A Comparative Study of Classifier Combination Methods Applied to NLP Tasks

Fernando Enríquez, José A. Troyano, Fermín L. Cruz, and F. Javier Ortega

Departamento de Lenguajes y Sistemas Informáticos
Universidad de Sevilla
Av. Reina Mercedes s/n 41012, Sevilla, Spain
{fenros,troyano,fcruz,javierortega}@us.es

Abstract. There are many classification tools that can be used for various NLP tasks, although none of them can be considered the best of all since each one has a particular list of virtues and defects. The combination methods can serve both to maximize the strengths of the base classifiers and to reduce errors caused by their defects improving the results in terms of accuracy. Here is a comparative study on the most relevant methods that shows that combination seems to be a robust and reliable way of improving our results.

Keywords: Classifier Combination, Machine Learning.

1 Introduction

Precision and diversity are introduced by Hansen and Salamon [5] as the sufficient and necessary requirements for carrying out the combination of two or more classifier systems successfully. On the other hand Dietterich justifies the combination from three points of view, namely statistical, computational, and representational, making clear that it covers much better the search space allowing us to get closer to the optimal solution.

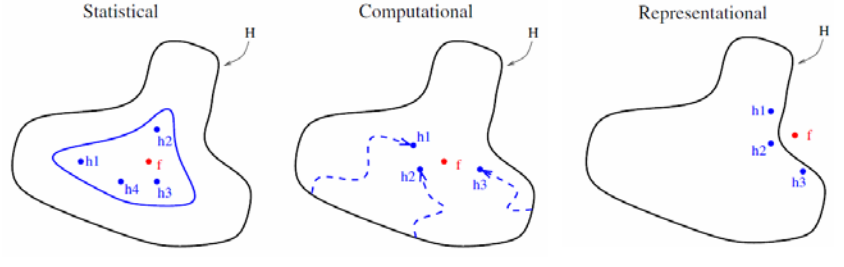

Fig. 1. Combination justification by Dietterich

In [7] the combination methods are organized in four levels based on where the combination is focused on considering the whole process. Therefore we can make

use of different collections of data (*data level*), different subsets of features used to represent the examples (*feature level*), different classifiers (*classifier level*) or different combining techniques (*combiner level*).

Still, not all existing combination methods are applicable to any set of classifiers. It is important to consider the type of information they produce as outputs. In [9] we find three possibilities: the 'abstract level' (the output is a single label or a subset of possible labels), the 'rank level' (all the tags or a subset of them are returned arranged in order of preference) and the 'measurement level' (the classifier assigns each tag a value indicative of the confidence you have in it.)

2 Combination and NLP

Since 1998, with the publication of papers like [4] and [2], many researchers developed their work using combination techniques to achieve better results on NLP tasks. Both articles were focused on POS tagging, and although many and varied works appeared afterwards, a comparative study considering a wider range of methods that could be useful to decide the most appropriate one for a particular scenery is still missing as far as we know.

After a literature review on a selection of seventy works that use some kind of combination for NLP tasks, we found the distribution of classifiers, combination methods and tasks shown in figure 2. Regarding classification and combination techniques we appreciate a very different allocation, as there is a greater balance in terms of frequency of use when we talk about classification algorithms although there are some methods that stand out slightly from the rest. In combination methods, however, voting methods and *stacking* account for most of the works, suggesting a possible lack of experimentation with other methods that could offer improvements in some NLP tasks or maybe some other characteristics that should be considered in some situations.

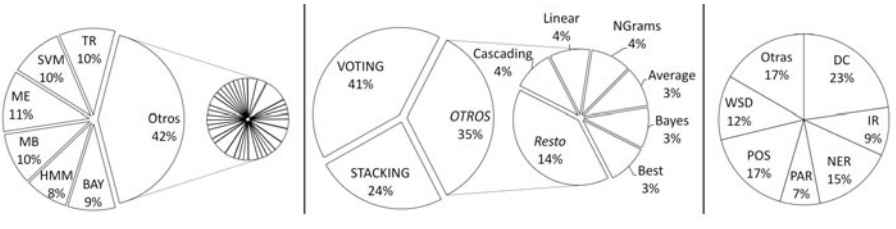

Fig. 2. Classifiers, combination methods and tasks in selected papers

With regard to the results presented in the papers, it is difficult to compare them because of the wide variety of classification and combination methods used, as well as the tasks and data, finding about fifty different datasets. Still we have prepared the table 1 which reflects the minimum, maximum and average improvements achieved in applying system combination to the NLP tasks faced by the different approaches.

Table 1. Summary of results obtained by selected papers

	min	max	mean
DC	0,01	8,10	2,02
NER	1,30	6,41	3,52
PAR	0,03	2,30	1,12
POS	-0,58	1,75	0,75
WSD	1,70	7,00	3,34

3 Comparative Study

In order to establish a more suitable setting to compare different combination methods, we performed experiments for a specific task and some specific individual base classifiers. We chose the POS task in which (thanks to the good performance of the base taggers) we can be sure that improvements obtained by the combination are not due to the low quality of the base classifiers. Three tools designed for this task have been used as base classifiers: TnT [1][1], TreeTagger [8][2] and MBT [3][3], adding a fourth classifier based on features and implemented using the software SVM^{light} [6][4]. The combination methods implemented are: Bayes (BAY), Behavior Knowledge Space (BKS), Stacked Generalization (SG), simple probabilistic combination (SPC), voting (VT) and bagging (BAG). It is also allowed to use the output of a method as input to another combination level, as if it were a base classifier, resulting in a cascading (CAS) scheme. In this case we tested with two combination levels, using a combination method to receive the outputs of other methods that work with the labels proposed by the base classifiers. All methods were evaluated using five very different corpus that differ from each other not only by their sizes but also in the language and the number of tags used in each case.

Table 2. Results obtained

CORPUS	Language	Classifiers				Combination						
		FV	MBT	TnT	TT	BAY	BKS	SG	SPC	VT	BAG	CAS
Brown	English	96,18	95,82	96,55	95,64	0,39	0,63	0,64	0,51	0,49	0,32	0,67
Floresta	Portuguese	96,52	95,81	97,02	96,66	0,55	0,72	0,78	0,60	0,63	0,36	0,71
Susanne	English	92,26	91,16	93,61	91,27	0,67	1,36	1,26	1,16	0,71	0,81	1,52
Talp	Spanish	94,59	94,80	95,82	95,62	0,96	1,08	1,10	1,10	0,76	0,75	1,18
Treebank	English	96,28	95,67	96,21	95,52	0,27	0,47	0,59	0,44	0,45	0,35	0,55
MEAN		95,17	94,65	95,84	94,94	0,57	0,85	0,87	0,76	0,61	0,52	0,93

[1] http://www.coli.uni-sb.de/thorsten/tnt
[2] http://www.ims.uni-stuttgart.de/Tools/DecisionTreeTagger.html
[3] http://ilk.uvt.nl/mbt/
[4] http://www.cs.cornell.edu/People/tj/svm_light/

Table 2 shows the results obtained by the classifiers and the improvements achieved by different combining methods. We can see that the improvements are significant in all cases being the stacking method the one with the best results obtained, showing it can be considered the best suited for dealing with different types of data because of its adaptability. Also cascading, with its two levels of combination, shows great robustness that should be taken into account. However there are also simpler methods that deserve our attention, like behavior knowledge space that achieved a noteworthy success and can be very useful in systems where speed is a critical requirement and not only the final accuracy.

References

1. Brants, T.: Tnt. a statistical part-of-speech tagger. In: In Proceedings of the 6th Applied NLP Conference ANLP 2000, pp. 224–231 (2000)
2. Brill, E., Wu, J.: Classifier combination for improved lexical disambiguation. In: Proceedings of the 17th International Conference on Computational Linguistics, pp. 191–195 (1998)
3. Daelemans, W., Zavrel, J., Berck, P., Gillis, S.: Mbt: A memorybased part of speech tagger-generator. In: Proceedings of the 4th Workshop on Very Large Corpora, pp. 14–27 (1996)
4. Halteren, H.V., Zavrel, J., Daelemans, W.: Improving data driven wordclass tagging by system combination. In: Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, vol. 1, pp. 491–497 (1998)
5. Hansen, L., Salamon, P.: Neural network ensembles. IEEE Transactions on Pattern Analysis and Machine Intelligence 12(10), 993–1001 (1990)
6. Joachims, T.: Making large-Scale SVM Learning Practical, ch. 11. MIT Press, Cambridge (1999)
7. Kuncheva, L.I., Whitaker, C.J.: Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. Machine Learning 51, 181–207 (2003)
8. Schmid, H.: Probabilistic part-of-speech tagging using decision trees. In: Proceedings of the Conference on New Methods in Language Processing (1994)
9. Xu, L., Krzyzak, A., Suen, C.Y.: Methods of combining multiple classifiers and their application to handwriting recognition. IEEE Transactions on Systems, Man, and Cybernetics 22, 418–435 (1992)