Experiences with the LVQ algorithm in multilabel text categorization

A. Montejo-Ráez[1], M.T. Martín-Valdivia[1], and L.A. Ureña-López[1]

Jaén University, Jaén E-23071 Spain,
Departamento de Informática
{amontejo,maite,laurena}@ujaen.es

Abstract. Text Categorization is an important information processing task. This paper presents a neural approach to a text classifier based on the Learning Vector Quantization (LVQ) algorithm. We focus on multilabel multiclass text categorization. Experiments were carried out using the High Energy Physics (HEP) text collection. The HEP collection is an highly unbalanced collection. The results obtained are very promising and show that our neural approach based on the LVQ algorithm behaves robustly over different parameters.

1 Introduction

Text Categorization (TC) is an important task for many Natural Language Processing (NLP) applications. Given a set of documents and a set of categories, the goal of a categorization system is to assign to each document a set (possibly empty) of categories that the document belongs to. The simplest case includes only one class and the categorization problem is a decision problem or binary categorization problem (given a document, the goal is to determinate if the document is 'relevant' or 'not relevant').

On the other hand, the single-label categorization problem consist on assign exactly one category to each document while the multi-label categorization problem assigns from 0 to m categories to the same document. For a deeper and exhaustive discussion on different text categorization problem see [Sebastiani, 2002].

This work studies the multi-label categorization problem using the Learning Vector Quantization (LVQ) algorithm. The LVQ algorithm is a competitive neural learning algorithm that allows assign a category (from a set of categories) to a document (single-label problem). In order to accomplish the multi-label categorization problem we have used the LVQ algorithm as a binary classifier integrated in the TECAT Toolkit [Montejo-Ráez, 2006]. TECAT stands for TExt CATegorization. It is a tool for creating multi-label automatic text classifiers. With TECAT you can experiment with different collections and classifiers in order to build a multi-labeled automatic text classifier and implements the Adaptive Selection of Base Classifiers (ASBC) as approach to the problem.

The paper is organized as follows. First, the HEP collection is introduced. Then, we describe briefly the TECAT Toolkit as a tool to solve the multilabel

6 Conclusions

A neural algorithm for multi-label has been studied. This algorithm is based in the LVQ learning method, integrating it into the ASBC approach. Some experiments have been carried out on a high unbalanced collection: the *hep-ex* partition of the HEP corpus. The results obtained show that, despite the complexity of the collection, the robustness of the algorithm remains against different configuration parameters.

As future work we plan to apply this algorithm on a more comparable collection in the text-categorization domain, specifically Reuters-21578. For this collection, results on applying LVQ with a codebook vector per class are available [Martín-Valdivia et al., 2004]. Also, since the ASBC algorithm allows to select among a set of possible classifiers, we will test our approach with a wide range of parameterizations of the LVQ algorithm on every class, letting the system decide which parameters are the best of each class.

Acknowledgements

This work has been partially supported by a grant from the Spanish Government, project TEX-MESS (TIN2006-15265-C06-03)

References

[Ezhela et al., 2001] Ezhela, V. et al. (2001). Citations as a mean for discovery and automatic indexing of the scientific texts with new knowledge for a given subject.

[Kohonen, 1995] Kohonen, T. (1995). *Self-organization and associative memory.* Springer-Verlag, 2 edition.

[Martín-Valdivia et al., 2004] Martín-Valdivia, M., García-Vega, M., García-Cumbreras, M., and Ureña López, L. (2004). Text categorization using the learning vector quantization algorithm. In *Proceedings of Intelligent Information Systems. New Trends in Intelligent Information Processing and Web Mining (IIS:IIPWM-04)*, Zakopane, Poland. Springer-Verlag.

[Montejo-Ráez, 2006] Montejo-Ráez, A. (2006). *Automatic Text Categorization of Documents in the High Energy Physics Domain.* PhD thesis, University of Granada.

[Montejo-Ráez et al., 2004] Montejo-Ráez, A., Steinberger, R., and Ureña López, L. A. (2004). Adaptive selection of base classifiers in one-against-all learning for large multi-labeled collections. In et al., V. J. L., editor, *Advances in Natural Language Processing: 4th International Conference, EsTAL 2004*, number 3230 in Lectures notes in artifial intelligence, pages 1–12. Springer.

[Montejo-Ráez and Ureña López, 2006a] Montejo-Ráez, A. and Ureña López, L. (2006a). Binary classifiers versus adaboost for labeling of digital documents. *Sociedad Española para el Procesamiento del Lenguaje Natural*, (37):319–326.

[Montejo-Ráez and Ureña López, 2006b] Montejo-Ráez, A. and Ureña López, L. (2006b). Selection strategies for multi-label text categorization. *Lecture Notes in Artificial Intelligence*, (4139):585–592.

[Sebastiani, 2002] Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Comput. Surv.*, 34(1):1–47.