On-line Learning With Reject Option

G. J. Pérez, M. Santibáñez, R. M. Valdovinos, J. R. Marcial, M. Romero, R. Alejo

***Abstract*— On-line learning is a training paradigm that allows the processing of constant data flows, so that learning adapts to new knowledge. However, due to the nature of the study problem, it is possible that in the clustering obtained there are data complexities (outliers, atypical patterns, noisy, etc.) that deteriorate the performance of the model in the classification stage. Due to the above, an alternative to cope data complexities is the use of algorithms that allow to detect reject options to filter noisy pattern. In this research the neighborhood-based reject option is implemented in an on-line learning process, with the intention of improving the clustering quality and thus increasing the precision indexes obtained with the nearest neighbor's rule in the classification stage. Likewise, to validate the quality of the clustering generated, internal and external analysis metrics are used. The experimental results show the viability of the proposal when analyzed on real data.**

***Keywords*— Preprocessing, On-line Learning, Clustering, Classification, Data Mining.**

I. INTRODUCCIÓN

EL APRENDIZAJE *On-line* trabaja basado en flujo constante de datos, el cual debe ser procesado y clasificado en tiempo real [1], aprovechando el conocimiento generado mientras el clasificador realiza su trabajo. Para lograr lo anterior, el aprendizaje *on-line* se ejecuta sobre lotes de datos, en los que se presenta parte del total de los datos de entrenamiento, con los que el modelo realiza la clasificación de datos nuevos [2].

En su funcionamiento general, el aprendizaje *on-line* considera que no se tiene conocimiento sobre la distribución en clases de sus datos de entrada, por lo que con la ayuda de algoritmos de agrupamiento o *clustering* se logra identificar de forma automática la distribución de los datos en grupos, mismos que se transformarán en clases. Posteriormente, conforme van llegando más patrones, éstos son incorporados al grupo con el que se identifique mayor semejanza. Finalmente, cuando se tiene un patrón nuevo que se desea clasificar, se hace uso de la distribución obtenida hasta el momento para realizar el reconocimiento hacia el grupo con que muestre mayor semejanza [3].

Para validar la calidad del agrupamiento obtenido, en la literatura se proponen métodos de análisis de condensado y separabilidad de grupos. No obstante, dependiendo de la distribución de los datos, se pueden obtener agrupamientos con complejidades de datos (desbalance, solapamiento, *outliers*, etc.) que pueden deteriorar el rendimiento del clasificador en la etapa de reconocimiento [4].

Los algoritmos de preprocesado orientados a identificar, eliminar y/o disminuir los efectos negativos que las complejidades de datos tienen en el rendimiento del clasificador son los de limpieza [5], condensado [6] y reetiquetado de patrones también conocido como Opción de Rechazo [7].

Esta investigación se enfoca a analizar la pertinencia de utilizar algoritmos de limpieza del conjunto de datos (CD) posterior a un proceso de agrupamiento dentro de una estrategia de aprendizaje *On-line*. De forma específica, se aplicará un algoritmo de reetiquetado o rechazo basado en la regla de los K-vecinos, los cuales realizan el reetiquetado de aquellos patrones que se encuentren mal ubicados o bien el rechazo (eliminación) de patrones que presentan inconsistencia respecto a los grupos obtenidos por el algoritmo de agrupamiento.

II. MÉTODOS Y HERRAMIENTAS

2.1 Algoritmo Batchelor & Wilkins

Batchelor & Wilkins es un algoritmo heurístico incremental que emplea un único parámetro (θ), el cual representa la fracción de la distancia media entre agrupaciones, utilizado para calcular un umbral de distancia que determina si se crea o no un nuevo agrupamiento [8] tal como se observa en el Algoritmo 1.

Algoritmo 1: Batchelor & Wilkins
Entradas: θ = Fracción de la distancia media entre agrupaciones
1. Seleccionar un patrón al azar y convertirlo en el centro del primer agrupamiento.
2. Seleccionar el patrón más alejado del primer agrupamiento y convertirlo en el centro del segundo agrupamiento.
3. Calcular las distancias de todos los patrones a los centros y determinar la distancia mínima a un centro para cada patrón.
4. Obtener el patrón más alejado de los agrupamientos existentes (Máximo de las distancias mínimas de los patrones a los agrupamientos)
5. Si la distancia del patrón escogido es mayor que el umbral especificado, crear un nuevo agrupamiento con el patrón seleccionado.
6. Repetir los pasos 3, 4 y 5 hasta que ya no se creen nuevos agrupamientos.
7. Asignar cada patrón a su agrupamiento más cercano.

El algoritmo funciona de la siguiente manera: El primer centro se crea tomando un patrón al azar del CD, posteriormente un segundo centro es generado utilizando el patrón más alejado del primer centro. A partir de este punto, se obtiene el patrón más alejado a los grupos existentes, si la distancia del patrón escogido al conjunto de centros es mayor al umbral establecido (θ) se crea un nuevo grupo con dicho patrón, lo anterior se repite mientras se generen nuevos grupos. Finalmente se asigna cada patrón a su agrupamiento

G. J. Pérez, Universidad Autónoma del Estado de México, Facultad de Ingeniería. Cd. Universitaria, Toluca, México, gerardo-jpa@hotmail.com.
M. Santibáñez, Universidad Autónoma del Estado de México, Facultad de Ingeniería. Cd. Universitaria, Toluca, México, monicass_isc@hotmail.com.
R. M. Valdovinos, Universidad Autónoma del Estado de México, Facultad de Ingeniería. Cd. Universitaria, Toluca, México, li_rmvr@hotmail.com.
J.R Marcial, Universidad Autónoma del Estado de México, Facultad de Ingeniería. Cd. Universitaria, Toluca, México, jrmarcialr@uaemex.mx.
M. Romero, Universidad Autónoma del Estado de México, Facultad de Ingeniería. Cd. Universitaria, Toluca, México, mrh1601@yahoo.com.
R. Alejo, Instituto Tecnológico de Estudios Superiores de Jocotitlán, Jocotitlán, México, ralejoll@hotmail.com
Corresponding author: Gerardo Javier Pérez Álvarez