**II Conferencia Internacional de Procesamiento de la Información (CIPI 2019)**

**Algoritmo de envoltorio para el aprendizaje multi-instancia: resultados iniciales**

***Wrapper algorithm for multiple-instance learning: early results***

**Luis Quintero-Domínguez1,2, Carlos Morell2, Alberto Verdecia-Cabrera2,3, Yojacni Companioni García1**

1- Universidad de Sancti Spíritus “José Martí Pérez”, Cuba. lqdominguez@uniss.edu.cu

2- Universidad Central “Marta Abreu” de Las Villas, Cuba. cmorellp@uclv.edu.cu

3- Universidad de Granma, Cuba. averdeciac@udg.co.cu

**Resumen:** El aprendizaje multi-instancia es una generalización del aprendizaje supervisado, donde cada ejemplo se representa por una bolsa etiquetada formada por un conjunto de instancias. Varios métodos de aprendizaje multi-instancia transforman cada bolsa en una única instancia para luego aplicar métodos de aprendizaje supervisado estándar. En este trabajo se presenta un nuevo método de aprendizaje multi-instancia que transforma los datos multi-instancia inspirado en la minería de textos. El método propuesto transforma los datos multi-instancia en una representación atributo-valor tradicional mediante la creación de un corpus de documentos formados por palabras artificiales para reducir la pérdida de información durante el proceso de transformación. Además, se evaluó experimentalmente el método propuesto utilizando nueve conjuntos de datos multi-instancia y dos métodos de aprendizaje que transforman los datos multi-instancia a una representación atributo-valor tradicional. El estudio experimental realizado indica que, en términos de precisión de la clasificación, el método propuesto es competitivo con los métodos de aprendizaje utilizados en la comparación.

***Abstract:*** *Multi-instance learning is a generalization of supervised learning, where each example is represented by a labeled bag composed by a set of instances. Several multi-instance learning methods transform each bag into a single instance and then apply standard supervised learning methods. This paper presents a new multi-instance learning method that transforms the multi-instance data and is inspired by text mining. The proposed method transforms the multi-instance data into a traditional attribute-value representation by creating a corpus of documents formed by artificial words to reduce the loss of information during the transformation process. In addition, the proposed method was empirically evaluated using nine multi-instance datasets and two learning methods that transform the multi-instance data into a traditional attribute-value representation. The empirical study indicates that, in terms of classification accuracy, the proposed method is competitive with the learning methods used in the comparison*.

**Palabras Clave:** Aprendizaje multi-instancia; Bolsa-de-palabras; Método envoltorio.

***Keywords:*** *Multiple-instance learning; Bag-of-words; Wrapper method*.

# 1. Introducción

El aprendizaje multi-instancia es una generalización del aprendizaje proposicional estándar también llamado aprendizaje atributo-valor y fue introducido por primera vez en (Dietterich et al., 1997). Mientras que en el aprendizaje estándar un ejemplo se representa por un vector de tamaño fijo de pares atributo-valor que tiene asociado una etiqueta de clase, en el aprendizaje multi-instancia un ejemplo se representa por una bolsa de vectores atributo-valor y la etiqueta de clase está asociada a la bolsa completa.

El aprendizaje multi-instancia ha despertado un interés creciente debido fundamentalmente a la gran variedad de problemas reales que pueden ser modelados de manera muy natural como problemas multi-instancia. Estos problemas incluyen clasificación de textos (Andrews et al., 2002), recuperación y clasificación de imágenes (Chen et al., 2006; Chen and Wang, 2004), predicción de actividad farmacológica (Dietterich et al., 1997), recomendación de páginas web índice (Sánchez Tarragó et al., 2014) y predicción de rendimiento académico (Zafra and Ventura, 2012).

Desde la introducción del aprendizaje multi-instancia, el número de métodos de clasificación multi-instancia ha crecido considerablemente. Muchos autores han propuesto categorías para tratar de capturar las características distintivas de estos métodos (Amores, 2013). Recientemente (Herrera et al., 2016) propusieron tres categorías principales:

* Métodos basados en instancias: estos son algoritmos donde el proceso de aprendizaje ocurre a nivel de las instancias.
* Métodos basados en bolsa: incluye clasificadores que trabajan directamente en el espacio de las bolsas.
* Métodos basados en mapeo (envoltorios): son clasificadores que aplican una transformación a los datos del problema multi-instancia de forma que puedan aplicarse los algoritmos del aprendizaje supervisado tradicional para obtener la solución.

Existen métodos pertenecientes a la categoría de envoltorio que transforman los problemas multi-instancia en problemas de aprendizaje tradicional remplazando cada bolsa con un vector de atributos consistente en un resumen estadístico derivado de las instancias en la bolsa. Estos métodos pueden provocar pérdida de información al transformar los problemas multi-instancia originales, lo cual afecta la eficacia de la clasificación (Fu and Robles-Kelly, 2009).

Aquí se presenta un método de aprendizaje multi-instancia perteneciente a la categoría de los basados en mapeo. El método propuesto, llamado MIBoW, está inspirado por las técnicas de minería de textos y otros campos donde se ha utilizado la representación de bolsa de palabras (Aggarwal and Zhai, 2012; Peng et al., 2016; Perovšek et al., 2015). MIBoW pretende lograr una reducción de la pérdida de información durante la transformación de los datos multi-instancia en una representación atributo-valor tradicional. MIBoW puede ser visto como una transformación del conjunto de datos multi-instancia en un corpus de documentos, donde cada bolsa se convierte en un documento descrito por un conjunto de palabras artificiales que serán los atributos en el conjunto de datos transformado.

En este trabajo se muestra la evaluación experimental inicial realizada para evaluar la eficacia del método propuesto, donde se utilizaron nueve conjuntos de datos y dos métodos de aprendizaje multi-instancia basados en mapeo. Los resultados experimentales indican que el método propuesto es competitivo con los métodos de aprendizaje utilizados.

# 2. Metodología

En esta sección se da una pequeña introducción al aprendizaje multi-instancia, se presenta el método propuesto y se describe el estudio experimental realizado.

## 2.1 Clasificación multi-instancia

En clasificación multi-instancia, un ejemplo de entrenamiento es una bolsa que contiene múltiples instancias descritas por vectores atributo-valor y tiene asociada una única etiqueta de clase (Herrera et al., 2016). Formalmente, en la clasificación multi-instancia, un ejemplo es un par , donde es un multi-conjunto (bolsa) de instancias y es la etiqueta de clase del ejemplo. Una bolsa se define como un multi-conjunto , porque múltiples copias de la misma instancia pueden estar incluidas en una bolsa. Las instancias ) son vectores del espacio -dimensional formado por el producto vectorial de los atributos que describen a las instancias y es el conjunto de etiquetas de clase. La tarea de clasificación multi-instancia es encontrar una función que, a partir de un conjunto de entrenamiento permita predecir la clase de un ejemplo no visto con anterioridad.

Los métodos de clasificación multi-instancia generalmente asumen la existencia de cierta relación entre las instancias y la etiqueta de clase de la bolsa. A esta relación se le conoce como hipótesis multi-instancia. Actualmente existe una variedad de hipótesis multi-instancia que han sido introducidas a medida que se han desarrollado nuevos métodos de solución para los problemas multi-instancia (Amores, 2013). La primera hipótesis que se empleó para definir el aprendizaje multi-instancia fue la hipótesis estándar (Dietterich et al., 1997). La hipótesis estándar establece que una bolsa será positiva si y solo si contiene alguna instancia positiva. Es decir, si la bolsa es negativa todas sus instancias serán negativas, si la bolsa es positiva al menos una de sus instancias será positiva. Formalmente, dada una función , capaz de estimar las etiquetas de clase de una instancia, se puede describir la hipótesis estándar como:

## 2.2 Método MIBoW

Esta sección describe los pasos principales del método MIBoW. Primero, se realiza una transformación del conjunto de datos multi-instancia en un corpus de documentos representados en el formato Bolsa-de-Palabras (*Bag-of-Words*, BoW). Cada bolsa de instancias se transforma en un documento textual descrito por palabras artificiales, que se construyen mediante la combinación de los nombres de los atributos con su valor: [*nombre del atributo*]\_[*valor del atributo*]. Cabe resaltar, que los atributos necesitan ser discretizados con anterioridad, para que los valores numéricos no provoquen la generación de un número excesivo de palabras artificiales. El método propuesto recorre cada instancia de la bolsa y genera las palabras artificiales con cada uno de los atributos, para conformar el conjunto de palabras que formarán el documento correspondiente a la bolsa.

Luego se construyen los vectores de pares atributo-valor que representarán a las bolsas de la representación multi-instancia original. Para esto, cada una de las palabras artificiales que se generaron en el corpus de documentos se considera como un atributo en la nueva representación. El valor asociado a cada documento (ejemplo) para una palabra (atributo) es la frecuencia con que dicha palabra ocurrió en el documento. Después, al documento se le asocia la misma etiqueta de clase que presentaba la bolsa a la cual representa.

Finalmente, luego de transformar el conjunto de datos multi-instancia a la nueva representación atributo-valor, se entrena un algoritmo de aprendizaje tradicional que obtendrá un modelo capaz de clasificar una bolsa no vista con anterioridad luego de ser transformada a la nueva representación.

## 2.3 Configuración del estudio experimental

Esta sección presenta el estudio experimental inicial realizado para evaluar la eficacia del método propuesto. Con este objetivo se utilizaron nueve conjuntos de datos multi-instancia que se describen en la Tabla 1. Para la comparación, además del algoritmo propuesto, se emplearon dos métodos de aprendizaje multi-instancia, que al igual que MIBoW, realizan una transformación de los datos a una representación atributo-valor tradicional: SimpleMI (Dong, 2006) y MIWrapper (Frank and Xu, 2003).

Tabla Características de los conjuntos de datos utilizados en la experimentación.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Dataset | Atributos | Bolsas positivas | Bolsas negativas | Total bolsas |
| AntDrugs5 | 5 | 198 | 202 | 400 |
| Atoms | 10 | 125 | 63 | 188 |
| Chains | 24 | 125 | 63 | 188 |
| Corel01vs02 | 9 | 100 | 100 | 200 |
| Corel02vs03 | 9 | 100 | 100 | 200 |
| Corel03vs04 | 9 | 100 | 100 | 200 |
| Corel04vs05 | 9 | 100 | 100 | 200 |
| EastWest | 24 | 10 | 10 | 20 |
| TREC9Sel-1 | 299 | 200 | 200 | 400 |

Tanto el método propuesto como SimpleMI y MIWrapper, transforman los conjuntos de datos multi-instancia a una representación atributo-valor para luego utilizar métodos de clasificación tradicional. Por este motivo la comparación experimental se realizó utilizando los clasificadores base RandomForest y SMO.

Se utilizó la herramienta Weka para realizar la evaluación experimental y la medida empleada para medir la eficacia de los métodos fue la precisión de la clasificación. Además, se discretizaron los conjuntos de datos utilizando la subdivisión del rango de cada atributo en 10 intervalos de igual longitud. Los métodos de aprendizaje se utilizaron con los valores de los parámetros por defecto del Weka.

# 3. Resultados y discusión

Como se mencionó anteriormente en el estudio experimental se comparó el método propuesto MIBoW, con los métodos SimpleMI y MIWrapper . Estos métodos tienen que ser utilizados en combinación con un algoritmo de clasificación tradicional pues transforman los datos multi-instancia a una representación atributo-valor tradicional. Para esto se utilizaron RandomForest y SMO.

Tabla Resultados de la evaluación experimental (RF-RandomForest)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | RandomForest | | | SMO | | |
| Dataset | MIBoW-RF | SimpleMI-RF | MIWrapper-RF | MiBoW-SMO | SimpleMI-SMO | MIWrapper-SMO |
| AntDrugs5 | **72.25** | 58.50 | 71.75 | 69.25 | 60.00 | **71.25** |
| Atoms | **79.71** | 66.49 | 66.49 | **68.71** | 66.49 | 66.49 |
| Chains | **89.36** | 73.42 | 77.63 | **84.56** | 77.19 | 69.62 |
| Corel01vs02 | 84.50 | 70.00 | **86.00** | **87.00** | 72.50 | 82.50 |
| Corel02vs03 | **81.00** | 75.00 | 80.00 | **84.00** | 74.00 | 71.50 |
| Corel03vs04 | **95.50** | 90.00 | 88.50 | **94.00** | 90.00 | 74.00 |
| Corel04vs05 | **100.00** | 99.00 | 95.50 | **100.00** | 98.50 | 90.00 |
| EastWest | **80.00** | 70.00 | 65.00 | 70.00 | **80.00** | 60.00 |
| TREC9Sel-1 | 73.50 | 50.25 | **78.50** | **82.50** | 50.25 | 75.00 |

La Tabla 2 muestra los resultados de la evaluación experimental en términos de precisión de la clasificación. Analizando las combinaciones con RandomForest, se puede apreciar que MIBoW obtiene el mejor valor de precisión de la clasificación en siete de los nueve conjuntos de datos. Adicionalmente, para comprobar si estas diferencias son estadísticamente significativas, se realizaron pruebas estadísticas siguiendo la metodología propuesta por (Demšar, 2006) y (García and Herrera, 2008) para comparar varios clasificadores sobre varios conjuntos de datos. En la Figura 1 se muestra la comparación entre las combinaciones con RandomForest utilizando la prueba de Friedman y el procedimiento de Shaffer para el análisis post hoc con un valor . En esta figura se puede apreciar que la combinación con MIBoW es significativamente superior a las que se realizan con SimpleMI y MIWrapper.

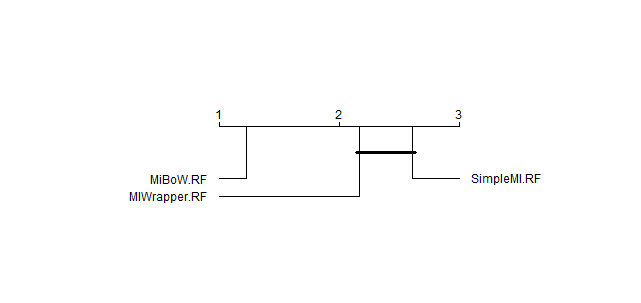


Figura Cmparación utilizando RandomForest

Analizando las combinaciones con SMO, se puede apreciar que MiBoW también obtiene el mejor valor de precisión de la clasificación en siete de los nueve conjuntos de datos. De manera similar a la metodología seguida con las combinaciones con RandomForest para comprobar si estas diferencias son estadísticamente significativas, se utilizó la prueba de Friedman y el procedimiento de Shaffer para el análisis post hoc con un valor . En la Figura 2 se puede apreciar que la combinación con MIBoW obtuvo el primer lugar en el *ranking* de Friedman y es significativamente superior a las que se realizan con SimpleMI y MIWrapper.

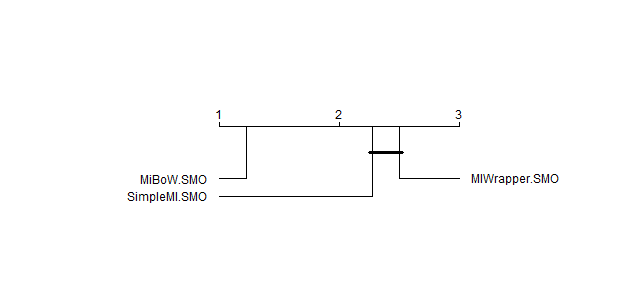


Figura Comparación utilizando SMO.

# 4. Conclusiones

En este trabajo se presenta un nuevo método de aprendizaje multi-instancia basado en mapeo, llamado MIBoW. El método propuesto está inspirado en las técnicas de minería de texto, en especial la representación Bolsa-de-Palabras. MIBoW transforma los datos multi-instancia en una representación atributo-valor tradicional mediante la creación de un corpus de documentos formados por palabras artificiales para reducir la pérdida de información durante el proceso de transformación. El estudio experimental realizado indica que, en términos de precisión de la clasificación, el método propuesto es superior a otros métodos que transforman los datos multi-instancia en una representación atributo-valor.

Como trabajo futuro se planea explorar el efecto de la utilización de métodos de pesado de palabras típicos de la minería de texto como es el caso del TF-IDF. Además, se pretende aumentar el estudio experimental utilizando otros métodos de aprendizaje y conjuntos de datos multi-instancia, para explorar con mayor detalle las ventajas y posibles limitaciones de MIBoW.

# 5. Referencias bibliográficas

1. Aggarwal, C.C., Zhai, C. (Eds.), 2012. Mining Text Data. Springer US, Boston, MA. https://doi.org/10.1007/978-1-4614-3223-4
2. Amores, J., 2013. Multiple instance classification: Review, taxonomy and comparative study. Artificial Intelligence 201, 81–105. https://doi.org/10.1016/j.artint.2013.06.003
3. Andrews, S., Tsochantaridis, I., Hofmann, T., 2002. Support vector machines for multiple-instance learning, in: Becker, S., Thrun, S., Obermayer, K. (Eds.), Advances in Neural Information Processing Systems. MIT Press, pp. 561–568.
4. Chen, Y., Bi, J., Wang, J.Z., 2006. MILES: Multiple-instance learning via embedded instance selection. IEEE Transactions on Pattern Analysis and Machine Intelligence 28, 1931–1947.
5. Chen, Y., Wang, J.Z., 2004. Image Categorization by Learning and Reasoning with Regions. Journal of Machine Learning Research 5, 913–939.
6. Demšar, J., 2006. Statistical Comparisons of Classifiers over Multiple Data Sets. Journal of Machine Learning Research 7, 1–30.
7. Dietterich, T.G., Lathrop, R.H., Lozano-Pérez, T., 1997. Solving the multiple instance problem with axis-parallel rectangles. Artificial Intelligence 89, 31–71. https://doi.org/10.1016/S0004-3702(96)00034-3
8. Dong, L., 2006. A comparison of multi-instance learning algorithms (Master Thesis). The University of Waikato.
9. Frank, E., Xu, X., 2003. Applying propositional learning algorithms to multi-instance data (Working Paper). Department of Computer Science, University of Waikato.
10. Fu, Z., Robles-Kelly, A., 2009. An instance selection approach to Multiple instance Learning, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition. Presented at the 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 911–918. https://doi.org/10.1109/CVPR.2009.5206655
11. García, S., Herrera, F., 2008. An Extension on “Statistical Comparisons of Classifiers over Multiple Data Sets” for all Pairwise Comparisons. Journal of Machine Learning Research 9, 2677–2694.
12. Herrera, F., Ventura, S., Bello-Pérez, R., Cornelis, C., Zafra Gómez, A., Sánchez-Tarragó, D., Vluymans, S., 2016. Multiple Instance Learning. Foundations and Algorithms. Springer International Publishing.
13. Peng, X., Wang, L., Wang, X., Qiao, Y., 2016. Bag of visual words and fusion methods for action recognition: Comprehensive study and good practice. Computer Vision and Image Understanding 150, 109–125. https://doi.org/10.1016/j.cviu.2016.03.013
14. Perovšek, M., Vavpetič, A., Kranjc, J., Cestnik, B., Lavrač, N., 2015. Wordification: Propositionalization by unfolding relational data into bags of words. Expert Systems with Applications 42, 6442–6456. https://doi.org/10.1016/j.eswa.2015.04.017
15. Sánchez Tarragó, D., Cornelis, C., Bello, R., Herrera, F., 2014. A multi-instance learning wrapper based on the Rocchio classifier for web index recommendation. Knowledge-Based Systems 59, 173–181. https://doi.org/10.1016/j.knosys.2014.01.008
16. Zafra, A., Ventura, S., 2012. Multi-instance genetic programming for predicting student performance in web based educational environments. Applied Soft Computing 12, 2693–2706. https://doi.org/10.1016/j.asoc.2012.03.054