

# DUDAS Y ANOTACIONES

## DUDAS

- Preguntar si es necesario meter los tipos de clustering *density-based*, *grid-based* y *model-based* o con *partitioning* y *micro-clusters*. -> [Centrarnos en partitioning, hierarchical, model-based \(probabilistic\) y micro-clusters](#).
  - **Partitioning algorithms:** construct a partition of a set of objects into  $k$  clusters, that minimize an objective function (e.g. the sum of squares distances to the centroid representative). Examples include k-means (Farnstrom et al., 2000), and k-medoids;
  - **Hierarchical algorithms.** (En otro sitio sale este tipo en lugar de micro-clustering)
  - **Micro-clustering algorithms:** divide the clustering process into two phases, where the first phase is online and summarizes the data stream in local models (micro-clusters) and the second phase generates a global cluster model from the micro-clusters. Examples of these algorithms include BIRCH (Zhang et al., 1996) and CluStream (Aggarwal et al., 2003);
  - **Density-based algorithms** are based on connectivity between regions and density functions. This type of algorithms find clusters of arbitrary shapes, e.g., DBSCAN (Birant and Kut, 2007), and OPTICS (Peter Kriegel et al., 2003). Se basa en la detección de en qué áreas existen concentraciones de puntos y dónde están separados por áreas vacías o con escasos puntos. Los puntos que no forman parte de un clúster se etiquetan como ruido.
  - **Grid-based algorithms:** based on a multiple-level granularity structure. View instance space as grid structures, e.g., Fractal Clustering (Barbará and Chen, 2000), and STING (Hinneburg and Keim, 1999). Grid based methods quantize the object space into a finite number of cells (hyper-rectangles) and then perform the required operations on the quantized space.
  - **Model-based algorithms:** find the best fit of the model to all the clusters. Good for conceptual clustering, e.g., COBWEB (Fisher, 1987), and SOM (Kaski and Kohonen, 1994). Model based clustering operates on the assumption that gene expression data originates from a finite mixture of underlying probability distributions (Ramoni et al. 2001). Each cluster corresponds to a different distribution, and generally, the distributions are assumed to be Gaussians.
- STREAM -> No sé que relación hay entre SSQ minimization y facility location -  $> \zeta \dots$  and instead evaluates an algorithm's performance by a combination of SSQ and the numbers of centers used?
- STREAM -> k-Median : N° of medians to be at most  $k$ ; Facility location -> Range of number of centers.  $\zeta$  Why facility location is more convenient if they

look for a  $z$  that gives a certain  $k$ ? -> ¿K-Median es NP-Hard pero utilizando Facility Location en el Local Search es factible?

- ¿El recurring concept drift de la propuesta RGNBC (Redes Bayesianas) tiene que ver con la definición de este concepto? -> En RGNBC puede que no tenga que ver con un concept drift recurrente puesto que habla sobre la llegada de nuevos atributos.
- ¿Por qué en la ecuación 9 se tienen en cuenta aquellos valores de instancias que no pertenecen a la clase  $a$ ? (Propuesta RGBNC)
- En la propuesta RGNBC, en la ecuación 12, ¿la evidencia no debería ser el sumatorio del numerador de  $\text{posterior}(\dots)$  en lugar del sumatorio de  $\text{posterior}(\dots)$ ? -> Para calcular cada uno de los  $\text{posterior}$  necesitas la evidencia que se calcula con la suma de todos los  $\text{posterior}$ .
- Paper “A survey on rough set theory and its applications” -> ¿BND(X) and NEG(X) definitions interchanged? Además, duda con respecto a la utilidad que tiene el paper RGNBC
- Paper “RGNBC” -> Página 5 último párrafo -> Debería de hablar sobre la varianza en lugar de sobre la media.
- ¿Error en la gráfica de la página 5 del paper “Classification of Massive ...”? ¿En la segunda también?
- Comentarle a los profesores lo de los diferentes surveys encontrados. Ejemplo: Ensemble (2017).
- No se entiende la función de CFIT en el paper “An effective pattern-based Bayesian classifier for evolving data stream”. Utilización: último párrafo de la página 5.
- Paper “Mining Complex Models from Arbitrarily Large Databases in Constant Time” -> Difficult to understand.
- Preguntar si es necesario que los métodos de clasificación sean también multietiqueta. -> Si
- Preguntar si también interesan papers que traten las redes bayesianas en general para flujos de datos, sin ser clasificadores bayesianos explícitamente.
- Preguntar por el paper que contiene el método denominado Globally Adaptive-MB-MBC (supervisado multidimensional) -> No lo podemos encontrar -> ¿Tiene un paper publicado?
- Método CPL-DS (semi-supervisado unidimensional) -> ¿tiene un paper publicado? -> ¿Classifying evolving data streams with partially labeled data?
- Diferencia entre los artículos de la sección Refereed journals y los de la sección Conference and workshop
- Types of concept drift:
  - Real concept drift (change of the target concept that the classifier is trying to predict) and virtual concept drift (change of the underlying data distribution).
  - According to [14], there are three possible sources of concept drifts: Conditional change (real concept drift), feature change (virtual concept drift) and dual change.

- Sudden drift, gradual drift, incremental drift, recurring concepts -> ¿Rate of change?



Tenerlo en cuenta a la hora de hablar de concept drift

- En la página 4 del paper “Classifying evolving data streams with partially labeled data (2011)” -> otherwise, the window size increases to include the more recent instances. -> ¿No debería ser to include the more out-of-date instances?
- Paper “Classifying evolving data streams with partially labeled data (2011)” -> Página 7 -> ¿Por qué se muestrea solo de la distribución empírica del instante de tiempo  $s$  y no también del instante de tiempo  $s+1$ ?
- Preguntar por las propuestas multi-etiqueta y multi-dimensional. Si hay que nombrarlas. -> Si
- Preguntar si me centro en los papers más recientes.
- Paper “Learning Decision Trees from Data Streams with Concept Drift (2016)” -> CEVOT extends EVO-Tree. This EVO-Tree has one split value in each node. ¿Why not more?

## ANOTACIONES

- Mirar libro para ver las relaciones entre métodos de clustering.
- Ver si el Sampling to Obtain Feasible Centers (apartado) de STREAM se puede comparar con el coresets nombrado en STREAMKM++.
- Comprobar si la carpeta Dynamic, Temporal and Continuous Time Bayesian Networks se puede incluir dentro de la carpeta de Bayesian networks que está dentro de la carpeta Classification for data streams.
- Mirar de las propuestas de clasificadores bayesianos a la hora de redactar el estado del arte relacionado con el Ensemble.
- Paper “MReC-DFS” -> Naive Bayes as a base learner -> Posible paper a introducir dentro de las propuestas de redes bayesianas.
- Una vez leídos bastantes papers -> Buscar menciones de papers dentro de otros papers (por fecha).

- A la hora de implementar un algoritmo -> SIMULAR FLUJO DE DATOS PARTIENDO LOS DATOS EN BLOQUES DE TAL FORMA QUE SIMULAMOS QUE LOS DATOS VAN LLEGANDO A MEDIDA QUE PASA EL TIEMPO.
- Añadir documento “Online Machine Learning in Big Data Streams (2018)” al grupo de surveys a consultar.
- Pedir libro
- Preguntar cómo estructurar la memoria.
- ¿Las referencias que no sean de comparativas hace falta que se ordenen por fecha?
- Ir mandando lo que vaya redactando del TFM.
- En la teoría de las redes bayesianas, ¿tengo que poner mucha notación? ¿O no hace falta?

## REUNIÓN 26 de abril 2019

- Introducir regresión logística en el estado del arte
  - Logistic regression for evolving data streams classification (2007)
- Introducir un apartado de notación.
- Revisar la bibliografía introducida en la memoria (aunque esté en formato bibtex, puede que los títulos no estén bien escritos, con las mayúsculas pertinentes).
- Cambiar apartado de redes bayesianas por clasificadores bayesianos. -> Introducir otro apartado en el mismo nivel que aprendizaje supervisado, no supervisado y series temporales para redes bayesianas para descubrimiento de conocimiento.
- Ver documentos de Javier Diaz para obtener información sobre clustering para data streams. -> Buscar en la página del CIG (2018)
- Cambiar máquinas de soporte vectorial por máquinas de vector soporte.
- A la hora de realizar la comparación entre propuestas, centrarnos en las diferencias en cuanto a metodología más que en los experimentos acerca de si un método es mejor que otro.
- Al haber una revisión de ensemble de 2017, no centrarnos en las propuestas que ya están en esta revisión, sino más en aquellas que no están (comparar las propuestas que no están entre ellas y con otras propuestas que están en los métodos de ensemble).
- Añadir en cada una de las comparaciones entre propuestas que tratan un mismo algoritmo de aprendizaje automático para flujos de datos una tabla resumen como las que se encuentra en alguna de las revisiones (por ejemplo si manejan el concept drift).
- Pedir TFM a alguna persona del grupo de investigación.
- Poner poco de teoría de cada uno de los algoritmos de aprendizaje automático.
- Con respecto a las propuestas que traten el aprendizaje semi-supervisado, al ser pocas, meterlas dentro de las demás propuestas.

