Entre Ciencia e Ingeniería, vol. 14, no. 28, julio-diciembre de 2020, páginas 40-45. DOI: https://doi.org/ 10.31908/19098367.2014.
ISSN 1909-8367 (Impreso), ISSN 2539-4169 (En línea)

Algoritmos Evolutivos Multiobjetivo aplicados a la Selección de Características en Microarrays de Datos de Cáncer[1]

Multiobjective Evolutionary Algorithms applied to Feature Selection in Microarrays Cancer Data

J. S. Dussaut, I. Ponzoni, A. C. Olivera y P. J. Vidal

Recibido: octubre 20 de 2020 – Aceptado: diciembre 27 de 2020.

***Resumen*—El análisis de microarrays de expresión de genes es un tópico actual para el diagnóstico y clasificación del cáncer humano. Un microarray de datos de expresión de genes consiste en una matriz de miles de características de las cuales la mayoría es irrelevante para clasificar patrones de expresiones de genes. La elección de un subconjunto mínimo de características para clasificación es una tarea dificultosa. En este trabajo, se realiza una comparación entre dos algoritmos evolutivos multiobjetivo aplicados a conjuntos de expresiones de genes populares en la literatura (linfoma, leucemia y colon). Con el objetivo de remover las características con fuerte correlación se realiza una etapa de preprocesamiento. Se muestra un análisis extenso y detallado de los resultados obtenidos para los algoritmos multiobjetivo seleccionados.**

***Palabras clave*— Algoritmos Evolutivos Multiobjetivo, Expresión de genes, Microarrays de Cáncer, Selección de características.**

***Abstract*— Microarray analysis of gene expression is a current topic for diagnosing and classification of human cancer. A gene expression data microarray consists of an array of thousands of features of which most are irrelevant for classifying patterns of gene expressions. Choosing a minimal subset of features for classification is a difficult task. In this work, a comparison is made between two multi-objective evolutionary algorithms applied to sets of gene expressions popular in the literature (lymphoma, leukemia, and colon). In order to remove the strongly correlated characteristics, a pre-processing stage is performed. An extensive and detailed analysis of the results obtained for the selected multi-objective algorithms is shown.**

***Keywords*— Cancer Microarrays, Feature Selection, Gene Expression, Multiobjective Evolutionary Algorithms.**

I. Introducción

Actualmente los avances tecnológicos han permitido el procesamiento de todo tipo de datos. En particular, la tecnología de microarrays ha logrado el estudio de diferentes tipos de enfermedades, en particular cáncer. Un microarray (o chip de ácido ribonucleico, chip de ARN) consiste en una matriz de dos dimensiones con información de grandes cantidades de material biológico. Esta tecnología se ha vuelto muy popular entre los investigadores para el análisis de genes y su clasificación en expresiones génicas diferenciales [1], lo que permite descubrir posibles objetivos farmacológicos y biomarcadores.

Teniendo en cuenta que muchos de los genes de los microarrays, los cuales de aquí en más llamaremos características, son redundantes o irrelevantes para la enfermedad estudiada, resulta necesaria una selección de estas características. La selección de características (*Feature Selection*, FS) es un problema de optimización combinatorial del tipo NP-Difícil que consiste en seleccionar un subconjunto de características *d* de un conjunto de características *n* para utilizar en la definición de un modelo matemático [2].

Existe múltiples enfoques para este problema en la literatura. En particular, varias heurísticas han sido utilizadas para abordarlo [3]-[10], como así también algoritmos evolutivos [11]. Los algoritmos evolutivos son naturalmente aplicables a la resolución de la selección de características dado su mecanismo para producir múltiples soluciones en una

[1] Producto derivado de los proyectos de investigación Proyecto Tipo 1 B081 y proyecto PIP es 112-2017-0100829.

J. S. Dussaut, Universidad Nacional del Sur, Bahía Blanca, Argentina, e-mail: jsd@cs.uns.edu.ar.

I. Ponzoni, CONICET, Universidad Nacional del Sur, Bahía Blanca, Argentina, e-mail: ip@cs.uns.edu.ar.

A. C. Olivera, CONICET, Universidad Nacional de Cuyo, Mendoza, Argentina, e-mail: acolivera@conicet.gov.ar.

P. J. Vidal, CONICET, Universidad Nacional de Cuyo, Mendoza, Argentina, e-mail: pjvidal@conicet.gov.ar.

Como citar este artículo: Dussaut, J. S., Ponzoni, I., Olivera, A. C. y Vidal, P. J. Algoritmos Evolutivos Multiobjetivo aplicados a la Selección de Características en Microarrays de Datos de Cáncer, Entre Ciencia e Ingeniería, vol. 14, no.28, pp. 40-45, julio-diciembre, 2020.
DOI: https://doi.org/ 10.31908/19098367.2014.

sola ejecución, lo que permite encontrar un conjunto de soluciones no dominadas con el compromiso entre el número de características seleccionadas y el rendimiento de la clasificación [11]. Más aún, la utilización de Algoritmos Evolutivos Multiobjetivo (AEMOs) son un campo activo de investigación [12]-[14].

En el presente trabajo, un exhaustivo análisis entre dos algoritmos multiobjetivo para selección de características es realizado. El objetivo es minimizar el número de características sin perder precisión en la clasificación. Teniendo en cuenta que la minimización de características y la maximización de la precisión de la clasificación son objetivos contrapuestos el enfoque multiobjetivo puede ser utilizado. Para los experimentos se utilizan conocidas bases de datos de expresión de genes: colon [15], linfoma [16] y leucemia [17]. Con el objetivo de reducir características fuertemente correlacionadas, una etapa de preprocesamiento es realizada antes de aplicar los algoritmos. La selección de los algoritmos a utilizar para la comparación se realizó teniendo en cuenta no solo selección de características sino problemas de decisión binarios. Los algoritmos utilizados son: *Non-dominated Sorting Genetic Algorithm, version II* (NSGA-II) y *Multiobjective Cross-generational elitist selection, Heterogeneous recombination, y Cataclysmic mutation* (MOCHC). Los resultados muestran que ambos algoritmos evolutivos obtienen resultados prometedores para el FS, pero en particular, el algoritmo diseñado para problemas binarios (MOCHC) es el que obtiene los mejores resultados con respecto al NSGA-II considerando métricas populares en la literatura.

II. Tecnología de microarrays y Expresión de Genes

Un *gen* es un fragmento de ADN (Ácido desoxirribonucleico) que contiene toda la información requerida para producir una proteína en un organismo vivo [18]. Diferentes variedades de células expresan diversos subconjuntos de sus genes. En este contexto, los experimentos con microarrays de expresión permiten la captura de niveles de expresión en miles de genes simultáneamente [19]. Estos experimentos consisten principalmente en la observación de cada gen varias veces bajo diferentes condiciones o, alternativamente, enfocando cada gen en un entorno, pero con diferentes tipos de tejidos como, por ejemplo, tejidos de cáncer.

La clasificación de datos de microarrays es un procedimiento supervisado de aprendizaje que infiere el diagnóstico a partir de una muestra de expresión fenotípica [20]. Matemáticamente, dado un conjunto de genes y su correspondiente etiqueta de clases de cada muestra, el objetivo es descubrir cómo discriminar entre las diferentes clases utilizando los datos de expresión de genes de manera tal que, cuando se tiene una nueva muestra de microarray de datos, pueda ser extraída la clase correspondiente. El modelo de clasificación se construye analizando las muestras descriptas en el conjunto de genes denominadas características. Se supone que cada muestra pertenece a una clase predefinida llamada etiqueta de clase. En este trabajo, abordamos un problema de clasificación de dos clases para los datos de expresión génica. Las características contienen coeficientes de expresiones de genes y experimentos de microarrays que corresponden a muestras de pacientes. El problema se encuentra en desarrollar un clasificador adecuado para el diagnóstico genético utilizando muestras de entrenamiento disponibles de pacientes sanos y con cáncer. El principal problema es cómo seleccionar estos clasificadores dado que la mayoría de los genes incluidos en las muestras son irrelevantes o redundantes para la discriminación [20], [21].

En este trabajo nuestro principal objetivo es encontrar un conjunto mínimo de características sin detrimento en la precisión de la clasificación. Hay que tener en cuenta que no siempre son objetivos contrapuestos, es decir, si características redundantes o irrelevantes son quitadas del conjunto seleccionado, la precisión es mejorada, en la mayoría de los casos una concesión se debe realizar. Se recomienda el conjunto mínimo de características para lograr clasificadores con buenas propiedades de interpretación y generalización. De esto surge, que los algoritmos en este trabajo utilizarán dos funciones objetivo:

- $F_1 = Número\ de\ características\ seleccionadas$
- $F_2 = Precisión\ de\ la\ clasificación\ del\ cromosoma$

F_1 es minimizada mientras que F_2 es maximizada, se busca encontrar soluciones con un buen compromiso entre ambos objetivos.

III. Optimización Multiobjetivo para Selección de Características

En esta sección, se describe la estrategia de optimización propuesta utilizando algoritmos de optimización multiobjetivo y el algoritmo de *k*-vecinos más próximos (*k*-NN, *k*-Nearest Neighbors). Los algoritmos evolutivos multiobjetivo [23] (AEMOs) son algoritmos evolutivos concebidos con el propósito de resolver problemas con objetivos en conflicto. AEMOs han obtenido resultados precisos al resolver problemas del mundo real en diferentes áreas de investigación. A diferencia de los métodos tradicionales para optimización multiobjetivo los AEMOs permiten encontrar un conjunto de soluciones en una sola ejecución utilizando el concepto de Pareto.

En este trabajo, proponemos comparar dos algoritmos evolutivos multiobjetivo que han sido aplicados para problemas de optimización en diferentes áreas en la resolución del problema de selección de características de microarrays de datos de cáncer. Tres conjuntos de datos de diferentes tipos de cáncer son utilizados para el testeo de los AEMOs. Una etapa de preprocesamiento de las características (features) es realizada con el objetivo de remover características fuertemente correlacionadas. Esto reduce los conjuntos de datos en dos: conjunto de entrenamiento y conjunto de testeo. Con el objetivo de obtener la precisión (F_2) de cada solución, un algoritmo *k*-vecinos más próximos (*k*-NN) es utilizado para la clasificación [24]. El esquema general de la metodología utilizada se puede observar en la Fig.1.

A. *Etapa de Pre-procesamiento.*

El pre-procesamiento de los datos (conjuntos Colon,

leucemia y Linfoma) consiste en dos fases. En la primera, se realiza una normalización de cada característica c_i para cada muestra x_i aplicando la ecuación (1).

$$c'j(x_i) = \frac{(c_j(x_i) - mín_j)}{(máx_j - mín_j)}, \forall i \quad (1)$$

donde $máx_i$ and min_i son el valor máximo y el valor de la expresión génica respectivamente de las características c_i sobre los conjuntos de datos [25]. Luego, se realiza un proceso de eliminación de los predictores correlacionados para mejorar la precisión y acelerar el entrenamiento. Para ello se utiliza el paquete *Caret* de R utilizando la función *cor*() en cada conjunto de datos [26].

Fig. 1. Esquema general de la estrategia propuesta.

B. *k-vecinos más cercanos*

k-NN es un algoritmo no paramétrico que almacena todos los casos disponibles y clasifica los nuevos basado en una medida de similitud (comúnmente distancia). Un caso es clasificado por mayoría de sus vecinos, la clase será la más común entre sus *k* vecinos más cercanos medida por la función de distancia. Si *k*=1 entonces la clase es la del vecino más cercano [27]. Para nuestra estrategia, se utiliza distancia Euclidea con el objetivo de predecir la clase de una muestra en el conjunto de prueba, considerando las características seleccionadas (con un 1) y el número de vecinos (*k*=5).

C. *Algoritmos Evolutivos Multiobjetivo*

En esta sección, introduciremos las principales características de los AEMOs utilizados.

Non-dominated Sorting Genetic Algorithm, version II (NSGA-II) [28] es un AEMO del estado del arte que ha sido aplicado de manera satisfactoria en muchas áreas. NSGA-II aplica dominancia de Pareto para el cálculo de la función de aptitud (fitness) construyendo frentes de soluciones. La búsqueda evolutiva en NSGA-II supera a su versión previa (NSGA) al utilizar: *i*) un ordenamiento no dominado elitista que reduce la complejidad del chequeo de dominancia; *ii*) una técnica de hacinamiento (*crowding*) para la conservación de la diversidad; y *iii*) un fitness que considera los valores de dicha distancia *crowding*.

Multiobjective Cross-generational elitist selection, Heterogeneous recombination, and Cataclysmic mutation (MOCHC) [29], es la versión multiobjetivo del Algoritmo CHC. CHC fue diseñado especialmente para trabajar con soluciones codificadas en binario como selección de características. En un paso de CHC, un nuevo conjunto de soluciones es construida seleccionando pares de soluciones de la población (padres) y recombinándolos. Esta selección es realizada de tal manera que individuos similares no pueden aparearse entre ellos. CHC aplica solamente un operador de recombinación llamado HUX (*half uniform crossover*). HUX toma dos padres y decide aleatoriamente un intercambio para aquellos bits en los que sus alelos difieren. Los bits de la cadena para los cuales ambos padres tienen el mismo valor no se cambian. En MOCHC, las soluciones se ordenan utilizando una clasificación y un estimador de distancia de hacinamiento similar a los utilizados en NSGA-II.

D. *Codificación de la solución*

La codificación de la selección de características es simple, cada cromosoma codifica las características seleccionadas utilizando un vector binario donde cada posición representa una característica del conjunto de datos. El valor 1 en la posición *i* implica que la característica *i* es seleccionada mientras que 0 indica que *i* no es seleccionada. El tamaño de la solución es proporcional al número de características luego del preprocesamiento.

IV. Experimentos

En esta sección introducimos los experimentos realizados para el problema. Los conjuntos de datos en este trabajo son de tres tipos de cáncer distintos:

Conjunto de datos de Cáncer de Colon [15]: tiene genes 2000 genes; 40 muestras de tejido, de los 62 totales, fueron biopsias de tumores etiquetas como negativas y 23 son biopsias normales etiquetadas como positivas.

Conjunto de datos de Linfoma [16]: posee 4026 genes; 24 de las 42 muestras de tejido están etiquetadas como centro

germinal tipo B y 23 como centro tipo B activado.

Conjunto de datos de Leucemia [17]: contiene 7129 genes; 47 muestras de las 72 muestras de tejido se han tomado de pacientes con leucemia linfoblástica aguda y 25 de pacientes con leucemia mieloide aguda.

Todos los conjuntos fueron preprocesados utilizando el paquete de R Caret (*Classification And REgression Training*) [26] y ClusterSim (*Searching for Optimal Clustering Procedure for a DataSet*) [25] a fin de normalizar y reducir la dimensionalidad en columnas de las matrices de datos.

La normalización de las columnas se realiza primero, usando ClusterSim en cada uno de los conjuntos de datos descritos. Luego, la biblioteca Caret permite encontrar las columnas altamente correlacionadas y eliminarlas de los datos. Esto resultó en la reducción del conjunto de datos del cáncer de Colon, de 2000 a 224 genes; el conjunto de datos de linfoma, de 4026 a 2840 genes y el conjunto de datos de leucemia de 7129 a 5816 genes. Los conjuntos de datos reducidos se dividen luego con Caret, en conjuntos de entrenamiento (75%) y de prueba (25%) para ejecutar y validar los diferentes métodos de clasificación descritos.

Todos los experimentos se ejecutaron sobre una CPU AMD FX(tm)-8320 8-Cores, con memoria de 16GB. El sistema operativo utilizado es Xenial Xerus 16.04 LTS. Para cada experimento se realizaron 20 ejecuciones independientes de cada algoritmo a fin de asegurar relevancia estadística. Los algoritmos fueron implementados bajo Java 1.8 utilizando jMetal framework 5.4 [30]. Para ambos, la condición de parada fue establecida en 25000 evaluaciones del fitness. En este estudio, se utilizaron los valores comunes encontrados en la literatura para los parámetros de cada algoritmo (Tabla I).

TABLA I
CONFIGURACIÓN UTILIZADA PARA NSGA-II Y MOCHC.

Algoritmo	Parámetro	Valor
NSGA-II	Tamaño de la Población	100
	Cruzamiento	SPX con Probabilidad de 90%
	Mutación	Bit-Flip con Probabilidad (1/n)
	Selección	Torneo Binario
MOCHC	Tamaño de la Población	100 (10x10)
	Cruzamiento	HUX con Probabilidad de 100%
	Mutación	Cataclismo con probabilidad de 35%
	Población Preservada	5
	Convergencia Inicial	25% del tamaño de la instancia
	Valor de Convergencia	3
	Selección	Aleatoria con tratamiento de incesto
	Criterio de Ordenamiento	Ranking y distancia crowding

V. RESULTADOS

En esta sección mostramos los resultados de la evaluación de los algoritmos para los tres conjuntos de datos utilizados. La Tabla II muestra la instancia evaluada y el algoritmo testeado en la columna uno y dos. Las columnas tres y cuatro muestran el promedio y desvío estándar de las características seleccionadas y la precisión de la clasificación respectivamente. La última columna muestra el resultado para el análisis estadístico entre ambos algoritmos: una flecha hacia abajo indica que MOCHC tiene valores estadísticamente más altos que NSGA-II.

TABLA II
RESULTADOS OBTENIDOS COLON, LINFOMA Y LEUCEMIA.

Instancia	Algoritmo	Nro. de Características Seleccionadas	Porcentaje Precisión (%)	Nivel de Significancia
Colon	NSGA-II	22.40 ± 1.84	100.00 ± 0.00	↓
	MOCHC	1.60 ± 0.52	100.00 ± 0.00	
Linfoma	NSGA-II	1049.50 ± 10.05	90.00 ± 0.00	↓
	MOCHC	297.17 ± 32.27	95.00 ± 5.48	
Leucemia	NSGA-II	822.00 ± 12.08	90.00 ± 0.00	↓
	MOCHC	234.00 ± 8.10	100.00 ± 0.00	

Se puede observar en la Tabla II que el MOCHC registra reducciones significativas en el número de características seleccionadas en contraposición con NSGA-II. Ambos AEMOs alcanzaron el 100% de exactitud para el conjunto de datos de cáncer de colon. De la misma forma, estos resultados son confirmados por los valores estadísticos obtenidos por el MOCHC. MOCHC es el que mejor explora el espacio de búsqueda en función de los resultados obtenidos. Esto se debe probablemente al operador especial de la técnica CHC que evita la convergencia prematura al obtener un equilibrio adecuado entre la exploración del espacio de búsqueda y la explotación de las características locales de las soluciones.

La Fig. 2 compara la evolución de los objetivos con respecto a las evaluaciones realizadas de los AEMOs. El número de características se muestran a la izquierda, el porcentaje alcanzado de precisión se muestra en el eje derecho y el porcentaje de evaluaciones realizadas se presenta en la parte inferior. El número de características seleccionadas disminuye a medida que aumentan las evaluaciones, mientras que la precisión aumenta con el incremento de la evaluación. Las líneas de puntos con círculos (∘) se refieren al porcentaje de valores de precisión, mientras que las líneas continuas con cuadros (□) son la evolución de la cantidad de entidades.

En la Fig. 2 se observa para que el porcentaje de precisión escala rápidamente para colon. Cuando se alcanza el 60% de las evaluaciones, MOCHC alcanza el 100% de precisión. Por otro lado, la reducción del número de características evoluciona moderadamente. Vale la pena señalar que NSGA-II evoluciona bien hasta que se alcanza el 50% de las evaluaciones; sin embargo, a partir de ese momento, el número de características se estanca. Como resultado, se obtiene el peor número de características (una cantidad mayor).

Para las otras dos instancias (Linfoma y Leucemia) se observan comportamientos muy similares. En cuanto al número de características seleccionadas, ambos algoritmos evolucionan de manera similar al principio, pero el MOCHC comienza a separarse (reduciendo la cantidad) del resto a partir del 20% de las evaluaciones, logrando el número mínimo de características. De forma general, los resultados numéricos revelan la capacidad de ambos AEMOs para obtener resultados competitivos, pero MOCHC alcanza los mejores resultados para todos los conjuntos de datos.

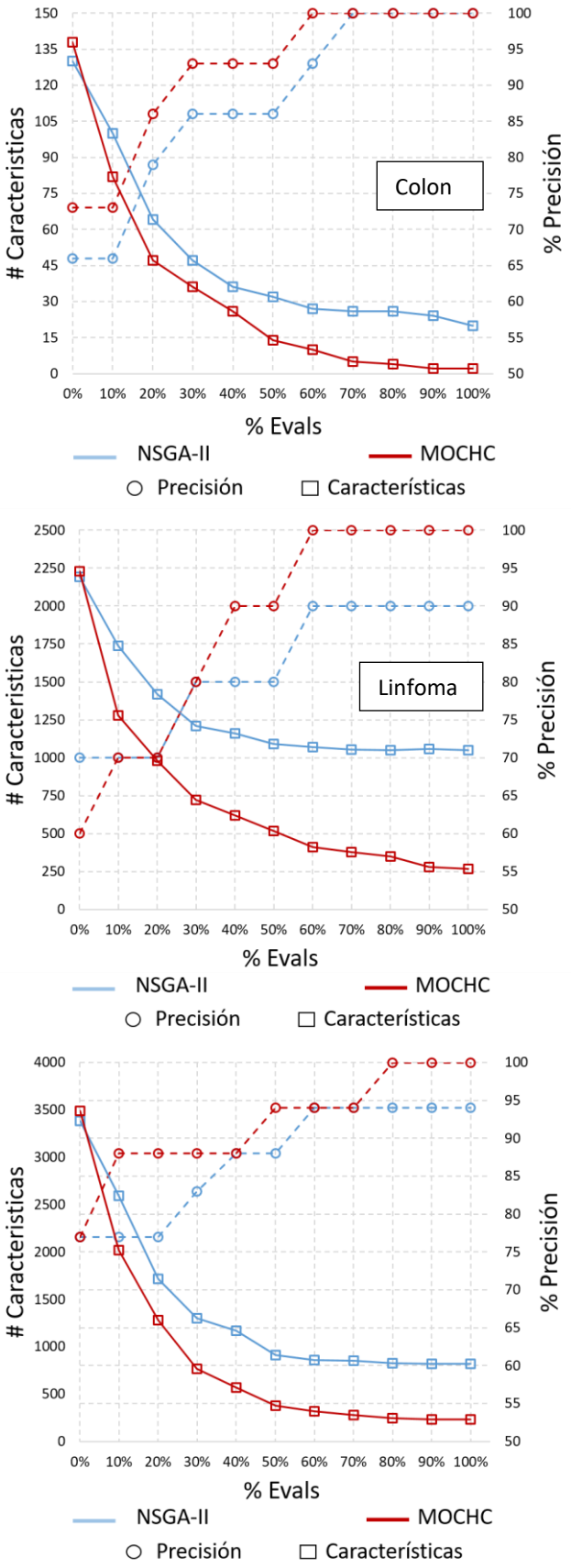

Fig. 2. Evolución de los objetivos para los conjuntos de datos.

VI. Conclusiones

En este trabajo, abordamos el problema de seleccionar un número de características lo más reducida posible que maximice la precisión de la clasificación para microarrays de datos de cáncer. Se evaluaron dos algoritmos del estado del arte, uno de ellos el NSGA-II, un algoritmo clásico en la literatura multiobjetivo, y el MOCHC, un algoritmo especialmente diseñado para problemas con representación binaria. Se utilizaron tres populares conjuntos de datos de cáncer para evaluar estos algoritmos. A fin de realizar una clasificación precisa seleccionando la menor cantidad de características se utilizó la técnica *k*-vecinos más cercanos. Una etapa de preprocesamiento permitió eliminar características redundantes e irrelevantes. Luego, se aplicaron ambos algoritmos. Se pudo comprobar que el MOCHC mejora notablemente los resultados con respecto al NSGA-II. En términos generales, ambos algoritmos demuestran resolver de forma satisfactoria el problema. En particular, MOCHC, al ser un algoritmo diseñado especialmente para problemas binarios tiene claras ventajas con respecto a NSGA-II.

Para trabajos futuros, nuestra idea es aplicar MOCHC a otros tipos de conjuntos de datos de cáncer y otros microarrays de expresiones de genes. También, la precisión de la clasificación podría ser testeada utilizando otro tipo de clasificadores.

Agradecimientos

Dussaut y Ponzoni agradecen al Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET) por el proyecto PIP es 112-2017-0100829. Olivera y Vidal agradecen a la Universidad Nacional de Cuyo (UNCuyo) por el Proyecto Tipo 1 B081 y a la Facultad de Ingeniería de la UNCuyo.

Referencias

[1] S. Selvaraj and J. Natarajan, “Microarray data analysis and mining tools,” *Bioinformation*, vol. 6, no. 3, pp. 95–99, 2011.

[2] P. M. Narendra and K. Fukunaga, “A branch and bound algorithm for feature subset selection,” *IEEE Transactions on Computers*, vol. C-26, no. 9, pp. 917–922, Sept 1977.

[3] M. Dash and H. Liu, “Feature selection for classification,” *Intelligent data analysis*, vol. 1, no. 3, pp. 131–156, 1997.

[4] H. Liu and Z. Zhao, “Manipulating data and dimension reduction methods: Feature selection,” in *Encyclopedia of Complexity and Systems Science*. Springer, 2009, pp. 5348–5359.

[5] H. Liu, H. Motoda, R. Setiono, and Z. Zhao, “Feature selection: An ever-evolving frontier in data mining,” in *Feature Selection in Data Mining*, 2010, pp. 4–13.

[6] A. W. Whitney, “A direct method of nonparametric measurement selection,” *IEEE Transactions on Computers*, vol. 100, no. 9, pp. 1100–1103, 1971.

[7] T. Marill and D. Green, “On the effectiveness of receptors in recognition systems,” *IEEE transactions on Information* Theory, vol. 9, no. 1, pp. 11–17, 1963.

[8] P. Pudil, J. Novovicov˘a,´ and J. Kittler, “Floating search methods in feature selection,” *Pattern recognition letters*, vol. 15, no. 11, pp. 1119– 1125, 1994.

[9] Q. Mao and I. W.-H. Tsang, “A feature selection method for multivariate performance measures,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 9, pp. 2051–2063, 2013.

[10] F. Min, Q. Hu, and W. Zhu, “Feature selection with test cost constraint,” *International Journal of Approximate Reasoning*, vol. 55, no. 1, pp. 167– 179, 2014.

[11] B. Xue, M. Zhang, W. N. Browne, and X. Yao, “A survey on evolutionary computation approaches to feature selection,” *IEEE Trans. Evol. Comput.*, vol. 20, no. 4, pp. 606–626, 2016.
[12] C. S. R. Annavarapu, S. Dara, and H. Banka, “Cancer microarray data feature selection using multi-objective binary particle swarm optimization algorithm,” *EXCLI Journal*, vol. 15, pp. 460–473, 2016.
[13] A. Hasnat and A. U. Molla, “Feature selection in cancer microarray data using multi-objective genetic algorithm combined with correlation coef-ficient,” in *2016 International Conference on Emerging Technological Trends (ICETT)*, 2016, pp. 1–6.
[14] M. M. Mafarja and S. Mirjalili, “Hybrid whale optimization algorithm with simulated annealing for feature selection,” *Neurocomputing*, vol. 260, pp. 302 – 312, 2017.
[15] U. Alon, N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack, and J. Levine, “Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays,” *Proceedings of the National Academy of Sciences*, vol. 96, no. 12, pp. 6745–6750, 1999.
[16] A. A. Alizadeh, M. B. Eisen, R. E. Davis, C. Ma, I. S. Lossos, Rosenwald, J. C. Boldrick et al., “Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling,” *Nature*, vol. 403, no. 6769, pp. 503–511, 2000.
[17] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri et al., “Molecular classification of cancer: class discovery and class prediction by gene expression monitoring,” *Science*, vol. 286, no. 5439, pp. 531– 537, 1999.
[18] J. S. Dussaut, C. A. Gallo, F. Cravero, M. J. Martínez, J. A. Carballido, and I. Ponzoni, “Gernet: a gene regulatory network tool,” *Biosystems*, vol. 162, pp. 1–11, 2017.
[19] J. A. Carballido, C. A. Gallo, J. S. Dussaut, and I. Ponzoni, “On evolutionary algorithms for biclustering of gene expression data,” *Current Bioinformatics*, vol. 10, no. 3, pp. 259–267, 2015.
[20] P. G. Kumar, T. A. A. Victoire, P. Renukadevi, and D. Devaraj, “Design of fuzzy expert system for microarray data classification using a novel genetic swarm algorithm,” *Expert Systems with Applications*, vol. 39, no. 2, pp. 1811–1821, 2012.
[21] R. K. Singh and M. Sivabalakrishnan, “Feature selection of gene expression data for cancer classification: a review,” *Procedia Computer Science*, vol. 50, pp. 52–57, 2015.
[22] S. Shahbeig, M. S. Helfroush, and A. Rahideh, “A fuzzy multi-objective hybrid tlbo–pso approach to select the associated genes with breast cancer,” *Signal Processing*, vol. 131, pp. 58–65, 2017.
[23] K. Deb, *Multi-Objective Optimization using Evolutionary Algorithms*. John Wiley & Sons, 2001.
[24] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification* (*2nd Ed*). Wiley, 2001.
[25] M. Walesiak, A. Dudek, and M. A. Dudek, “clustersim package,” 2011.
[26] M. Kuhn, “Caret package,” *Journal of Statistical Software*, vol. 28, no. 5, pp. 1–26, 2008.
[27] N. S. Altman, “An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression,” *The American Statistician*, vol. 46, no. 3, pp. 175–185, 1992.
[28] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan, “A fast and elitist multiobjective genetic algorithm: NSGA-II,” *IEEE Trans. Evol. Comput.*, vol. 6, no. 2, pp. 182–197, 2002.
[29] A. J. Nebro, E. Alba, G. Molina, F. Chicano, F. Luna, and J. J. Durillo, “Optimal antenna placement using a new multi-objective chc algorithm,” in *9th annual conference on Genetic and evolutionary computation*. New York, NY, USA: ACM Press, 2007, pp. 876–883.
[30] J. J. Durillo and A. J. Nebro, “jMetal: A java framework for multi-objective optimization,” *Advances in Engineering Software*, vol. 42, pp. 760–771, 2011.