Título: NEW INTERNAL AND EXTERNAL VALIDATION INDICES FOR CLUSTERING IN BIG DATA

Nombre: Luna Romera, José María

Universidad: Universidad de Sevilla

Departamento: Lenguajes y sistemas informáticos

Fecha de lectura: 17/10/2019

Mención a doctor europeo: concedido

Programa de doctorado: Programa de Doctorado en Ingeniería Informática por la Universidad de Sevilla

Dirección:

> **Director:** María del Mar Martínez Ballesteros

> **Director:** Jorge García Gutiérrez

> **Tutor/Ponente:** JOSÉ CRISTÓBAL RIQUELME SANTOS

Tribunal:

> **presidente:** KARINA GIBERT OLIVERAS

> **secretario:** Cristina Rubio Escudero

> **vocal:** Francisco Martínez Álvarez

> **vocal:** JOSÉ ANTONIO TROYANO JIMENEZ

> **vocal:** Alberto Cano Rojas

Descriptores:

> ANALISIS DE DATOS

> INTELIGENCIA ARTIFICIAL

El fichero de tesis ya ha sido incorporado al sistema

> https://hdl.handle.net/11441/90302

Localización: ESCUELA INTERNACIONAL DE DOCTORADO (EIDUS)

Resumen: Esta tesis, presentada como un compendio de artículos de investigación, analiza el concepto de índices de validación de clustering y aporta nuevas medidas de bondad para conjuntos de datos que podrían considerarse Big Data debido a su volumen. Además, estas medidas han sido aplicadas en proyectos reales y se propone su aplicación futura para mejorar algoritmos de clustering.

El clustering es una de las técnicas de aprendizaje automático no supervisado más usada. Esta técnica nos permite agrupar datos en clusters de manera que, aquellos datos que pertenezcan al mismo cluster tienen características o atributos con valores similares, y a su vez esos datos son disimilares respecto a aquellos que pertenecen a los otros clusters. La similitud de los datos viene dada normalmente por la cercanía en el espacio, teniendo en cuenta una función de distancia. En la literatura existen los llamados índices de validación de

clustering, los cuales podríamos definir como medidas para cuantificar la calidad de un resultado de clustering. Estos índices se dividen en dos tipos: índices de validación internos, que miden la calidad del clustering en base a los atributos con los que se han construido los clusters; e índices de validación externos, que son aquellos que cuantifican la calidad del clustering a partir de atributos que no han intervenido en la construcción de los clusters, y que normalmente son de tipo nominal o etiquetas.

En esta memoria se proponen dos índices de validación internos para clustering basados en otros índices existentes en la literatura, que nos permiten trabajar con grandes cantidades de datos, ofreciéndonos los resultados en un tiempo razonable. Los índices propuestos han sido testeados en datasets sintéticos y comparados con otros índices de la literatura. Las conclusiones de este trabajo indican que estos índices ofrecen resultados muy prometedores frente a sus competidores.

Por otro lado, se ha diseñado un nuevo índice de validación externo de clustering basado en el test estadístico chi cuadrado. Este índice permite medir la calidad del clustering basando el resultado en cómo han quedado distribuidos los clusters respecto a una etiqueta dada en la distribución. Los resultados de este índice muestran una mejora significativa frente a otros índices externos de la literatura y en datasets de diferentes dimensiones y características.

Además, estos índices propuestos han sido aplicados en tres proyectos con datos reales cuyas publicaciones están incluidas en esta tesis doctoral. Para el primer proyecto se ha desarrollado una metodología para analizar el consumo eléctrico de los edificios de una smart city. Para ello, se ha realizado un análisis de clustering óptimo aplicando los índices internos mencionados anteriormente. En el segundo proyecto se ha trabajado tanto los índices internos como con los externos para realizar un análisis comparativo del mercado laboral español en dos periodos económicos distintos. Este análisis se realizó usando datos del Ministerio de Trabajo, Migraciones y Seguridad Social, y los resultados podrían tenerse en cuenta para ayudar a la toma de decisión en mejoras de políticas de empleo. En el tercer proyecto se ha trabajado con datos de los clientes de una compañía eléctrica para caracterizar los tipos de consumidores que existen. En este estudio se han analizado los patrones de consumo para que las compañías eléctricas puedan ofertar nuevas tarifas a los consumidores, y éstos puedan adaptarse a estas tarifas con el objetivo de optimizar la generación de energía eliminando los picos de consumo que existen la actualidad.