Técnicas de reducción, algoritmos resistentes al ruido o ambos. Opciones para el manejo de rasgos clasificatorios en la atribución de autoría.

Antonio Rico Sulayes

Universidad Autónoma de Baja California, Calzada Universidad 14418, Parque Industrial Internacional, Tijuana, Baja California, México
antonio.rico@uabc.edu.mx

Resumen. Un problema en las tareas clasificatorias es el manejo de los rasgos que caracterizan las clases. Cuando la lista de rasgos es larga, se puede usar un algoritmo resistente al ruido de rasgos irrelevantes, o reducir dichos rasgos. La atribución de autoría, tarea que asigna un texto anónimo a un sujeto en una lista de posibles autores, ha sido ampliamente abordada como una tarea de clasificación automática de texto. En dicha tarea, los *n-gramas* pueden producir listas largas de rasgos incluso en corpus pequeños. A pesar de esto, falta una investigación que exponga los efectos de usar algoritmos resistentes al ruido, reducir los rasgos, o combinar ambas opciones. Este trabajo responde a esta carencia utilizando contribuciones a foros de discusión relacionados con el crimen organizado. Los resultados obtenidos muestran que los clasificadores evaluados, en general, se benefician con la reducción de rasgos, y que gracias a dicha reducción, incluso algoritmos clásicos superan a clasificadores de punta considerados altamente resistentes al ruido.

Palabras clave: Atribución de autoría, Rasgos clasificatorios, Algoritmos resistentes al ruido, Reducción de rasgos.

1 Listas de rasgos en la atribución de autoría

Si se define a la atribución de autoría como la asignación de un texto anónimo a un sujeto dentro de una lista de posibles autores, esta tarea constituye un problema de clasificación de textos. Ahora bien, si este problema se aborda utilizando métodos automatizados, el problema le compete a la clasificación automática de textos, área atendida por la recuperación de información. Como un problema de la clasificación automática de textos, la atribución de autoría utiliza dos elementos primordiales. Por un lado, requiere una selección de rasgos clasificatorios, que discriminan a los elementos de las diversas clases. Por otro lado, la atribución de autoría utiliza un método de clasificación que procesa los rasgos. En este contexto, el método clasificatorio se emplea para atribuir un cierto texto a un sujeto en específico.

Respecto del primer elemento esencial para la atribución de autoría, la selección de rasgos clasificatorios, los investigadores dedicados a dicha tarea han propuesto nue-

vos rasgos de manera constante por varias décadas. A finales del siglo pasado, más de 1,000 rasgos diferentes fueron identificados en más de 300 trabajos dedicados a esta tarea clasificatoria [1]. Este número de rasgos se ha incrementado dramáticamente en años recientes debido a la introducción de rasgos textuales que son etiquetados de manera automática, como por ejemplo, los n-gramas. Los n-gramas fácilmente producen listas de varios millares de elementos incluso en corpus (colecciones de textos de lenguaje natural) relativamente pequeños, como los que típicamente se utilizan en la atribución de autoría [2,3,4].

En contraposición a la proliferación de rasgos en la atribución de autoría, los resultados de investigaciones recientes surgieren que la selección de los mismos es el elemento primordial en la mejora de resultados para esta tarea [5,6,7]. Según las investigaciones mencionadas, esta selección es incluso más importante que la puesta a punto de los algoritmos de clasificación. En la clasificación de textos en general, la motivación para el uso de listas reducidas de rasgos obedece a que los rasgos altamente discriminatorios son más eficientes y obtienen una mayor precisión en los resultados [8]. Estos rasgos discriminatorios evitan el ruido de las listas extensas, las cuales incluyen rasgos redundantes o poco discriminatorios. Este tipo de rasgos es particularmente ineficiente cuando se aplica a nuevos conjuntos de datos. Una respuesta que se ha dado a este problema en la atribución de autoría son los algoritmos avanzados para la clasificación de textos, tales como las máquinas de vectores de soporte (SVMs, por sus siglas en inglés), que pueden compensar el ruido de las largas listas de rasgos [3], [9]. La segunda respuesta es el uso de técnicas para la reducción de dichas listas, como la selección de rasgos con mayor frecuencia o con índices altos de información mutua. Esta solución también ha sido extensamente empleada en la atribución de autoría [10]. Frente a estas dos posibilidades en el manejo de rasgos clasificatorios para la atribución de autoría (la utilización de algoritmos que compensan el ruido de los rasgos no discriminatorios o redundantes y la reducción alternativa de las listas de rasgos con recursos externos al clasificador), la literatura especializada no ha comparado los resultados de ambas opciones.

Este artículo responde a esta carencia, comparando los métodos de clasificación y de reducción de rasgos más comunes en la atribución de autoría. Además este trabajo introduce un método de reducción de rasgos nunca antes usado en esta tarea. En la evaluación de las diferentes combinaciones entre clasificadores y técnicas de reducción de rasgos, el presente trabajo utiliza datos extraídos de medios sociales relacionados con el crimen organizado en México. El artículo concluye mostrando que los métodos de clasificación con una larga tradición en la atribución de autoría [1] se pueden combinar con técnicas de reducción de rasgos (tanto técnicas conocidas como de nueva aplicación en este contexto) y que esas combinaciones igualan y superan los resultados obtenidos por clasificadores de última generación.

2 Los medios sociales y el crimen organizado

La guerra contra el tráfico de drogas en México se caracterizó desde un inicio por un creciente y luego constante número de muertes [11,12,13,14]. Además de esta característica, esta lucha también se ha distinguido por el continuo uso de comunicaciones escritas enviadas por miembros del crimen organizado tanto al gobierno como

a la población en general [7]. Entre estas comunicaciones se encuentran los comentarios publicados por usuarios en medios sociales dedicados a este tema. Este tipo de medios sociales han proliferado en años recientes en este país, con algunos resultados lamentables, como el asesinato de sus usuarios por grupos criminales y el eventual cierre de algunos de estos sitios [15,16,17]. A pesar de estos hechos, este tipo de medios sigue existiendo hasta el día de hoy [18].

El presente trabajo utiliza las contribuciones de usuarios publicadas en uno de los primeros sitios relacionados con el crimen organizado en México. Este sitio, creado en abril de 2010, albergó en su inicio un foro de discusión dedicado a este tema [7]. Las contribuciones de este foro utilizadas aquí fueron recuperadas copiando todos los mensajes publicados durante el primer medio año de vida del foro. Esto permitió recuperar 41,751 mensajes publicados en 4,205 conversaciones. Una vez depurados todos los mensajes recuperados (eliminando copias y mensajes de usuarios anónimos) se identificaron 37,571 mensajes creados por usuarios registrados en el foro. Estos mensajes pertenecen a 1,026 usuarios diferentes y contienen un total de 2,128,049 instancias de palabras o tokens.

2.1 Datos experimentales

Con los datos recuperados del foro de discusión referido arriba, se crearon varios corpus para explorar los efectos de la reducción de rasgos en combinación con diversos clasificadores, comunes en la atribución de autoría. Entre los 1,026 usuarios que produjeron mensajes usando una cuenta de usuario, se seleccionaron aquellos que tenían un mínimo de 40 mensajes individuales con un mínimo de 2,000 palabras de texto original en la suma de todos sus mensajes. Con estos dos criterios de selección, se identificaron 106 usuarios del foro que cumplían con dichos criterios.

El mínimo de palabras de texto original que se puso como requisito para seleccionar a los usuarios del foro (2,000 palabras) se encuentra en el extremo inferior de lo que estudios previos han utilizado en la atribución de autoría. Por ejemplo, entre los investigadores que reportan este dato experimental algunos han empleado 2,000, 8,000, 15,000, 33,000, 40,000, y 55,000 palabras, [19], [20], [21], [22], [23], [2], respectivamente. Esta cantidad de texto es utilizada como datos de entrenamiento para representar a cada sujeto en el conjunto de autores potenciales durante la clasificación. El otro criterio de selección de sujetos como autores potenciales (un mínimo de 40 mensajes), se utilizó para descartar usuarios esporádicos con escasos mensajes de cierta extensión. Muestreando aleatoriamente los 106 seleccionados, se identificaron 40 usuarios con los que se construyeron 39 corpus. En estos corpus, el número de sujetos va de un mínimo de 2 hasta un máximo de 40.

Dividiendo las 2,000 palabras de texto original de cada sujeto, se construyeron 4 sub-muestras de aproximadamente 500 palabras cada una. Estas sub-muestras fueron construidas agregando aleatoriamente mensajes, de entre todas las contribuciones de cada uno de los usuarios del foro, conservando la integridad de los mensajes individuales. Cada una de estas sub-muestras se utilizó como una unidad en los datos de prueba empleados para la clasificación. El rango en el tamaño de estas sub-muestras de prueba (478-541 palabras) se encuentra también en el extremo inferior de lo que

estudios previos que utilizan mensajes íntegros han empleado. Por ejemplo, algunos estudios usan 99-608, 100-1000, 500-2000, 600, 628-1342, y 7,500+ palabras en las unidades de sus datos de prueba, [19], [24], [22], [23], [21] y [20], respectivamente. Debido a que las contribuciones de cada usuario seleccionado fueron agregadas en 4 sub-muestras, los experimentos realizados con cada uno de los 39 corpus comprendieron desde 8 atribuciones (en el corpus más pequeño con 2 autores) hasta 160 atribuciones (en el corpus más grande con 40 sujetos).

3 Atribución de autoría como clasificación automática de textos

La atribución de autoría ha sido ampliamente abordada, tanto por investigadores de la recuperación de la información, así como de la lingüística y la estilística forense [5]. En todas estas disciplinas, establecer rasgos de comparación y clasificación entre los textos de diversos autores es una necesidad esencial. Sólo a partir de estos rasgos se pueden discriminar los textos, los cuales son tratados como anónimos durante la puesta a prueba del método de clasificación. Ahora bien, en el caso particular de la recuperación de información, el procesamiento de los rasgos mencionados es realizado siempre de manera automática con un algoritmo clasificador. Además, algunos de los clasificadores comunes en esta área, los cuales son considerados de última generación, han surgido dentro del aprendizaje automático.

3.1 Rasgos clasificatorios de autoría

En cuanto a los rasgos que se utilizaron para realizar la clasificación, se partió de una selección previa de rasgos léxicos, sintácticos y estructurales. Los rasgos léxicos incluyeron una lista de todos los unigramas de palabras (equivalente a todos los tipos o formas léxicas diferenciadas). Como esta lista es dependiente del corpus del que es extraída, el tamaño de la misma varió entre 1,402 tipos para el corpus más pequeño, hasta 13,089 para el corpus más grande. Cabe aclarar que la puntuación fue removida de las unidades léxicas a las que se unía y los signos separados fueron utilizados como unigramas léxicos independientes, un procedimiento común en la atribución de autoría [9], [20], [25] y [26]. En cuanto a los rasgos sintácticos, se utilizó una lista previamente recabada (para otra tarea clasificatoria) con elementos léxicos funcionales pluriverbales, es decir, con más de una palabra. Estos elementos se componen principalmente de una preposición más otros elementos léxicos, como ‘después de(l)’ o ‘lejos de(l)’, o de una conjunción combinada con otras palabras, como ‘después de que’ o ‘mientras que’. La lista predeterminada de elementos léxicos funcionales pluriverbales, cuyas instancias fueron etiquetadas en los corpus, tenía 132 elementos en total, con 68 bigramas, 56 trigramas, and 7 tetragramas. Finalmente, los rasgos estructurales estaban dados por una lista preseleccionada de 19 elementos, varios de ellos previamente utilizados por el autor de este artículo [7]. Entre estos rasgos se incluyen diversos rasgos de formato del texto como el uso de subrayados, negritas, imágenes, colores y tamaños especiales de letras. Igualmente, estos rasgos comprenden elementos propios de la comunicaciones electrónicas, como el uso hipervínculos, activos y

no activos, emoticones, en imágenes y representados con caracteres del teclado, así como la reduplicación excesiva de signos de puntuación, como suelen emplearse los signos de admiración. Dada esta preselección de rasgos de autoría, el número de rasgos etiquetados (todos automáticamente) en los 39 corpus va de 1,553 para el corpus con 2 autores solamente, hasta 13,249, para el corpus con el máximo de 40 sujetos.

3.2 Algoritmos de clasificación en la atribución de autoría

Las investigaciones sobre atribución de autoría han empleado una variedad de algoritmos clasificatorios. En una revisión exhaustiva de 32 trabajos dedicados a esta tarea y publicados en la última década, se identificaron 23 algoritmos diferentes de clasificación [10]. Aunque muchos de los clasificadores identificados aparecen sólo una vez en la literatura, algunos algoritmos han sido utilizados en varias investigaciones. Entre estos algoritmos se encuentran diferentes implementaciones del árbol de decisión, C4.5, formas varias del análisis bayesiano (como el multivariante y el basado en el modelo de Bernoulli), diferentes tipos de redes neuronales (como las artificiales y las llamadas redes neuronales de retropropagación [9]) y las SVMs. También son comunes en la atribución de autoría algunos clasificadores propios de la estadística, como el análisis discriminante (AD) y los clasificadores basados en la prueba de Chi cuadrado. Además, hay que mencionar que 10 de los 32 trabajos revisados en el estudio citado utilizan más de un algoritmo clasificador y comparan los resultados obtenidos por los diferentes algoritmos seleccionados.

En cuanto a los clasificadores puestos a prueba en este trabajo, se escogieron los 4 algoritmos que han dado los mejores resultados en la atribución de autoría, según la revisión bibliográfica exhaustiva antes mencionada [10]. Los 4 clasificadores son el AD, el análisis bayesiano multivariante (ABM), el análisis bayesiano de Bernoulli (ABB) y las SVMs. A estos clasificadores se ha sumado el algoritmo más comúnmente usado como base de referencia en esta tarea, el árbol de decisión C4.5, en su implementación para Weka, J4.8.

3.3 Técnicas de reducción de rasgos en la atribución de autoría

El estudio mencionado en la sección anterior [10], reporta también un abundante uso de técnicas de reducción de rasgos en las investigaciones dedicadas a la atribución de autoría. De los 32 estudios examinados, 17 fueron identificados por su uso de alguna técnica de reducción de rasgos, o de algún método de evaluación de rasgos que permite reducir el conjunto de los mismos. Con menos variación que el uso de clasificadores, las técnicas de reducción de rasgos identificadas incluyen la ganancia de información (GI), la frecuencia (relativa, absoluta o normalizada), el análisis de componentes principales, algunos métodos de evaluación de rasgos de la estadística en general (el análisis de la varianza, ANOVA, el análisis de la covarianza, ANCOVA, y la ANOVA de dos vías), y dos métodos paso a paso, la distancia de Mahalanobis y la Lambda de Wilks. Un estudio también utiliza una lista de palabras vacías para eliminarlas de su lista completa de rasgos. Aunque sólo dos estudios comparan más de una

técnica de reducción, el de uso de dos de ellas es notoriamente más común que el de las demás: la frecuencia, empleada en seis estudios, y la GI, en tres.

En este trabajo, se escogieron las dos técnicas de reducción más comunes, la frecuencia y la GI, que son también aquellas con que se han reportado los mejores resultados en los estudios comparativos. La frecuencia aquí ha sido expresada como frecuencia absoluta con un mínimo de 4 instancias, número igual al número de sub-muestras por autor. Además se decidió incluir una técnica de reducción nueva en la atribución de autoría, la llamada "selección de sub-conjuntos de rasgos basada en correlaciones" (CFS, por sus siglas en inglés). Esta técnica de reducción, descrita por primera vez en [27], se incluyó porque fue diseñada con la intención explícita de mejorar el rendimiento de algoritmos basados en el análisis bayesiano. Como se mencionó arriba, dos versiones de estos algoritmos, ABM y ABB, fueron utilizados aquí.

Por otro lado, la lista completa de rasgos sin reducción constituyó la cuarta alternativa para combinar con cada algoritmo de clasificación. En esta cuarta opción es trabajo exclusivo del clasificador compensar el ruido de los rasgos no discriminatorios o redundantes. Dados los cinco clasificadores y las cuatro listas de rasgos (tres reducidas y una sin ninguna reducción), hay un total de 20 configuraciones aplicadas a los 39 corpus. Todas estas opciones dan origen a un total de 780 experimentos.

4 Resultados

Respecto de la precisión cuyos promedios se reportan a continuación, ésta representa la proporción de verdaderos positivos o asignaciones correctas de sub-muestras de prueba a sus verdaderos autores. Para la obtención de dicha precisión en experimentos individuales, se aplicaron un clasificador y una lista de rasgos (reducida o sin ninguna reducción) en la asignación de todas las sub-muestras de prueba a su respectivos autores. Además, el cálculo de la precisión en los experimentos individuales se llevó a cabo por medio de un diseño de validación cruzada. En cuanto a las figuras presentadas a continuación en la Tabla 1, éstas representan el promedio de la precisión obtenida por la combinación de un clasificador y una lista de rasgos al aplicarse paulatinamente a los 39 corpus.

Tabla 1. Resultados promedio de la clasificación en los 39 corpus

Técnica de Reducción	Clasificador				
	C4.5	AD	ABM	ABB	SVMs
Ninguna	0.456	0.313	0.743	0.831	0.457
Frecuencia	0.489	0.466	0.942	0.821	0.829
GI	0.671	0.700	0.947	0.848	0.775
CFS	0.660	0.811	0.940	0.820	0.726

En su conjunto, la Tabla 1 muestra los resultados en promedio, sobre todos los 39 corpus, de los 5 clasificadores en combinación con las 4 listas de rasgos. Como se puede observar en la tabla, los mejores resultados obtenidos con el algoritmo de base de referencia, C4.5, se logran al combinar el mismo con la técnica de reducción por GI. Lo mismo sucede con los clasificadores basados en análisis bayesianos, el ABM y el ABB. En los tres clasificadores que obtienen sus mejores resultados al combinarse con la lista reducida por GI, la precisión promedio con todos los corpus incrementa de 0.456 sin ninguna reducción a 0.671 para C4.5, de 0.743 a 0.947 para el ABM, y de 0.831 a 0.848 para el ABB. El mejor resultado en promedio sobre todos los corpus para el clasificador del AD se da con la técnica de reducción de CFS. Con esta técnica, la precisión promedio mejora de 0.313 sin ninguna reducción a 0.811. Finalmente, con el clasificador de última generación, basado en SVMs, el mejor resultado en promedio se obtiene utilizando la frecuencia absoluta mínima de 4 instancias. Con este criterio, la precisión promedio mejora de 0.457 sin ninguna reducción a 0.829.

5 Conclusiones

Con la perspectiva global que ofrece la Tabla 1, es posible regresar a la lista de opciones para el manejo de rasgos clasificatorios mencionada en el título de este trabajo. De esa lista se desprende la pregunta: ¿Cuál es la mejor opción para manejar listas largas de rasgos en la atribución de autoría: aplicar técnicas de reducción de rasgos, usar algoritmos de clasificación resistentes al ruido o combinar estas dos opciones? En cuanto a la primera respuesta a la pregunta, la aplicación de técnicas de reducción de rasgos, la Tabla 1 muestra que todos los clasificadores se benefician claramente de dicha aplicación. Al mismo tiempo, se aprecian tres tendencias en el beneficio obtenido por medio de las diversas técnicas de reducción. Por un lado, dos clasificadores, C4.5 y el AD, observan una mejoría moderada con la lista reducida por medio de la frecuencia. La mejoría va de 0.456 a 0.489 para C4.5 y de 0.313 a 0.466 para el AD. Sin embargo, la mejora en la precisión de estos dos clasificadores es mayor con las técnicas que requieren un cálculo más elaborado y que, probablemente de manera más relevante, producen listas de rasgos mucho más cortas. En este sentido cabe mencionar que el número de rasgos en la lista reducida con el criterio de frecuencia va de 229 en el corpus con dos autores a 2,168 en el corpus con 40 autores, mientras que GI y CFS generan listas considerablemente más pequeñas, con rangos de 23-75 rasgos y de 7-27 rasgos, respectivamente. La segunda tendencia que se observa en la aplicación de técnicas de reducción está dada por otros dos clasificadores, el ABM y las SVMs. Estos algoritmos se benefician notoriamente con todas las técnicas de reducción de rasgos. Finalmente, el ABB exhibe una tercera tendencia ya que no siempre se beneficia con la reducción de rasgos. Sin embargo, logra una mejoría moderada, respecto de la lista sin reducción, al alimentarse con la lista reducida por GI.

A continuación, la Figura 1 muestra la precisión obtenida por la combinación de la lista de rasgos seleccionados por GI y el ABM. Esta combinación obtuvo la precisión promedio más alta sobre todo los corpus. En la misma figura, se ha incluido la precisión obtenida por el ABM sin ninguna reducción de rasgos.

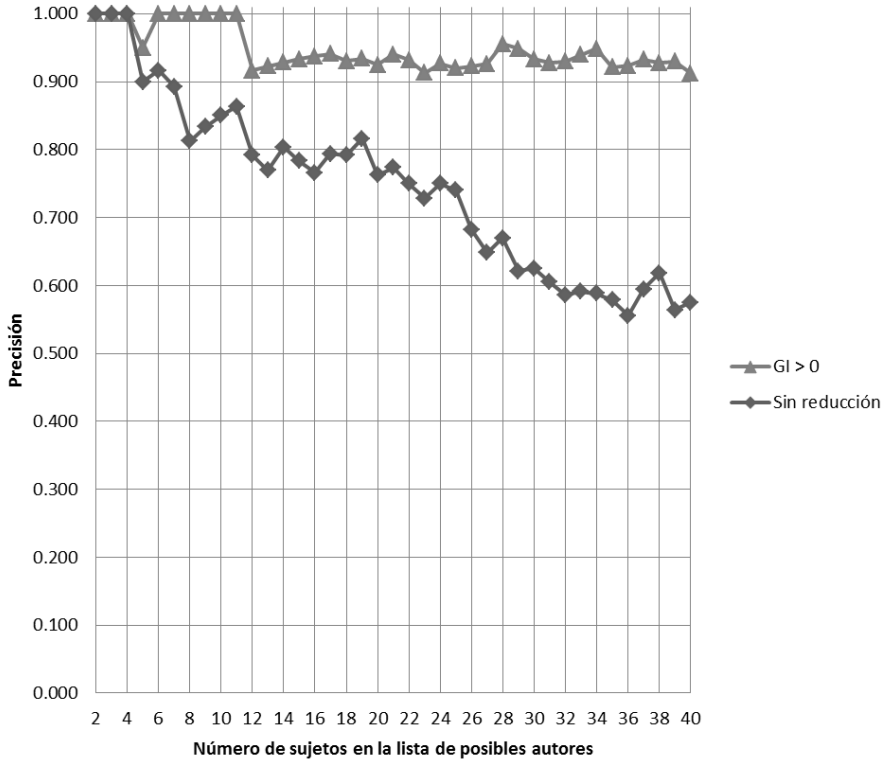

Fig. 1. Efecto de la reducción de rasgos por GI con el ABM

La línea superior en la Figura 1, con vértices resaltados con triángulos, corresponde a la precisión del ABM con la lista reducida con GI, mientras que la línea inferior, con vértices resaltados con rombos, describe la precisión del ABM al usar la lista no reducida. Estas dos líneas permiten apreciar el claro efecto positivo de la reducción de rasgos en la combinación más exitosa, en este estudio, de una técnica de reducción y un clasificador. Como se observa en la línea superior, la aplicación de la técnica de reducción de rasgos por GI le permite al clasificador del ABM mantener una precisión constante por encima de 0.900 a lo largo de los todos los experimentos con los 39 corpus, los cuales incluyen como muestra la figura desde dos hasta 40 autores.

Respecto de la segunda opción para manejar rasgos clasificatorios, la utilización de clasificadores resistentes al ruido, este trabajo ha evaluado un algoritmo de última generación en el aprendizaje automático, las SVMs. Este algoritmo es considerado como especialmente resistente al ruido en el contexto de la atribución autoría [3], [9]. En este estudio, las SVMs han mostrado por un lado que son en efecto altamente resistentes al ruido, ya que obtienen su mejor resultado promedio con las listas reducidas con el criterio de frecuencia, las cuales son comparativamente largas. Sin embargo, la combinación de un modelo tradicional de clasificador, el ABM, en combinación

con cualquiera de los tres métodos de reducción de rasgos (frecuencia, GI y CFS), tiene la capacidad de superar los mejores resultados obtenidos por las SVMs. La superioridad del ABM, en su mejor desempeño logrado con GI, por encima del mejor rendimiento de las SVMs, al combinarse con la lista reducida por frecuencia, puede apreciarse en la Figura 2, a continuación.

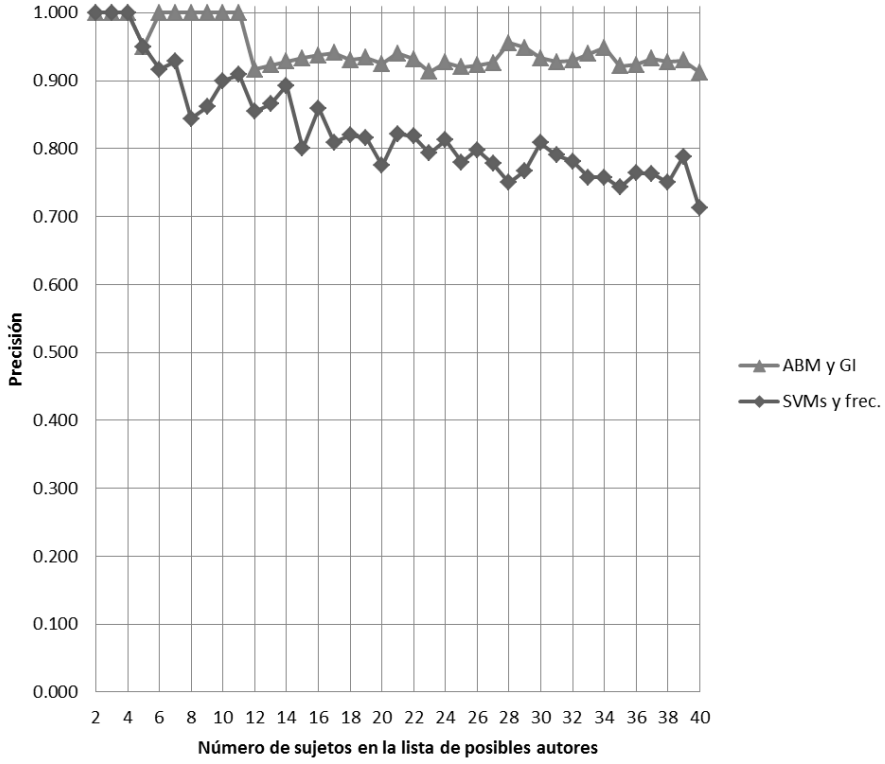

Fig. 2. Precisión promedio más alta obtenida por el ABM y las SVMs

La tercera opción para el manejo de listas largas de rasgos es la utilización simultánea de algoritmos altamente resistentes al ruido y técnicas de reducción. La conveniencia de utilizar o no esta opción se desprende de los últimos dos puntos discutidos. Por un lado, todos los clasificadores se beneficiaron de las técnicas de reducción de rasgos. Por otro lado, el algoritmo altamente resistente al ruido demostró que obtuvo el mayor beneficio al usar listas reducidas comparativamente largas. Sin embargo, ha sido la combinación de un clasificador que es no es considerado particularmente resistente al ruido y una lista notoriamente reducida (y depurada de aquellos rasgos que insertan ruido) la que ha obtenido los mejores resultados en la evaluación de las múl-

tiples combinaciones de clasificadores y técnicas de reducción. Por tanto, esta tercera opción ha sido superada, al igual que la segunda antes discutida.

Referencias

1. Rudman, J. The State of Authorship Attribution Studies: Some Problems and Solutions. Computers and the Humanities. 31, 351-365 (1998)
2. Koppel, M., Schler, J., & Argamon, S. Computational Methods in Authorship Attribution. Journal of the American Society for Information Science and Technology. 60(1), 9-26 (2009)
3. Abbasi, A., & Chen, H. Applying Authorship Analysis to Extremist-Group Web Forum Messages. IEEE Intelligent Systems. 20(5), 67-75 (2005)
4. Gamon, M. Linguistic Correlates of Style: Authorship Classification with Deep Linguistic Analysis Features. En actas de the 20th International Conference on Computational Linguistics: Vol.4, pp. 611-617. Stroudsburg, PA: Association for Computational Linguistics (2004)
5. Juola, P. Authoship Attribution. Hanover, MA: Now Publishers (2008)
6. Koppel, M., Schler, J., & Messeri, E. Authorship Attribution in Law Enforcement Scenarios. En C.S. Gal, P. Kantor, & B. Saphira (Eds.), Security Informatics and Terrorism: Patrolling the Web, pp.111-119. Amsterdam: IOS (2008)
7. Rico-Sulayes, A. Statistical Authorship Attribution of Mexican Drug Trafficking Online Forum Posts. International Journal of Speech, Language and the Law. 18(1), 53-74 (2011)
8. Manning, C. D., Raghavan, P., & Schütze, H. Introduction to Information Retrieval. New York, NY: Cambridge (2008)
9. Zheng, R., Li, J., Chen, H. & Huang, Z. A Framework for Authorship Identification of Online Messages: Writing-Style Features and Classification Techniques. Journal of the American Society for Information Science and Technology. 57(3): 378-393 (2006)
10. Rico-Sulayes, A. (2012). Quantitative Authorship Attribution of Users of Mexican Drug Dealing Related Online Forums. Tesis doctoral, Georgetown University.
11. Agar, M. Mexican Drug War Deaths Top 47,500. The Telegraph (2012, enero 12)
12. Córdoba, J., & Luhnow, D. In Mexico, Death Toll in Drug War Hits Record. The World Street Journal (2011, enero 13)
13. Davison, J., & Stastna, K. Mexico's cartels: Behind the drug war. CNN (2014, febrero 25)
14. Planas, R. A Murder Every Half Hour in Mexico's Drug War. NYDailyNews.com: Daily News (2012, enero 13)
15. Borderland Beat, http://www.borderlandbeat.com/2011/09/nuevo-laredo-silent-war.html
16. Goodman, J. D. In Mexico, Social Media Become a Battleground in the Drug War. The New York Times: The Lede (2011, septiembre 15)
17. Stevenson, M. Woman Decapitated in Mexico for Web Posting. Associated Press (2011, septiembre 24)
18. Blog del Narco, http://www.blogdelnarco.com/
19. Chaski, C. E. Who's At The Keyboard? Authorship Attribution in Digital Evidence Investigations. International Journal of Digital Evidence. 4(1), 1-13 (2005)
20. Baayen, H., van Halteren, H., Neijt, A., & Tweedie, F. An Experiment in Authorship Attribution, pp. 29-37. En actas de JADT 2002: Sixth International Conference on Textual Data Statistical Analysis (2002)

21. Spassova, M. S. El Potencial Discriminatorio de las Secuencias de Categorías Gramaticales en la Atribución Forense de Autoría de Textos en Español. Tesis doctoral, Universitat Pompeu Fabra, Barcelona (2009)
22. Stamatatos, E., Fakotakis, N., & Kokkinakis, G. Computer-Based Authorship Attribution without Lexical Measures. Computers and the Humanities. 35, 193-214 (2001)
23. Burrows, J. Delta: A Measure of Stylistic Difference and a Guide to Likely Authorship. Literary and Linguistic Computing. 17(3), 267-86 (2002)
24. Corney, M. Analysing E-mail Text Authorship for Forensic Purposes. Tesis de maestría, Queensland University of Technology (2003)
25. Orebaugh, A., & Allnutt, J. Classification of Instant Messaging: Communications for Forensics Analysis. The International Journal of Forensic Computer Science. 4(1), 22-28 (2009)
26. Tambouratzis, G., & Vassiliou, M. Employing Thematic Variables for Enhancing Classification Accuracy within Author Discrimination Experiments. Literary and Linguistic Computing. 22(2), 207-224 (2007)
27. Hall, M. A. Correlation-based Feature Selection for Machine Learning. Tesis de maestría, The University of Waikato (1999)