Available online at www.sciencedirect.com
ScienceDirect
Neural Networks 20 (2007) 748–756
Neural Networks
www.elsevier.com/locate/neunet

The learning vector quantization algorithm applied to automatic text classification tasks

M.T. Martín-Valdivia*, L.A. Ureña-López, M. García-Vega

Department of Computing, University of Jaén, Campus Las Lagunillas s/n, Edificio A3, Jaén, E-23071, Spain

Received 20 October 2005; received in revised form 12 December 2006; accepted 12 December 2006

Abstract

Automatic text classification is an important task for many natural language processing applications. This paper presents a neural approach to develop a text classifier based on the Learning Vector Quantization (LVQ) algorithm. The LVQ model is a classification method that uses a competitive supervised learning algorithm. The proposed method has been applied to two specific tasks: text categorization and word sense disambiguation. Experiments were carried out using the REUTERS-21578 text collection (for text categorization) and the SENSEVAL-3 corpus (for word sense disambiguation). The results obtained are very promising and show that our neural approach based on the LVQ algorithm is an alternative to other classification systems.

Keywords: Learning vector quantization (LVQ); Word sense disambiguation (WSD); Text categorization (TC); SENSEVAL; Reuters-21578 text collection; Natural language processing (NLP)

1. Introduction

Document classification can be thought of as a problem of mapping the space between an input document and an output class. Neural networks can learn nonlinear mappings from a set of training patterns.

A Neural Network (NN) is an interconnected assembly of simple processing elements (called units or nodes), whose structure is inspired on animal neurons. Despite a large number of successful applications of NNs in a variety of areas (see Rumelhart, Widrow, and Lehr (1994) for a survey of practical applications), their use in Natural Language Processing (NLP) tasks has not been explored sufficiently. In fact, there are no so many references as in other applications, for example in areas like optimization or pattern classification. However, NNs present various properties that NLP could take advantage of, such as massively parallel architecture, noise tolerance, self-organization and generalization. In fact, recently, interest in connecting both disciplines is growing spectacularly, as shown in Dale, Moisl, and Somers (2000).

In this paper, we discuss a neural classifier based on the Kohonen model which uses competitive supervised learning. In particular, we use the Learning Vector Quantization (LVQ) algorithm to accomplish two text classification tasks: Text Categorization (TC) and Word Sense Disambiguation (WSD).

The paper is organized as follows. First, we introduce briefly the automatic text classification task. Then, we describe the LVQ algorithm and the information representation model used in our experiments. After this, we show our evaluation environment and results for the two tasks considered (TC and WSD). Finally, we discuss our conclusion and future research.

2. Automatic text classification

Automatic text classification is one of the main tasks of NLP. Various approaches have been explored, such as support vector machine (Joachims, 1998), Naive Bayes learning methods (Lewis & Ringuette, 1994) or linear text classifiers (Lewis, Schapire, Callan, & Papka, 1996).

2.1. Neural text classification

Neural networks have also been applied to automatic text classification tasks. Certainly, the most widely used NN in

* Corresponding author. Tel.: +34 953 212898; fax: +34 953 212472.
E-mail address: maite@ujaen.es (M.T. Martín-Valdivia).

0893-6080/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.neunet.2006.12.005

Table 6
Official results for English all words

	System	Precision	Recall
1	GAMBL-AW-S	0.652	0.652
2	SenseLearned-S	0.646	0.646
...	...	...	
9	LVQ-UJAEN	0.590	0.590
...	...	...	
25	AutoPSNVs-U	0.359	0.359
26	DLSI-UA-all-Nosu	0.280	0.280

For the ELS task, this edition of SENSEVAL showed a predominance of kernel-based methods (e.g. SVM) which were used by most of the systems. For example, the best system (Htsa3) uses Regularized Least-Squares Classification (RLSC) as a learning method, which is based on kernels and Tikhonov regularization. The second system (ITC-IRST) works with the kernel function to integrate diverse knowledge sources. However, our system is the only one based on a neural approach in SENSEVAL-3.

Regarding the EAW task, the two best systems apply Memory Based Learning (MBL) using TiMBL, but there are significant differences in the performance and the approaches of the systems.

5. Conclusions and future work

This paper presents a neural approach to automatic text classification, specifically, we use the LVQ algorithm. This neural network is a supervised learning algorithm based on the Kohonen model and we use it to automatic classify a document collection according to its content.

The proposed method has been applied to accomplish two different but closely related tasks: First, we use the LVQ algorithm to categorize the REUTERS-21578 text collection; The second experiment has been carried out using the SENSEVAL-3 corpus in order to generate a disambiguator system based on the LVQ algorithm. The experiments show that the LVQ model performs successfully in both tasks.

The results obtained encourage us to continue working with the LVQ algorithm in other NLP tasks such as named entity classification, automatic summary generation or question and answering systems. We intend to apply the SOM model to carry out unsupervised text classification tasks.

Acknowledgement

This work has been supported by the Spanish Government (MCYT) with grant FIT-150500-2003-412.

References

Apte, C., Damerau, F., & Weiss, S. (1994). Automated learning of decision rules for text categorization. *Information Systems*, *12*(3), 233–251.

Baeza-Yates, R., & Ribiero-Neto, B. (1999). *Modern information retrieval*. Addison-Wesley.

Brown, P., Della-Pietra, S., Della-Pietra, V., & Merce, J. (1991). Word sense disambiguation using statistical methods. In *Proceedings of the 29th annual meeting of the association for computational linguistics*.

Chen, H., Houston, A., Sewell, R., & Schatz, B. (1998). Internet browsing and searching: User evaluations of category map and concept space techniques. *Journal of the American Society for Information Science*, *49*, 582–603.

Dale, R., Moisl, H., & Somers, H. (2000). *Handbook of natural language processing*. New York: Marcel Dekker, Inc.

Dittenbach, M., Merkl, D., & Rauber, A. (2001). Hierarchical clustering of document archives with the growing hierarchical self-organizing map. In *Proceedings of international conference on artificial neural networks* (pp. 500–505).

Eyheramendy, S., Genkin, A., Ju, W., Lewis, D., & Madigan, D. (2003). Sparse bayesian classifiers for text categorization. *Technical report. DIMACS working group on monitoring message streams*.

Fellbaum, C. (1998). *WordNet: An electronic lexical database*. The MIT Press.

García-Vega, M., García-Cumbreras, M., Martín-Valdivia, M., & Ureña López, L. (2004). The University of Jaen word sense disambiguation system. In R. Mihalcea, & P. Edmonds (Eds.), *Proceedings of Senseval-3: The third international workshop on the evaluation of systems for the semantic analysis of text* (pp. 121–124).

Goren-Bar, D., Kuflik, T., Lev, D., & Shoval, P. (2001). Automating personal categorization using artificial neural networks. In *Proceedings of user modeling 2001* (pp. 188–198).

Guerrero, V., Moya, F., & Herrero, V. (2002). Document organization using Kohonen's algorithm. *Information Processing and Management*, *38*, 79–89.

Honkela, T. (1997). Self-organizing maps in natural language processing. *Ph.D. thesis*. Helsinki University.

Honkela, T., Kaski, S., Lagus, K., & Kohonen, T. (1996). Newsgroup exploration with WEBSOM method and browsing interface. *Technical report*. Helsinki University of Technology, Espoo, Finland.

Honkela, T., Kaski, S., Lagus, K., & Kohonen, T. (1997). WEBSOM self-organizing maps of document collections. In *Workshop on self-organizing maps* (pp. 310–315). Finland, Espoo.

Hung, C., & Wermter, S. (2004). Neural network based document clustering using wordnet ontologies. *International Journal of Hybrid Intelligent Systems*, *1*(3), 127–142.

Hung, C., Wermter, S., & Smith, P. (2004). Hybrid neural document clustering using guided self-organization and wordnet. *IEEE Intelligent Systems*, *19*(2), 68–77.

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. In *Proceedings of European conference on machine learning* (pp. 137–142).

Kaski, S., Kangas, J., & Kohonen, T. (1998). Bibliography of self-organizing map (SOM) papers: 1981–1997. *Neural Computing Surveys*, *1*, 1–176.

Kilgarriff, A. (1997). What is word sense disambiguation good for? In *Proceedings of natural language processing pacific Rim symposium*.

Kilgarriff, A. (1998). Senseval: An exercise in evaluating word sense disambiguation programs. In *Proceedings of International conference on language resources and evaluation*.

Kilgarriff, A., & Palmer, M. (2000). Introduction to the special issue on Senseval. *Computers and the Humanities*, *24*, 1–13.

Kohonen, T. (1995). *Self-organization and associative memory*. Berlin: Springer-Verlag.

Kohonen, T., Kaski, S., Lagus, K., Salojarvi, J., Honkela, J., Paatero, V., & Saarela, A. (2000). Self organization of a massive document collection. *IEEE Transactions on Neural Networks*, *11*, 574–585.

Lagus, K. (2002). Text retrieval using self-organized document maps. *Neural Processing Letters*, *15*, 21–29.

Lai, K., & Lam, W. (2001). Automatic textual document categorization using multiple similarity-based models. In *Proceedings of the 1st SIAM international conference on data mining*.

Lewis, D. (1992). Representation and learning in information retrieval. *Ph.D. thesis*. Department of Computer and Information Science, University of Massachusetts.

Lewis, D., & Ringuette, M. (1994). Comparison of two learning algorithms for text categorization. In *Proceedings of the 3rd annual symposium on document analysis and information retrieval*.

Lewis, D., Schapire, R., Callan, J., & Papka, R. (1996). Training algorithms for linear text classifiers. In *Proceedings of the 19th annual international ACM SIGIR conference on research and development in information retrieval*.