Maestría en Ingeniería Matemática
Propuesta de Tesis

FACULTAD DE INGENIERIA

Título de la propuesta

Comparación de particiones

Identificación de los proponentes

Mathias Bourel, Instituto de Matemática y Estadística Prof. Rafael Laguardia, Facultad de Ingeniería.
Badih Ghattas, Institut de Mathématiques de Marseille, Université d'Aix-Marseille, France.

Contacto: mbourel@fing.edu.uy

Área Temática

Estadística / Aprendizaje Automático.

Perfil esperado del estudiante

El estudiante deberá tener solidas bases en Probabilidad y Estadística, en particular en Análisis Multivariado, y en lenguaje de programación R.

Resumen

En este proyecto nos interesamos a los índices de comparación de particiones y a sus propiedades matemáticas. El objetivo es de implementar y de verificar un test de hipótesis entre dos o más particiones del mismo espacio de datos obtenidas a partir de un método de aprendizaje no supervisado.
Existen actualmente varios índices para comparar dos particiones y los mismos se basan en comparar pares de individuos en cada una de ellas. Podemos citar por ejemplo el índice de Rand, el índice de Rand ajustado, el índice de Jaccard, y otros índices que no siempre se pueden considerar para cualquier tipo de partición.
Un primer objetivo es elaborar un estado del arte en cuanto a las propiedades estadísticas de estos índices. Queremos después considerar un índice de clasificación erronea definido de la manera siguiente.

Notamos las clases obtenidas en la primera partición por $r = 1, \ldots, R$, las clases obtenidas en la segunda partición por $s = 1, \ldots, S$, por $y_1, \cdots, y_n$ la etiqueta de cada observación en la primera partición y por $\hat{y}_1, \cdots, \hat{y}_n$ la etiqueta de cada observación en la segunda. Si Σ es el conjunto de permutaciones de $\{1, \cdots, S\}$ y $\mathcal{A}$ el conjunto de arreglos de S elementos en $\{1, \cdots, R\}$, el índice de clasificación erronea de la segunda partición respecto de la primera es:

$$MCE = \begin{cases} \underset{\sigma \in \Sigma}{\text{Min}} \ \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{y_i \neq \sigma(\hat{y}_i)\}} & \text{si } S \leq R \\ \underset{\sigma \in \mathcal{A}}{\text{Min}} \ \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{\sigma(y_i) \neq \hat{y}_i\}} & \text{si no} \end{cases}$$

Nos proponemos estudiar las propiedades estadísticas de este índice, y en particular su distribución. Por otro lado procederemos a realizar varias simulaciones con distintos parámetros experimentales: tamaño de la muestra, clases balanceadas y desbalanceadas, cantidad de clases, etc.

References

[Fowlkes and Mallows(1983)] E. B. Fowlkes and C. L. Mallows. A Method for Comparing Two Hierarchical Clusterings. *Journal of the American Statistical Association*, 78(383):553–569, 1983.

[Hubert and Arabie(1985)] Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of Classification*, 2(1):193–218–218, 1985.

[Meila(2007)] M. Meila. Comparing clusterings - an information based distance. *Journal of Multivariate Analysis*, 98(5):873–895, 2007.

[Meila(2003)] Marina Meila. Comparing Clusterings by the Variation of Information. pages 173–187. 2003.

[Meila(2005)] Marina Meila. Comparing clusterings: an axiomatic view. In *ICML '05: Proceedings of the 22nd international conference on Machine learning*, pages 577–584, New York, NY, USA, 2005. ACM.

[Mitchell and Mancoridis(2001)] Brian S. Mitchell and Spiros Mancoridis. Comparing the Decompositions Produced by Software Clustering Algorithms Using Similarity Measurements. In *ICSM*, pages 744–753, 2001.

[Pfitzner et al.(2009)Pfitzner, Leibbrandt, and Powers] Darius Pfitzner, Richard Leibbrandt, and David Powers. Characterization and evaluation of similarity measures for pairs of clusterings. *Knowledge and Information Systems*, 19(3):361–394, 2009.

[Youness and Saporta(2010)] Genane Youness and Gilbert Saporta. Comparing partitions of two sets of units based on the same variables. *Advances in Data Analysis and Classification*, 2010.