to solve nonlinear problem, we map the input data into feature spaces by kernel trick. Traditionally, the mapping is implicitly represented by specifying a kernel function as the inner product between each pair of samples in the feature space (Shawe-Taylor and Cristianini, 2004). Recently, Kernel Sparse Representation (KSR) method (Gao *et al*., 2010) has shown the necessity in face recognition, but it is difficult in optimization of sparse matrix. Conversely, the mapping in this paper, is given in an explicit form as describe in Xiong *et al*. (2005). If the rank of the kernel matrix K is r, it can be decomposed as:

$$K_{N\times N} = Q_{N\times r}\Lambda_{r\times r}Q^t_{r\times N} \quad (4)$$

where, is a diagonal matrix consisting of the *r* positive eigenvalues of K and Q consists of the corresponding orthonormal eigenvectors. Thus, the explicit mapping also called the Empirical Kernel Mapping (EKM) in Xiong *et al*. (2005), is given as $\Phi^e \chi \rightarrow F^r$:

$$x \rightarrow \Lambda^{-1/2}Q^t[ker(x, x_1),...,ker(x, x_N)]^t \quad (5)$$

Let $B = KQA^{-1/2}$and then the dot product matrix $\{\Phi^e(\chi_i)\}^N_{i=1}$ of generated by EKM can be calculated as:

$$BB^t = KQA^{-1/2}\Lambda^{-1/2}Q^1K = K \quad (6)$$

That is exactly equal to the kernel matrix in the Implicit Kernel Mapping (IKM) and, the mapped samples generated by EKM and IKM, respectively, have the same geometrical structure. In Xiong *et al*. (2005), it is shown that comparing EKM with IKM, the former is easier to access and easier to study the adaptability of a kernel to the input space than the latter. This is why we select EKM here.

In this study, the samples explicitly mapped into, where is called new view of the original input space, corresponding to kernel. Then we substitute the mapped features and basis to the formulation of sparse representation, and arrive at empirical kernel sparse representation (EKSR), the optimization problem is expressed as follows:

$$\min_s \|S_i\|_l \quad \text{s.t.} \quad \left\|\Phi^e(x_i) - S_i \bullet \{\Phi^e(x_i)\}^m_{i=1}\right\| < \varepsilon \quad (7)$$

Equation 7 is convex and the l_i minimization problem also can be solved by standard l_i-magic software. Then we can calculate the weight matrix $S = [S_1, S_2,\cdots, S_m]^T$ based on Eq. 7.

Kernel sparsity score: Although the sparse reconstructive weight matrix has no labels of data, it can reflect intrinsic geometric properties of the data and contain natural discriminating information (Qiao *et al*., 2010).

Let $X = [x_1, x_2,\cdots x_m] \in R^{n\times m}$ be the data matrix including all the training samples in its columns, x_{ij} represents the jth character of the ith sample. We expect to reconstruct each sample x_i by EKSR, using as few samples as possible. A sparse reconstructive weight vector $Si = [S_{i1},\cdots S_{i(I-1)}, 0, S_{i(I+1)},\cdots S_{im}]^T$ for each x_i can be calculated through Eq. 7, where the elements S_{ij} ($j \neq I$) denote the contribution of each x_j to reconstruct the sample x_i. That is:

$$x_i = S_{i1}x_1 + ... + S_{i(i-1)}x_{i-1} + S_{i(i+1)}x_{i+1} + ... + S_{in}x_n \quad (8)$$

Algorithm: Kernel sparsity score based on semantic

Input: Document set $X = [x_1, x_2, .., x_m] \in R^{n \times m}$
Output: Feature subset
Step 1: Calculate term probability distribution of document p (z| d) with PLSA
Step 2: Explicitly map $\{x_i\}^m_{i=1}$ into $\{\Phi^e(x_i)\}^m_{i=1}$ by kernel as shown in Eq. 5
Step 3: Construct weight matrix S using Eq. 7
Step 4: Calculate the kernel sparsity score using Eq. 10, and ascend n feature

Fig. 1: Se-KSS algorithm

After computing the weight vector S_i for each x_i, I = 1, 2, Y, m. We can define the sparse reconstructive weight matrix $S = (\tilde{S}_{ij})_{m \times m}$ as follows:

$$S = \left[\tilde{S}_1, \tilde{S}_2, ..., \tilde{S}_m\right]^T \quad (9)$$

Extended to the entire data, we construct evaluate mechanism (Dan, 2010) by the reconstruction error as follows:

$$KS\_Score_j = \sum_{i=1}^{m}(x_{ij} - \sum_{r=1}^{m} S_{ij} x_{ij})^2 \quad (10)$$

The ability of preserving sparse structure is the stronger as the value of KS_Score is the smaller. We select excellent feature subsets by the value of KS_Score, called Kernel Sparsity Score (KSS) algorithm. Based on the above discussion, we summarize the proposed algorithm as shown in Fig. 1.

EXPERIMENT AND RESULT

As described above, DF, IG, Chi and TDE (Song *et al*., 2010.) have been successfully applied to text classification. In what follows, we test the performance of our proposed algorithm compared with four popular algorithms on two datasets.

The datasets: Retuers is a subset of Retuers 21578 which has 6 different classes and contains 6223 documents for training and 2428 documents for test. Sogou is provided for classifying Chinese by Sogou Lab. We select a subset which contains 5000 documents of 5 classes, where there are 1000 samples per subject. At the moment of pretreatment, we use ICTCLAS provided by CAS for participle.

Experimental results:

- Based on K-NN classifier. To verify the effect of different topics, we evaluate the performance of the proposed method with several topics in the series of experiments using the K-NN classifier, where we set neighborhood size k = 30. In experiments, the frame of multi-class classification is one against rest. We show the performance of the number of PLSA = s topics in Fig. 2
- To verify the effectiveness of the proposed method, we evaluate the mentioned methods based on two classifiers including KNN and SVM, where the number of topics with Se-KSS selects the optimization from Fig. 2. The experimental results are presented in Table 1. For SVM, we simply use the linear kernel

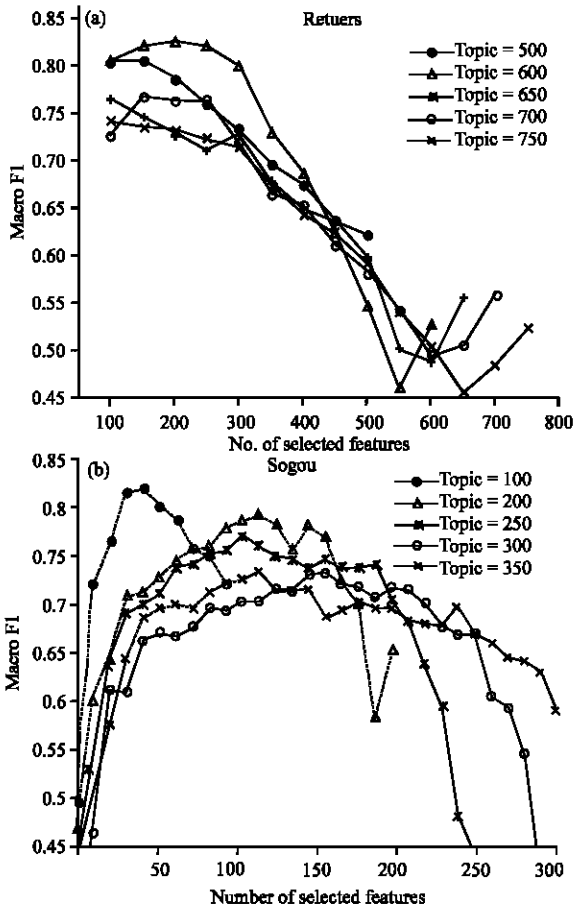

Fig. 2(a-b): The performance of k-NN classifier on the different number of PLSA = s topics

Table 1: The average F_1-Measure of different classifiers based on different feature selection

Datasets	DF	IG	CHI	TDE	Se-KSS
Reuters					
KNN	0.7772	0.7819	0.7986	0.8003	0.8238
SVM	0.8346	0.8421	0.8437	0.8648	0.9157
Sogou					
KNN	0.7886	0.7983	0.8012	0.7994	0.8226
SVM	0.8248	0.8372	0.8415	0.8536	0.8849

From above, we can draw the t as follows: (1) Se-KSS has excellent performance in classification on both datasets. This shows that by reconstructing the sparse weight matrix of data, the F_1-Measure can be improved. From Fig. 2, we know that the performance of Se-KSS is under the influence of the number of topics. According to different datasets, we should select appropriate topics. (2) DF is simple to perform, but it generally performs as well as IG and CHI. In order to improve efficiency, we can use it instead of IG and CHI. Because TDE uses information entropy to describe the distribution, its performance exceeds DF, IG and CHI. (3) . The classifiers also affect the F_1-Measure performance significantly. However, the proposed Se-KSS can generally achieve better performance than other methods based on KNN or SVM.

CONCLUSION

We introduced a non-parametric feature selection method based on Empirical Kernel Sparse Representation (EKSR) for text classification, which gets the sparse weight matrix in a high dimensional feature space mapped by empirical kernel function. Then the excellent features are acquired automatically and capture spatial information among documents. The result of experiment has shown the performance of the proposed algorithm outperforms the compared methods on the two datasets used here without label information.

However, for PLSA model, there is no way to assign probability to an unseen document and it is prone to overfitting. We can consider using Latent Dirichlet Allocation (LDA) (Blei *et al*., 2003.), which is a well-defined generative probabilistic model and generalizes easily to new documents. Secondly, the use of label information should be considered, one way is to reconstruct weight matrix using Tree Group Lasso (Liu and Ye, 2010). In the future work, we will try to overcome this limitation to further improve its performance.

ACKNOWLEDGMENTS

This study was partly supported by National Natural Science Foundation of China 60603029.

REFERENCES

Blei, D.M., A.Y. Ng and M.I. Jordan, 2003. Latent dirichlet allocation. J. Machine Learn. Res., 3: 993-1022.

Dan, S., 2010. Research on feature selection algorithms based on pairwise constraints and sparse representation. Nanjing University of Aeronautics and Astronautics, Nanjing, China.

Gao, S., I.W.H. Tsang and L.T. Chia, 2010. Kernel sparse representation for image classification and face recognition. Proceedings of the 11th European Conference on Computer Vision: Part IV, September 5-11, 2010, Heraklion, Crete, Greece, pp: 1-14.

Hofmann, T., 1999. Probabilistic latent semantic indexing. Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, August 15-19, 1999, Berkeley, CA., USA., pp: 50-57.

Hofmann, T., 2001. Unsupervised learning by probabilistic latent semantic analysis. Machine Learn., 42: 177-196.

Huang, K. and S. Aviyente, 2007. Sparse Representation for Signal Classification. In: Advances in Neural Information Processing Systems, Scholkopf, B., J. Platt and T. Hofmann (Eds.). MIT Press, Cambridge, MA, pp: 609-616.

Liu, J. and J. Ye, 2010. Moreau-Yosida regularization for grouped tree structure learning. Proceedings of the 24th Annual Conference on Neural Information Processing Systems, December 6-9, 2010, Hyatt Regency, Vancouver, BC., Canada, pp: 1-9.

Mairal, J., F. Bach, J. Ponce, G. Sapiro and A. Zisserman, 2008a. Discriminative learned dictionaries for local image analysis. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, June 23-28, 2008, Anchorage, AK., USA., pp: 1-8.

Mairal, J., F. Bach, J. Ponce, G. Sapiro and A. Zisserman, 2008b. Supervised dictionary learning. Report No. RR-6652, National Institute for Research in Computer, Paris, France, pp: 15. http://www.di.ens.fr/willow/pdfs/RR-6652.pdf.

Mallat, S.G. and Z. Zhifeng, 1993. Matching pursuits with time-frequency dictionaries. IEEE Trans. Signal Process., 41: 3397-3415.

Murray, J.F. and K. Kreutz-Delgado, 2007. Visual recognition and inference using dynamic overcomplete sparse learning. Neural Comput., 19: 2301-2352.

Qiao, L., S. Chen and X. Tan, 2010. Sparsity preserving projections with applications to face recognition. Pattern Recogn., 43: 331-341.

Shawe-Taylor, J. and N. Cristianini, 2004. Kernel Methods for Pattern Analysis. Cambridge University Press, ISBN: 0521813972 England, pp: 462.

Sivic, J., B.C. Russell, A.A. Efros, A. Zisserman and W.T. Freeman, 2005. Discovering objects and their location in images. IEEE Int. Conf. Comput. Version, 1: 370-377.

Song, J., M. Xu and C. Fan, 2010. A text feature selection method using TFIDF based on entropy. Proceedings of the 9th International FLINS Conference, August 2-4, 2010, Emei, Chengdu, China, pp: 962-967.

Wright, J., A.Y. Yang, A. Ganesh, S.S. Sastry and Y. Ma, 2009. Robust face recognition via sparse representation. IEEE Trans. Pattern Anal. Mach. Intell., 31: 210-227.

Xiong, H., M.N.S. Swamy and M.O. Ahmad, 2005. Optimizing the kernel in the empirical feature space. IEEE Trans. Neural Networks, 16: 460-474.