Estudio Comparativo sobre Métodos de Combinación de Clasificadores en PLN

Fernando Enríquez, José A. Troyano, Fermín 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

Resumen Existen múltiples herramientas de clasificación que pueden ser utilizadas para diversas tareas del PLN, aunque ninguna de ellas puede considerarse la mejor en términos generales ya que cada una posee una lista particular de virtudes y defectos. Los métodos de combinación pueden servirnos tanto para rentabilizar al máximo las virtudes de los clasificadores base, obteniendo mejores resultados en términos de precisión, como para disminuir los errores provocados por sus defectos. Aquí se presenta un estudio comparativo sobre los más relevantes.

Keywords: Combinación de Clasificadores, Aprendizaje Automático

1. Fundamentos de la Combinación

En Hansen y Salamon [4] se establecen la precisión y la diversidad como requisitos necesarios y suficientes para llevar a cabo con éxito la combinación de dos o más sistemas de clasificación. Por su parte Dietterich [2] justifica la combinación desde tres puntos de vista como son el estadístico, el computacional, y el de representación, dejando claro que se cubre mucho mejor el espacio de búsqueda para aproximarnos a la solución óptima.

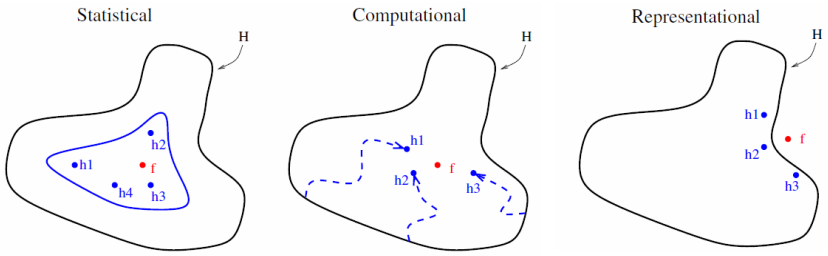

Figura 1. Justificación para la combinación según Dietterich.

En [6] se organizan los métodos de combinación en base a cuatro niveles que representan el punto del proceso donde recae el peso de la combinación,

pudiendo hacer uso de diferentes colecciones de datos (*data level*), diferentes subconjuntos de características empleadas para representar los ejemplos (*feature level*), diferentes clasificadores (*classifier level*) o distintas técnicas de combinación (*combiner level*).

Aún así, no todos los métodos de combinación existentes son aplicables a cualquier conjunto de clasificadores. Es importante considerar el tipo de información que estos producen como salida. En [7] se describen tres posibilidades: el 'abstract level' (la salida es una única etiqueta o un subconjunto de las etiquetas posibles), el 'rank level' (se devuelven las etiquetas o un subconjunto de ellas ordenadas según el orden de preferencia) y el 'measurement level' (el clasificador atribuye a cada etiqueta un valor indicativo de la confianza que se tiene en ella).

2. La Combinación y el PLN

A partir de 1998, con la publicación de los trabajos [3] y [1], fue cuando un mayor número de investigadores desarrollaron sus trabajos aplicando las técnicas de combinación a tareas del PLN. Ambos artículos se dedicaban al etiquetado POS, y aunque les sucedieron múltiples y variados trabajos, se echa en falta un estudio comparativo que abarque un mayor número de métodos y sirva para guiar al investigador a la hora de seleccionar el más adecuado.

Tras un análisis bibliográfico sobre una selección de setenta trabajos que hacen uso de alguna técnica de combinación en tareas de PLN, comprobamos la distribución de clasificadores, métodos y tareas que se muestra en la figura 2. En cuanto a las técnicas de clasificación y de combinación apreciamos un uso dispar, ya que en lo referente a los algoritmos de clasificación, si bien hay algunos métodos que destacan ligeramente del resto, existe un mayor equilibrio en cuanto a la frecuencia de uso. En los métodos de combinación sin embargo, los métodos de votación y *stacking* acaparan la mayor parte de los trabajos, dejando entrever una posible falta de experimentación con el resto de métodos que hemos comentado y que podrían ofrecer mejoras en algunas tareas del PLN.

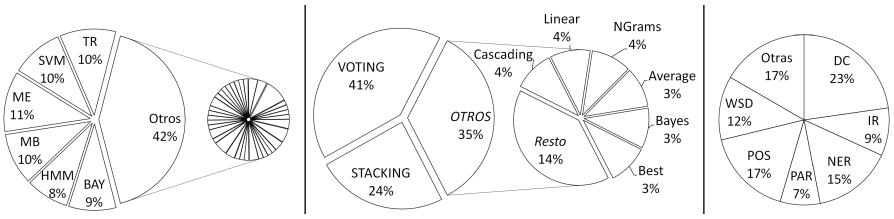

Figura 2. Clasificadores, métodos de combinación y tareas de las referencias seleccionadas.

En cuanto a los resultados presentados en los trabajos, resulta difícil establecer comparaciones debido a la gran variedad de métodos de clasificación y combinación, además de las tareas y los datos, contabilizándose cerca de 50 corpus

distintos. Aún así hemos confeccionado la tabla 1 donde se reflejan las mejoras mínima, máxima y media alcanzadas en los trabajos que aplican combinación de sistemas a alguna tarea del PLN.

	mínimo	máximo	media
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

Cuadro 1. Resumen de los resultados de las referencias seleccionadas.

3. Estudio Comparativo

Para lograr establecer un escenario más propicio para comparar los distintos métodos, hemos realizado experimentos de combinación para una tarea y clasificadores base concretos. Hemos elegido la tarea POS en la que (gracias al buen rendimiento de los etiquetadores base) podemos estar seguros de que las mejoras obtenidas por la combinación no son consecuencia de la baja calidad de los clasificadores base. Como base se han utilizado tres herramientas diseñadas para esta tarea, TnT, TreeTagger y MBT junto a un clasificador basado en características e implementado haciendo uso del software SVM^{light}[5][1] Los métodos de combinación implementados son: Bayes (BAY), *behavior knowledge space* (BKS), *stacked generalization* (SG), combinación simple de probabilidades (SPC), votación (VT) y bagging (BAG). Además se permite que la salida de un método se pueda volver a introducir como entrada en otro nivel de combinación, como si de un clasificador base se tratara, dando lugar a un esquema de *cascading* (CAS). En este caso hemos probado con dos niveles de combinación, utilizando un método de combinación para recibir las salidas del resto de métodos, que a su vez trabajan con las etiquetas propuestas por los clasificadores base. Todos los métodos han sido evaluados mediante cinco corpus muy diferentes, tanto por idioma como por su tamaño y por el conjunto de etiquetas que utilizan.

En la tabla 2 se muestran los resultados obtenidos por los clasificadores y las mejoras logradas por los diferentes métodos de combinación. Podemos comprobar que las mejoras son significativas en todos los casos, siendo *stacking* el método que mejores resultados obtiene, mostrándose como el que mejor se adapta a los diferentes tipos de datos. También *cascading*, con sus dos niveles de combinación, hace gala de una robustez que destaca al conseguir muy buen resultado. No obstante hay que destacar también los buenos resultados de métodos más

[1] http://www.cs.cornell.edu/People/tj/svm_light/

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

Cuadro 2. Resultados obtenidos.

sencillos, como el *behavior knowledge space*, que pueden resultar muy útiles en sistemas donde prima la velocidad en lugar de la precisión.

Referencias

1. E. Brill and J. Wu. Classifier combination for improved lexical disambiguation. *Proceedings of the 17th international conference on Computational linguistics*, pages 191–195, 1998.
2. T. G. Dietterich. Ensemble methods in machine learning. *In J. Kittler and F. Roli, editors, Multiple Classifier Systems, Lecture Notes in Computer Science*, 1857:1–15, 2000.
3. H.V. Halteren, J. Zavrel, and W. Daelemans. Improving data driven wordclass tagging by system combination. *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, 1:491–497, 1998.
4. L. Hansen and P. Salamon. Neural network ensembles. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(10):993–1001, 1990.
5. T. Joachims. *Making large-Scale SVM Learning Practical*, chapter 11. MIT Press, 1999.
6. L.I. Kuncheva and C.J. Whitaker. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine Learning*, 51:181–207, 2003.
7. L. Xu, A. Krzyzak, and C. Y. Suen. Methods of combining multiple classifiers and their application to handwriting recognition. *IEEE Transactions on Systems, Man, and Cybernetics*, 22:418–435, 1992.