Análisis del desempeño del algoritmo genético en la clasificación automática de documentos

Juan Manuel Zárate Sánchez, Hilarión Muñoz Contreras,
María Antonieta Abud Figueroa

División de Estudios de Posgrado e Investigación, Instituto Tecnológico de Orizaba, México

manuel_zasa86@hotmail.com, hmunozc189@msn.com
mabud@ito-depi.edu.mx

Resumen. Existen diversos algoritmos de clasificación para el proceso de clasificación automática de documentos, estos algoritmos buscan soluciones eficientes y rápidas, es por lo cual, el Algoritmo Genético es ideal para aplicar a este contexto, ya que es capaz de encontrar soluciones eficientes en unos cuantos segundos, por su capacidad de examinar el espacio de búsqueda en forma amplia y eficiente. Este Artículo se centra principalmente en analizar el desempeño del Algoritmo Genético para la clasificación automática de documentos. En este artículo se analizó el desempeño que tiene el algoritmo OlexGA, que posee como base un Algoritmo Genético, mediante diversas pruebas de minería de datos.

Palabras clave: algoritmo genético, clasificación texto, algoritmo evolutivo, OlexGA, minería de datos, WEKA.

1. Introducción

La clasificación automática de documentos se define como el proceso de separar los documentos en grupos o clases. El criterio de agrupamiento es acuerdo en las similitudes que existen entre ellos [1]. La clasificación automática de documentos se relaciona con dos disciplinas Informáticas :1) la Minería de Datos y 2) la Recuperación de Información, la Primera disciplina utilizan las técnicas que se ocupan para la clasificación de objetos, por último, la segunda disciplina utilizan los conceptos y métodos para el pre-procesamiento de documentos, es decir, esto consiste en la transformación de la información no estructurada de los documentos a información estructurada manejable para los algoritmos de clasificación [2].

Este artículo está estructurado de la siguiente manera. En la sección 2 se analizó el estado del arte de la clasificación automática de documentos aplicando un enfoque evolutivo. En la sección 3 se proporciona una breve descripción del proceso de clasificación automática de documentos. La sección 4 aborda como se realiza las pruebas, en la sección 5 se muestran los resultados de las pruebas, en la sección 6 se presentan la discusión de los resultados y por último, la sección 7 se tiene las conclusiones.

2. Estado del arte

En el contexto de la Minería de Datos, el uso de Algoritmo Genético para clasificar documentos ofrece múltiples ventajas como ahorro de tiempo, dinero y esfuerzo. Por tal razón, diversos trabajos en la literatura han abordado este enfoque.

Yolis et al. [3] describen las restricciones que presentan los algoritmos de clasificación para categorizar los documentos, argumentando que los tiempos de cómputo son inaceptables. Según la perspectiva del autor, una solución para resolver esta problemática es la aplicación del Algoritmo Genético. Este trabajo propone un algoritmo genético que presenta un operador cruzado y cuatro operadores de mutación, diseñados especialmente para solucionar la problemática de la clasificación automática de documentos, el diseño del algoritmo de clasificación se adapta a los conceptos que aplican los algoritmos genéticos. Se realizaron experimentos cuyos resultados confirman que el algoritmo propuesto es una buena solución para la clasificación de documentos. Teniendo como conclusión que los algoritmos genéticos son una herramienta poderosa para la resolución de problema donde las soluciones son amplias y optimización son complejas, sin embargo, el autor afirma que los algoritmos genéticos no son un método de solución universal de problemas, sino un paradigma que debe adaptarse correctamente al problema a resolver.

K. Premalatha et al. [4] describen un proceso de clasificación de documentos basado en el Algoritmo Genético con un operador de mutación simultánea y un número de tasa de mutación. En la mutación simultánea, el algoritmo genético, utiliza varios operadores de mutación en la producción de la próxima generación. El operador de mutación es representativo para el éxito de los algoritmos genéticos, ya que amplían las direcciones de búsqueda y evita la tendencia a una optimización local. El algoritmo propuesto aumentó notablemente el éxito para la clasificación de documentos, para comprobar esto se comparó con el algoritmo genético simple con el algoritmo de K-mean.

Meena et al. [5] exponen que las técnicas de agrupación tienen mejor optimización aplicando el enfoque del Algoritmo Genético. El algoritmo k-mean es un método popular para la categorización de documentos de texto, pero sus resultados se basan en la elección de los centros de conglomerados, esto fácilmente se traduce en optimización local. En este trabajo se propuso en utilizar un método dinámico para la agrupación basado en el Algoritmo Genético, el objetivo es encontrar fácilmente mejores centros de conglomerados y reduciendo las iteraciones mediante con otra técnica de optimización DDE (*Discrete Differential Evolution*). Se obtuvo como conclusión que hay una mejora en la categorización de documentos utilizando la combinación del algoritmo genético y DDE.

A. Casillas et al. [6] presentan un algoritmo que realiza la categorización de una colección de documentos, sin conocer previamente el número de agrupación. En este trabajo se diseñó e implementó un algoritmo genético que encuentra un valor próximo al óptimo número de agrupación k, con un costo computacional menor que con la regla de Calinski y Harabasz, este algoritmo obtiene un valor aproximado del óptimo número de agrupación k y resuelve la forma de agruparlos dentro de las agrupaciones de k. El autor evaluó este algoritmo con conjuntos de documentos, que son el resultado de salida de un motor de búsqueda ante consultas de un usuario, por lo cual se afirma que el algoritmo es capaz de realizar una agrupación en línea.

A. Villagra et al. [7] detallan en forma breve una de las líneas de investigación que se realizó en el Laboratorio de Tecnologías Emergentes (LabTEm) sobre Algoritmos Evolutivos y su aplicación como técnica alternativa y complementaria en tareas de Minería de Datos, concretamente en la clasificación de objetos. En la actualidad, se almacenan enormes cantidades de datos y no es fácil representar e interpretar los datos que están almacenados, por consiguiente es difícil obtener información relevante para la toma de decisiones basadas en dichos datos. La Minería de Datos implica "escavar" en esa inmensidad de datos, en búsqueda de patrones, asociaciones o predicciones que permitan transformar esa maraña de datos en información útil. Esta línea de investigación tuvo como resultado técnicas avanzadas de minería de datos basadas en el enfoque evolutivo.

Veronica et al. [8] presentan un Algoritmo Genético con el nombre Olex-GA, para clasificación de texto basados en reglas de inducción de la siguiente forma: "clasificación de documentos d bajo la categoría c si $t_1 \in d \; o \; \ldots o \; t_n \in d$ y no contenga ($t_{n+1} \in d \; o \; \ldots o \; t_{n+m} \in d$)", donde cada t_1 es un término. Olex-GA se basa en una eficiente representación binaria de varias reglas por individuos y utiliza como función de aptitud F-medida. El algoritmo de clasificación propuesto fue probado sobre la prueba estándar de los conjuntos Reuters 21578 y Ohsumed, y se comparó con varios algoritmos de clasificación.

Como se observa en la revisión del estado del arte, la utilización de Algoritmo Genético en la minería de datos es una buena solución para la clasificación automática de documentos.

3. Proceso de clasificación automática de documentos

Para comprender el proceso de clasificación automática de documentos se debe analizar el proceso normal que hace un clasificador humano; este proceso se define de la siguiente forma:

"... el acto de organizar el universo del conocimiento en algún orden sistemático. Se considera la actividad más importante de la mente humana. El hecho de clasificar reside en el dicotómico proceso de diferenciar los objetos que tienen ciertas características de otros objetos que no tienen y agruparlos en una clase los objetos que tienen características comunes" [9].

Esto se traduce en términos de gestión documental, que la clasificación de documentos manual consiste en tres simples pasos:

- Leer el documento para tener una idea precisa de su contenido.
- Resumir el contenido en un tema principal
- Establecer qué categoría se acerca más al tema principal y representar el documento mediante una notación propia de la clasificación, de tal forma que el almacenamiento esté en orden y que la recuperación del documento sea posible.

El clasificador automático no tiene la capacidad de enfrentar en una forma directa algunas tareas de la clasificación manual, por lo cual estas tareas son realizadas con diferentes métodos que lleven a un resultado similar. Por ejemplo, la tarea de leer el documento e identificar su contenido, el sistema de clasificador automático suele llevar a cabo este proceso mediante la utilización de métodos de indización automática. Es

evidente que el proceso de indización automática no tiene nada que ver con el proceso que realiza un clasificador humano, ya que este tiene una lectura secuencial, además, el sistema de clasificación automática no entiende o comprenden los contenidos de los documentos, pero lo representa en una forma que simule el entendimiento.

En la tarea de resumir el contenido en un tema principal, es un proceso fácil para un clasificador humano, pero para el sistema de clasificación automática es lógico pensar que no pueda asignar un tema principal a un contenido de un documento, como lo haría un clasificador humano, ya que el sistema de clasificación automático no comprende el contenido y desconoce el tema principal o su significado.

Por último, el proceso de establecer la clasificación de un documento de acuerdo con su tema principal es una tarea relativamente sencilla para un clasificador humano, ya que este comprobaría la pertenencia del tema principal del documento a cada una de las categorías principales de la clasificación hasta encontrar la correcta y después descender por las subcategorías hasta localizar la que más se adapte al tema principal del documento. Un clasificador automático no podrá llevar a cabo este proceso de forma directa, ya que no conoce el significado del tema principal del documento.

Por lo anterior descrito, es evidente que el trabajo de un clasificador humano no se logra llevar acabo directamente por un clasificador automático, además, la parte inteligente del sistema de clasificación automático está enfocado en la aplicación de un algoritmo que permite el aprendizaje.

Los algoritmos de clasificación asignan a los documentos una o varias etiquetas, categorías o clases prestablecidas, sin embargo, no tiene como objetivo emular el comportamiento de un clasificador humano, sino aprovecharse de él, para lograr las mismas conclusiones con diferentes métodos. Para lograr aprovecharse del clasificador humano es necesario utilizar un conjunto de documentos de entrenamiento. Este conjunto de documento de entrenamiento ha sido preclasificado por clasificadores humanos expertos en la materia.

El proceso de clasificación automática de documento suele contener tres etapas generales [10] :

1. Pre-procesamiento,
2. Construcción de Clasificador,
3. Clasificación de nuevos documentos.

A continuación se describirán las etapas de la categorización automática de documentos.

3.1.Preprocesamiento

En esta etapa de proceso de clasificación automática de documentos, los textos de los documentos se trasfiere a un formato compacto que se utiliza en las siguientes etapas. En este proceso incluye los siguientes métodos:

Stemming: Es un método reduce una palabra a su raíz, por ejemplo, tenemos una consulta sobre “bibliotecas” que también saldrán documentos que tenga la palabra “bibliotecario”, porque la raíz de las dos palabra es “bibliotec”.

Stopwords: es una técnica que elimina palabras de función, por ejemplo: preposiciones, artículos, conjunciones y otras palabras dependiente del dominio. Esta eliminación es acuerdo a un diccionario prestablecido y se aplica después del método de *Stemming*

Elección de atributos: cada palabra del texto será un atributo, esto solo tendrá dos valores 1 y 0, las palabras que sean irrelevante serán eliminadas, por ejemplo, las palabras que se repitan en todos los documentos.

Con estos métodos se construye la representación compacta de los documentos para los algoritmos de clasificación.

3.2.Construcción de clasificador

La gran parte de los algoritmos de clasificación llevan a cabo un tipo de entrenamiento o aprendizaje antes de realizar la tarea de clasificar nuevos documentos. Para realizar este tipo de aprendizaje se debe contar con un conjunto de documentos de entrenamiento ya clasificados por expertos. Los conjunto de documentos de entrenamiento son los documentos convertidos en conjuntos de atributos y valores por la etapa anterior, por lo cual, se utilizan para construir un clasificador que asignará categorías a nuevos documentos.

3.3.Clasificación

Este el último proceso de la clasificación automática de documentos, es la más rápida que los anteriores procesos, ya que todo el trabajo del proceso de clasificación automática de documentos recae en la etapas de pre-procesamiento y construcción del clasificador. Para clasificar un conjunto de documentos nuevos, previamente pre-procesado, se utiliza el clasificador construido en la anterior etapa.

4. Análisis del desempeño del clasificador basado en algoritmo genético

Una vez que se comprendió el proceso de Clasificación Automática de Documentos se procedió a la evaluación del desempeño del Algoritmo Genético para la clasificación de documentos. Cabe mencionar que este trabajo no tiene como objetivo desarrollar un sistema de clasificación de documentos, sino realizar un análisis del desempeño del Algoritmo Genético para la tarea de clasificación de texto y comparando con otros algoritmos de clasificación, por lo cual, se utilizó API (Application Programming Interface) de WEKA.

Los criterios para medir el desempeño del algoritmo genético giran en dos criterios: 1) El criterio de eficiencia hace referencia a las medidas que están relacionados con los tiempos computacionales requeridos para llevar a cabo los procesos construcción del clasificador y la evaluación del clasificador. 2) El criterio de efectividad se refiere a nuestra capacidad para medir el comportamiento que tiene el clasificador desde un punto de vista de la calidad de resultados obtenidos por él mismo. Para medir el desempeño de los algoritmos de clasificación se emplearon 3 tipos de prueba:

Percentage Split: En este caso, se dividirá el conjunto de entrenamiento en dos partes: los primeros 66% de los datos para construir el clasificador y el 33% finales, para hacer la evolución.

Cross-validation: Este tipo de prueba se realiza una validación cruzada estratificada del número de particiones dado (*Folds*). La validación cruzada consiste en: dado un número *n* se divide los datos en *n* partes y, por cada parte, se construye el clasificador con las *n-1* partes restantes y se prueba con esa partición.

Supplied test set: Este tipo de evaluación se construye el clasificador con un conjunto de documentos de entrenamiento y el clasificador se evalúa con otro conjunto de documentos de prueba para calcular sus aciertos.

4.1.Conjunto de documentos

En este trabajo se usó el conjunto de documentos denominado *Reuters 21758,* es una colección de documentos de noticias reales que aparecieron en cables de la agencia Reuters durante 1987. Los documentos fueron recopilados y categorizados manualmente por personal de la agencia y de la compañía Carnegie Group [11]. La colección *Reuters 215*78 es un estándar para la prueba de algoritmos de clasificación, por lo que los resultados obtenidos tendrán mucha más validez que si los experimentos se realizaran sobre un conjunto de datos recopilados sin seguir una metodología estándar. Además se utilizó en diversos trabajos [4], [12], [13], [14], [15], [16] y [17]. Antes de usar la colección de Reuters 21578 se tiene que realizar un pre-procesamiento para obtener una representación compacta de los documentos, en este trabajo se utilizó la representación compacta de un investigador del El Instituto de Teoría de la Información y Automatización de la Republica Checa, que lo puso a disposición en [18].

Esta representación está dividida en dos conjuntos: el primer conjunto de documentos de entrenamiento, que contienen 7769 documentos y el segundo es de conjunto de documentos de prueba (*test*), que contienen 3018 documentos, además, sólo se seleccionó diez clases que contienen la mayor cantidad de documentos para la prueba, estas clases se observan en la tabla 1.

Tabla 1. Las diez clases o clasificación con mayor cantidad de documentos de prueba

Clase	Número de documentos de esta clase en el conjunto de pruebas
EARN	1087
ACQ	719
CRUDE	189
MONEY-FX	179
GRAIN	149
INTEREST	131
TRADE	117
SHIP	89
WHEAT	71
CORN	56

Todos los conjuntos documentos entrenamiento y de prueba de la colección de Reuters se convirtió al formato ARFF, que es un formato de WEKA, para construir los diferentes clasificadores.

4.2.Algoritmos de clasificación

Existen diversos algoritmos de clasificación empleados para la clasificación de documentos, sin embargo, no son específicamente para este fin sino que se han propuesto para clasificar de todo tipos de objetos, es decir, algunos de esto se adaptan con más o menos para la clasificación de documentos. A continuación se muestran los algoritmos de clasificación empleados en los experimentos, cabe aclarar que estos algoritmos son los que han sido utilizados para la clasificación de documentos y están implementados en WEKA: OlexGA, Ridor, JRip, J48, ADTree, IBK, NaiveBayesMultinomial y LibSVM.

El Algoritmo OlexGA es un algoritmo de clasificación basado en el Algoritmo Genético, la descripción de este algoritmo se encuentra en [8], el cual se va analizar su desempeño por ser un algoritmo con enfoque evolutivo.

5. Resultados

En esta sección se presentan los resultados obtenidos de la realización de las pruebas descritas anteriormente, cada prueba se repitió 10 veces para sacar un promedio de cada algoritmo de clasificación.

En la prueba de *Percentage Split* se obtuvo los datos que se muestran en las tablas 2 y 3, con estos resultados se llega a la conclusión que el algoritmo OlexGA tiene en promedio de precisión de 96.14, sin embargo, algunos algoritmos tienen un promedio de precisión mayor que OlexGA, por ejemplo, JRIP Y NaiveBayesMultinomial, pero su promedio de tiempo es mayor que el OlexGA , por lo cual , este algoritmo de clasificación con enfoque evolutivo tiene un buen desempeño al dar resultados aceptables en pocos segundos (ver Tabla 3) .

Tabla 2. Precisión obtenida de la prueba de *Percentage Split*.

Clases	*OlexGA*	*Ridor*	*JRip*	*J48*	*ADTree*	*IBk*	*NaiveBayesM.*	*LibSVM*
ACQ	**86.51**	92.33	93.64	93.76	90.07	93.66	96.54	94.71
CORN	**99.53**	99.38	99.56	99.41	99.45	98.33	98.29	95.28
CRUDE	**97.38**	97.27	97.86	97.76	96.98	97.36	97.5	96.49
EARN	**95.53**	95.69	95.85	95.51	94.11	95.59	95.66	96.14
GRAIN	**98.89**	98.46	98.87	98.54	98.55	96.74	97.01	96.56
INTEREST	**96.59**	96.8	97.28	96.88	96.35	97.32	97.37	96.08
MONEY	**96.84**	96.21	97.13	96.57	95.63	97.06	96.63	95.8
SHIP	**98.81**	98.04	98.56	98.42	98.34	98.36	98.72	96.83
TRADE	**92**	96.61	97.12	96.95	96.56	96.61	96.38	96.15
WHEAT	**99.33**	99.38	99.47	99.33	99.44	97.79	97.97	94.91
PROMEDIO	**96.14**	**97.01**	**97.53**	**97.31**	**96.54**	**96.88**	**97.20**	**95.89**

Tabla 3. Tiempo obtenido para la construcción del clasificador y su evaluación en la prueba de *Percentage Split.*

Clases	***OlexGA***	*Ridor*	*JRip*	*J48*	*ADTree*	*IBk*	*NaiveBayesM*	*LibSVM*
ACQ	**4.928**	30.89	64.11	29.1	47.01	40.63	4.126	6.912
CORN	**5.457**	6.969	10.6	4.486	41.82	70.44	7.117	2.696
CRUDE	**4.715**	7.149	24.96	11.67	49.43	60.59	6.121	3.015
EARN	**4.535**	28.02	30.9	16.7	45.97	50.07	5.056	6.82
GRAIN	**4.617**	7.839	19.51	12.71	51.46	52.82	5.341	3.196
INTEREST	**5.748**	8.53	31.72	16.21	48.42	47.33	4.795	3.467
MONEY	**4.714**	9.27	28.8	20.78	46	55.5	5.619	4.494

Clases	***OlexGA***	*Ridor*	*JRip*	*J48*	*ADTree*	*IBk*	*NaiveBayesM*	*LibSVM*
SHIP	**5.298**	10.15	18.94	15.4	51.55	45.33	4.59	1.743
TRADE	**5.149**	8.27	21.95	9.387	45.86	56.63	5.723	3.477
WHEAT	**6.481**	5.475	8.764	5.754	37.49	58.4	5.918	2.808
PROMEDIO	**5.1642**	**12.25**	**26.02**	**14.21**	**46.50**	**53.77**	**5.44**	**3.86**

En la prueba de *Cross-validation* se obtuvo los siguientes resultados que se muestran en las tablas 4 y 5, con estos datos se llega a la conclusión que el algoritmo OlexGA tiene un promedio de precisión aceptable en un menor tiempo que los otros algoritmos de clasificación analizados.

Tabla 4. Precisión obtenida de la prueba de *Cross-validation*.

Clases	*OlexGA*	*Ridor*	*JRip*	*J48*	*ADTree*	*IBk*	*NaiveBayesM.*	*LibSVM*
ACQ	**86.7**	92.45	93.65	93.85	90.27	93.63	96.46	94.75
CORN	**99.6**	99.44	99.55	99.48	99.51	98.36	98.25	95.39
CRUDE	**97.56**	97.22	97.79	97.62	96.95	97.23	97.52	96.53
EARN	**95.5**	95.28	95.83	95.46	94.07	95.76	95.85	96.22
GRAIN	**98.89**	98.51	98.88	98.52	98.64	96.69	96.95	96.51
INTEREST	**96.59**	96.81	97.21	96.8	96.41	97.35	97.38	95.96
MONEY	**96.78**	96.29	97.04	96.47	95.6	96.97	96.53	95.88
SHIP	**98.77**	98.09	98.58	98.45	98.33	98.34	98.75	97.58
TRADE	**92.31**	96.76	97.24	97.07	96.6	96.63	96.34	96.34
WHEAT	**99.41**	99.4	99.45	99.38	99.41	97.91	97.99	95.43
PROMEDIO	**96.211**	**97.02**	**97.52**	**97.31**	**96.579**	**96.88**	**97.20**	**96.05**

Tabla 5. Tiempo obtenido para la construcción del clasificador y su evaluación en la prueba de *Cross-validation*

Clases	***OlexGA***	*Ridor*	*JRip*	*J48*	*ADTree*	*IBk*	*NaiveBayesM.*	*LibSVM*
ACQ	**19.96**	202.5	409	239.4	353.3	151.4	15.38	31.55
CORN	**23.18**	28.14	42.92	18.85	158.6	176.7	17.91	8.089
CRUDE	**21.03**	55.8	158.1	76.82	252.6	175.7	17.8	11.16
EARN	**24.21**	95.28	95.83	95.46	94.07	95.76	95.85	96.22
GRAIN	**24.01**	78.41	178.5	111.9	320.6	159.3	16.14	12.58
INTEREST	**22.77**	59.95	176.9	102.8	248.2	156.6	15.89	13.5
MONEY	**18.73**	60.89	179.2	123.6	228.3	179.6	18.19	18.82
SHIP	**19.05**	71.03	130.1	104.2	273.9	153.1	15.53	6.886
TRADE	**25.4**	62.26	195.5	84.96	234.2	182.2	18.43	14.09
WHEAT	**18.14**	24.14	43.7	24.41	127	168.4	17.06	8.334
PROMEDIO	**21.648**	**73.84**	**160.97**	**98.24**	**229.07**	**159.87**	**24.81**	**22.12**

En la prueba de *Supplied test set* se obtuvo los siguientes datos que se muestran en las tablas 6 y 7. En la tabla 7 se observa que el algoritmo OlexGA tiene un promedio de tiempo de 5 segundos, es evidente que el algoritmo de clasificación con enfoque evolutivo tiene un mejor tiempo que los demás algoritmos, además, en el promedio de precisión tiene un resultado aceptable.

Tabla 6. Precisión obtenida de la prueba de *Supplied test set*.

Clases	*OlexGA*	*Ridor*	*JRip*	*J48*	*ADTree*	*IBk*	*NaiveBayesM.*	*LibSVM*
ACQ	**86.4**	92.93	94.61	94.7	89.96	93.64	97.21	94.72
CORN	**99.49**	99.65	99.68	99.71	99.73	98.62	98.36	96.32
CRUDE	**97.76**	97	97.62	97.42	97.46	95.67	97.24	95.95
EARN	**97.33**	96.57	96.99	96.61	96.34	97.38	97.68	98.05
GRAIN	**99.25**	98.45	99.11	98.73	98.92	97	97.18	97.54
INTEREST	**96.38**	96.68	97.25	96.15	96.66	96.68	97.25	96.15
MONEY	**96.31**	95.84	96.46	95.93	95.11	95.39	96.6	95.38

Clases	*OlexGA*	*Ridor*	*JRip*	*J48*	*ADTree*	*IBk*	*NaiveBayesM.*	*LibSVM*
SHIP	**98.61**	98.14	98.81	98.58	98.89	96.6	98.71	97.41
TRADE	**93.65**	97.44	97.73	97.5	97.57	96.34	96.91	96.95
WHEAT	**99.53**	99.53	99.57	99.39	99.52	98.29	98.21	96.55
PROMEDIO	**96.47**	**97.22**	**97.78**	**97.47**	**97.01**	**96.56**	**97.53**	**96.50**

Tabla 7. Tiempo obtenido para la construcción del clasificador y su evaluación en la prueba de *Supplied test set.*

Clase	***OlexGA***	*Ridor*	*JRip*	*J48*	*ADTree*	*IBk*	*NaiveBayesM.*	*LibSVM*
ACQ	**6.042**	68.55	125	57.55	85.39	62.87	6.37	12.08
CORN	**4.603**	9.733	15.18	6.009	62.04	77.52	7.828	3.617
CRUDE	**4.502**	19.92	61.93	30.13	90	81.28	8.204	5.008
EARN	**4.753**	47.7	60.45	32.81	81.07	77.07	7.799	16.13
GRAIN	**5.729**	30.81	70.07	45.09	134.2	83.55	8.443	7.764
INTEREST	**5.903**	25.75	84.1	46.82	84.63	64.73	6.546	6.001
MONEY	**4.684**	26.57	77.18	48.86	82.64	74.32	7.507	7.806
SHIP	**4.557**	23.73	55.56	40.76	95.39	59.73	6.047	2.943
TRADE	**4.876**	19.25	49.8	23.88	79.34	73.35	7.408	6.339
WHEAT	**4.401**	9.17	16.98	10.05	39.34	74.64	7.538	3.807
PROMEDIO	**5**	**28.11**	**61.62**	**34.19**	**83.404**	**72.90**	**7.36**	**7.14**

6. Discusión de los resultados

Este trabajo tuvo la finalidad de analizar el desempeño que tiene el Algoritmo Genético para el problema general de la clasificación automática de documentos aplicando las técnicas de minería de datos que ofrece la herramienta de WEKA. Los resultados obtenidos de las tres pruebas, se puede deducir que se obtienen resultados aceptables, como se muestra en la figura 1.

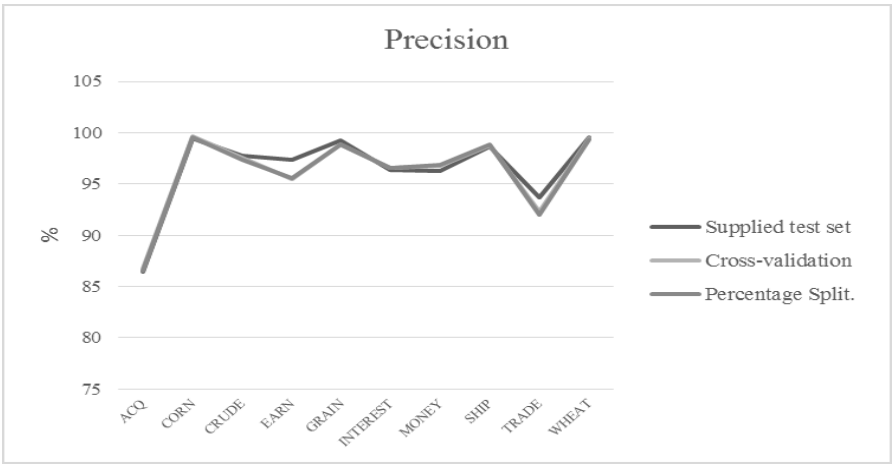

Fig. 1. Grafica de los resultados de presión en las tres pruebas.

Además se consigue estos resultados en un tiempo menor que la mayoría de los algoritmos de clasificación analizados, como se muestra en la figura 2, por lo cual, se llega a la conclusión que la aplicación de los Algoritmos Genético en la clasificación de automática de documentos se consiguen soluciones aceptables en un menor tiempo.

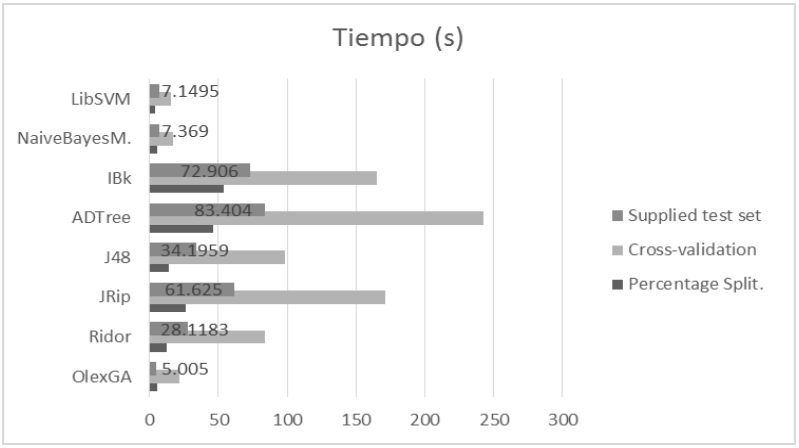

Fig. 1. Tiempo promedio de los algoritmos de clasificación en las tres pruebas.

7. Conclusión

Los algoritmos de clasificación de documentos están en creciente atención debido al aumento de la cantidad de información electrónica, y a la necesidad de accederla en el menor tiempo posible y con mayor eficacia. Si bien existen algoritmos que dan resultados aceptables en la clasificación de documentos, los tiempos de cómputo que requieren son inaceptables para las aplicaciones prácticas.

En las pruebas realizadas para este trabajo se confirma que los algoritmos genéticos son una poderosa herramienta para la solucionar el problema general de la clasificación automática de documentos, dando como resultado soluciones aceptables en un mejor tiempo.

Referencias

[1] Kaufman, L., Rousseeuw; J.: Finding groups in data: an introduction to cluster analysis. In: Applied Probability and Statistics, New York, Wiley Series in Probability and Mathematical Statistics (1990)

[2] Yolis, E.: Algoritmos genéticos aplicados a la categorización automática de documentos. Facultad de ingeniería universidad de Buenos Aires, Buenos Aires, (2003)

[3] Yolis, E., Britos, P., Sicre, J., Servetto, A., García-Martínez, R., Perichinsky. G.: Algoritmos genéticos aplicados a la categorización automática de documentos. In: *IX* Congreso Argentino de Ciencias de la Computación, pp. 1468–1479 (2003)

[4] Premalatha, A.M., Natarajan, K.: Genetic Algorithm for Document Clustering based on Simultaneous and Ranked Mutation. Journal of Modern Applied Science, vol. 3, no. 2, pp. 35–42 (2009)

[5] Meena, K., Singh, P.: Text Documents Clustering using Genetic Algorithm and Discrete Differential Evolution. International Journal of Computer Applications, vol. 43, no. 1 (2012)

[6] Casillas, A., González de Lena M., Mart, R.: Document Clustering into an unknown number of clusters using a Genetic Algorithm. In: Speech and Dialogue: 6th International Conference, vol. 6, pp. 43–49 (2003)

[7] Villagra, A., Pandolfi, D., Lasso, M., de San Pedro, M.: Algoritmos evolutivos y su aplicabilidad en la tarea de clasificación. In: VIII Workshop de Investigadores en Ciencias de la Computación (2006)

[8] Pietramala, A., Policicchio, V. L., Rullo, P., & Sidhu, I.: Genetic algorithm for text classification rule induction. In: Machine Learning and Knowledge Discovery in Databases, pp. 188–203 (2008)

[9] Chan, M.L.: Cataloging and classification: an introduction. New York: McGraw-Hill, p. 209 (1981)

[10] Abelleira, M., Pérez, A.: Minería de texto para la categorización automática de documentos. PhD in Computer Science por Carnegie Mellon University, Madrid, España (2010)

[11] U. of California, UCI Knowledge Discovery in Databases Archive, University of California, Irvine, 9 Septiembre 2005. [En línea]. Available: http://kdd.ics.uci.edu/databases/reuters21578/reuters21578.html. [Último acceso: 2014 Septiembre 01].

[12] Jiménez, R. S.: La documentación en el proceso de evaluación de Sistemas de Clasificación Automática. Documentación de las Ciencias de la Información, vol. 30, pp. 25–44 (2007)

[13] Goldenberg, D.: Categorización Automática de Documentos con Mapas Auto Organizados de Kohonen. Tesis de Magister en Ingeniería del Software, Universidad Politécnica de Madrid, España, (2007)

[14] Toutanova, K., Chen, F. R., Popat, K., Hofmann, T.: Text classification in a hierarchical mixture model for small training sets. In: Proceedings of the tenth international conference on Information and knowledge management, ACM, pp. 105–113 (2001)

[15] Tikk, D., Dong-Yang J., Lee-Bang, S.: Hierarchical text categorization using fuzzy relational thesaurus. Kybernetika, vol. 30, pp. 583–600 (2003)

[16] Sciarrone, F.: An extension of the q diversity metric for information processing in multiple classifier systems: a field evaluation. International Journal of Wavelets, Multiresolution and Information Processing, vol. 11, no. 6 (2013)

[17] Peleja, F., Pereira-Lopes, G., Silva, J.: Text Categorization: A comparison of classifiers, feature selection metrics and document representation. In: Proceedings of the 15th Portuguese Conference in Artificial Intelligence, pp. 660–674 (2011)

[18] Karciauskas, G.: Reuter's data preprocessed by Gytis Karciauskas, The Institute of Information Theory and Automation, 2014. [En línea]. Available: http://staff.utia.cas.cz/vomlel/reuters-data.html. [Último acceso: 02 Octubre 2014].