Procesamiento de Lenguaje Natural, Revista nº 48 marzo de 2012, pp 127-130 recibido 25-01-12 revisado 12-03-12 aceptado 18-03-12

Combinación de Sistemas mediante Aprendizaje Automático en Tareas de Procesamiento de Lenguaje Natural

System Combination using Machine Learning in NLP Tasks

Fernando Enríquez y **José A. Troyano**
Dep. de Lenguajes y Sistemas Informáticos
Universidad de Sevilla
Avda. Reina Mercedes s/n
41012 Sevilla
fenros@us.es

Resumen: La combinación de sistemas constituye un área de investigación ampliamente estudiada en el ámbito del Reconocimiento de Patrones, en donde se han desarrollado múltiples técnicas para aprovechar la diversidad de métodos de clasificación de los que se dispone actualmente gracias al Aprendizaje Automático. En el desarrollo de esta Tesis Doctoral se ha realizado un estudio de las técnicas de combinación existentes y su grado de implicación en tareas del PLN. Asimismo se han expuesto algunos trabajos sobre tareas concretas y un estudio comparativo con los resultados arrojados por muchas de estas técnicas implementadas y aplicadas sobre la tarea de etiquetado morfosintáctico. El uso de un gran número de corpus diferentes y los experimentos llevados a cabo nos han permitido extraer algunas conclusiones que creemos de gran utilidad de cara al uso de estas técnicas en el futuro dentro del PLN.
Palabras clave: Aprendizaje Automático, combinación de sistemas

Abstract: The combination of systems is an area of widely studied research in the field of Pattern Recognition, where many techniques have been developed for taking advantage of the diversity of classification methods that are currently available thanks to Machine Learning. During the work implied in this PhD Thesis we have carried out a study of the existing combination techniques and their implication in NLP tasks. Some works on concrete tasks have also been exposed as well as a comparative study with the results obtained by many of these techniques implemented and deployed over the POS-tagging task. By using many different corpora and making many different experiments we have been able to draw some conclusions that can be very helpful for using these techniques in the future inside NLP.
Keywords: Machine Learning, system combination

1. Introducción

La combinación de clasificadores parte de la idea de obtener el máximo provecho de los diferentes puntos de vista que pueden aportar distintos clasificadores enfrentados a un mismo problema. Asimismo, podemos identificar dos requisitos. Por un lado la diversidad, por la que necesitamos en principio varios clasificadores y/o varios corpus, y que diferentes componentes cometan diferentes errores. Y por otro lado la precisión, requisito por el cual debemos utilizar clasificadores con una precisión mayor del 50 % para que no se limiten a introducir "ruido" en el sistema final.

En nuestro estudio hemos puesto a prueba el requisito de la diversidad, analizando diferentes escenarios en los que podríamos emplear estas técnicas e incluyendo algunos que no suelen ser tenidos en cuenta precisamente por una aparente falta de diversidad. Identificamos en primer lugar un escenario "básico" con varias técnicas de clasificación y un corpus con el que entrenar, la situación más habitual para la que se aplica la combinación. En el extremo opuesto nos podemos encontrar con un solo corpus y un solo clasificador, con la que hemos experimentado generando diversidad en los datos, creando diferentes

ISSN 1135-5948

versiones de los mismos mediante técnicas como *bagging*, y/o generando diversidad en los clasificadores, empleando diferentes configuraciones. Por último hemos experimentado con la posibilidad de contar con una o más técnicas de clasificación y diferentes corpus de diferentes tipos para combinar dicha información heterogénea en un sistema que extraiga el máximo conocimiento del conjunto de recursos disponibles.

2. La Combinación y el PLN

El PLN no tardó en hacer uso de estos métodos de combinación para sus propias tareas de clasificación, creándose una sucesión de trabajos que desde finales de los noventa siguen aportando mejoras en los resultados.

No obstante, la cobertura que se ha llevado a cabo en PLN sobre el total de métodos de combinación existentes no ha estado nunca balanceada, evidenciándose una clara tendencia al uso de técnicas de votación y *stacking* frente al uso de otros métodos sobre los que apenas hay trabajos publicados. Como muestra de ello hemos llevado a cabo un estudio bibliográfico en el que se han seleccionado un gran número de trabajos de PLN que hacen uso de la combinación de clasificadores, extrayendo información sobre los clasificadores base utilizados, las técnicas de combinación aplicadas y otros muchos datos. De este estudio se desprende que mientras el tipo de clasificadores es muy variado, en efecto las técnicas de combinación muestran a la votación y el *stacking* como claros "favoritos" como se aprecia en la figura 1.

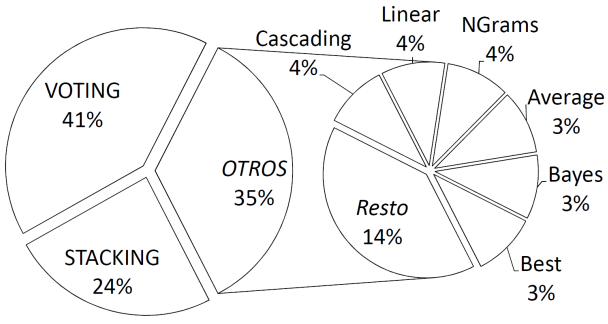

Figura 1: Métodos de combinación en PLN

Otras conclusiones que debemos resaltar extraídas de este estudio es que la combinación logra un mayor beneficio cuando se aplica a tareas difíciles debido al mayor margen de mejora que existe de inicio, y que según estos trabajos existe la posibilidad de mejorar los métodos más utilizados con otros mucho menos conocidos, de ahí que considerásemos necesaria la realización de un estudio comparativo entre diferentes técnicas.

3. Estudio Comparativo

Para lograr establecer un escenario más propicio para comparar los distintos métodos, hemos realizado experimentos de combinación empleando una tarea y clasificadores base concretos. Hemos elegido la tarea de etiquetado POS (Part-Of-Speech) en la que se clasifican las palabras en categorías morfosintácticas (verbo, nombre, etc.). En esta tarea y 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 mismos. Como base se han utilizado tres herramientas diseñadas para esta tarea, TnT (Brants, 2000), TreeTagger (Schmid, 1994) y MBT (Daelemans et al., 2003) junto a FV, un clasificador basado en características e implementado haciendo uso del software SVM^{light}(Joachims, 1999)[1]. Algunos de los métodos de combinación implementados son: Bayes(French, 1985) (BAY), *behavior knowledge space* (Huang y Suen, 1995) (BKS), *stacked generalization* (Wolpert, 1992) (SG) y votación (Arrow, 1951) (VT). 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* (Maudes, Rodríguez, y García-Osorio, 2007) (CAS). Si bien hay semejanzas con SG al contar este con un nivel adicional de clasificación, en el *cascading* se establece la posibilidad de encadenar por niveles los métodos de combinación que deseemos sin ser por sí mismo un método de combinación como es SG. 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 corpus muy diferentes, tanto por idioma como por su tamaño y por el conjunto de etiquetas que utilizan. Entre ellos figuran los corpus: Brown[2] (B), Flo-

[1] http://www.cs.cornell.edu/People/tj/svm_light/
[2] http://www.hit.uib.no/icame/brown/bcm.html

resta[3] (F), Susanne[4] (S), CLiC-TALP[5] (T) y Penn Treebank[6] (P). Brown, Susanne y Penn Treebank están en inglés, destacando el número de etiquetas utilizada por Susanne, 131, muy superior al resto. Por su parte, Floresta está en portugués y Talp en español.

	B	F	S	T	P	$\bar{x}$
TnT	96,55	97,02	93,61	95,82	96,21	95,84
TT	95,64	96,66	91,27	95,62	95,52	94,94
MBT	95,82	95,81	91,16	94,80	95,67	94,65
FV	96,18	96,52	92,26	94,59	96,28	95,17
BAY	0,39	0,55	0,67	0,96	0,27	0,57
BKS	0,63	0,72	1,36	1,08	0,47	0,85
SG	0,64	0,78	1,26	1,10	0,59	0,87
VT	0,49	0,63	0,71	0,76	0,45	0,61
CAS	0,67	0,71	1,52	1,18	0,55	0,93

Tabla 1: Resultados de los clasificadores base (*accuracy*) y mejoras obtenidas combinando respecto al mejor clasificador.

Todos los resultados de los clasificadores base están por encima del 90 % de precisión, rondando el 95 % de media. En la tabla 1 se reflejan las mejoras logradas por los diferentes métodos de combinación respecto al *accuracy* del mejor clasificador base participante. Podemos comprobar que las mejoras son significativas en todos los casos, siendo *stacking* el método de combinación que mejores resultados obtiene, mostrándose como el que mejor se adapta a los diferentes tipos de datos. Mejor aún es el resultado aplicando el esquema de ejecución de *cascading* con sus dos niveles de combinación. En su caso hace gala de una robustez que destaca al conseguir muy buenos resultados tras experimentar con diferentes técnicas en su segundo nivel de combinación. No obstante hay que destacar también a métodos más sencillos, como el *behavior knowledge space*, que pueden resultar muy útiles en sistemas donde prima la velocidad y simplicidad del método en lugar de la precisión.

También se realizaron experimentos adicionales que pretenden explorar diferentes aspectos de la combinación y distintos escenarios de los comentados en la introducción, mostrándose en la tabla 2 algunos de los resultados obtenidos.

[3]http://www.linguateca.pt/Floresta/
[4]http://www.grsampson.net/Resources.html
[5]http://clic.fil.ub.es/demos/
[6]http://www.ldc.upenn.edu/

En primer lugar quisimos comprobar qué ocurre al eliminar el mejor clasificador participante del esquema de combinación, tras lo que se consiguen mejoras mayores, confirmando que la combinación es un recurso muy valioso en situaciones adversas, en las que no contamos con los mejores clasificadores para la tarea encomendada.

Después vimos como afectaba la calidad del corpus en los resultados, para lo que pusimos en práctica las técnicas de combinación tras reducir el tamaño del corpus Penn Treebank de las 800.000 palabras que contiene aproximadamente (P_{800}), a 50.000 (P_{50}), 100.000 (P_{100}) y 200.000 (P_{200}). En los resultados SG se muestra como el más eficiente sufriendo menos la pérdida de información que los clasificadores base, aportando mayores mejoras respecto al mejor clasificador conforme disminuye el tamaño del corpus.

En tercer lugar combinamos con SG los resultados proporcionados por los tres posibles valores de los parámetros de ejecución 'n' y 'd' de TnT, que permiten elegir el tamaño de n-gramas y el método de suavizado respectivamente. En la tabla aparece la mejora respecto a la mejor versión de TnT de entre las participantes en la combinación (n-A y d-A) y también la mejora respecto a la versión que utiliza el valor por defecto del parámetro estudiado (n-B y d-B). Se observa que aunque el valor por defecto casi siempre es el que aporta mejores resultados, en algún caso se obtiene una recompensa adicional por utilizar combinación respecto a optar por los valores por defecto.

Finalmente mostramos los resultados de combinar los clasificadores FV generados mediante diferentes conjuntos de características o *features*. En cada versión de FV eliminamos un tipo de características: léxicas (FV-l), de n-gramas (FV-n), de prefijos y sufijos (FV-f). Posteriormente realizamos la combinación mediante SG de estas 3 versiones de FV (C-A) y de estas 3 versiones más la versión completa de FV (C-B).

Destacamos que en los dos últimos casos, combinando parámetros de TnT y versiones de FV, se muestra la eficacia de la combinación utilizando un único clasificador base y generando la diversidad de forma artificial.

Otra línea de experimentación ha consistido en combinar mediante SG información heterogénea. En este caso se combina la información aportada por varios clasificadores ba-

		B	F	S	T	P	$\bar{x}$
1	BAY	0,64	0,86	1,35	0,93	0,07	0,77
	BKS	0,86	0,98	2,19	1,09	0,15	1,05
	SG	0,94	0,96	1,95	1,11	0,31	1,05
	VT	0,66	0,82	1,24	0,85	0,08	0,73
		P_{800}	P_{200}	P_{100}	P_{50}	-	$\bar{x}$
2	BAY	0,27	0,33	0,34	0,25	-	0,30
	BKS	0,47	0,57	0,49	0,22	-	0,44
	SG	0,59	0,56	0,65	0,73	-	0,63
	VT	0,45	0,47	0,41	0,36	-	0,42
		B	F	S	T	P	$\bar{x}$
3	n-A	0,14	0,16	0,39	0,24	0,22	0,23
	d-A	0,12	0,13	0,34	0,05	0,19	0,17
	n-B	0,14	0,16	0,39	0,24	0,22	0,23
	d-B	0,13	0,13	0,34	0,29	0,19	0,22
		B	F	S	T	P	$\bar{x}$
4	C-A	0,35	0,09	0,08	0,11	0,19	0,16
	C-B	0,36	0,11	0,20	0,20	0,19	0,21
		B	C	P	-	-	$\bar{x}$
5	SG	0,34	0,49	0,31	-	-	0,38
	SGh	0,44	0,61	0,48	-	-	0,51

Tabla 2: Mejoras obtenidas sobre el *accuracy* de los clasificadores en los experimentos adicionales: 1. Eliminando el mejor clasificador, 2. Reduciendo el tamaño del corpus, 3. Combinando diferentes valores de los parámetros de TnT, 4. Combinando diferentes conjuntos de *features* con FV y 5. Combinando información heterogénea con SG (se muestra SG "normal" como referencia).

se utilizando diferentes corpus para etiquetar el contenido de un corpus "objetivo". Como ejemplo de los beneficios que se pueden obtener mostramos en la tabla 2 como se combinan TnT, TreeTagger y MBT entrenados con diferentes corpus para etiquetar uno de ellos. A los corpus Brown(B) y Penn Treebank (P) utilizados anteriormente añadimos el corpus de la competición CoNLL del 2000 de la que se ha extraído la etiqueta POS, dando lugar al corpus (C). Se añade además información léxica sobre las palabras a etiquetar, como por ejemplo si empieza con mayúscula o si contiene dígitos. En la primera columna mostramos las mejoras utilizando solo los clasificadores base sobre el corpus objetivo, sin información heterogénea para poder comparar con el sistema completo. Las mejoras demuestran que SG permite aprovechar todos los recursos a nuestra disposición de forma muy satisfactoria.

4. Conclusiones

En este trabajo se han estudiado los métodos de combinación desde diferentes puntos de vista, contemplando tanto los aspectos teóricos fundamentales, los datos históricos y las implicaciones prácticas derivadas de la experimentación propia. Todos los resultados aquí mostrados demuestran el gran potencial de la combinación en tareas de clasificación y por ello creemos que los métodos de combinación representan una gran oportunidad para mejorar los sistemas actuales y que el desarrollo de herramientas que saquen provecho de sus virtudes puede impulsar los resultados obtenidos en tareas de clasificación y por extensión, a las tareas del PLN que se basan o hacen uso de ellas.

Bibliografía

Arrow, K. 1951. *Social Choice and Individual Values*. Wiley, New York.

Brants, T. 2000. Tnt. a statistical part-of-speech tagger. *In Proceedings of the 6th Applied NLP Conference (ANLP00)*, páginas 224–231.

Daelemans, W., J. Zavrel, A.v.d. Bosch, y K.v.d. Sloot. 2003. Mbt: Memory-based tagger, reference guide. Informe Técnico 03-13, ILK.

French, S. 1985. Group consensus probability distributions: a critical survey. *Bayesian Statistics*, 2:183–202.

Huang, Y. S. y C. Y. Suen. 1995. A method of combining multiple experts for the recognition of unconstrained handwritten numerals. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17:90–93.

Joachims, T., 1999. *Making large-Scale SVM Learning Practical*, capítulo 11. MIT Press.

Maudes, J., J.J. Rodríguez, y C. García-Osorio. 2007. Cascading for nominal data. *Multiple Classifier Systems, Lecture Notes in Computer Science*, 4472:231–240.

Schmid, H. 1994. Probabilistic part-of-speech tagging using decision trees. *In Proceedings of the Conference on New Methods in Language Processing.*

Wolpert, D.H. 1992. Stacked generalization. *Neural Networks*, 5:241–259.