



Competición Vídeo: ¿Cómo puedo saber si el resultado de un clustering es lo suficientemente bueno?

José María Luna-Romera and José C. Riquelme Santos
Departamento de Lenguajes y Sistemas Informáticos
Universidad de Sevilla
Sevilla, España

Abstract—En este vídeo divulgativo se hace una introducción a una de las técnicas de análisis de datos más usada, el clustering. En él, se explica a través de sencillos ejemplos qué es el clustering y cómo se analizan las distintas soluciones que ofrece a través de diferentes índices de validación.

Index Terms—Análisis de clustering, validación de clustering

I. HERRAMIENTAS Y MATERIALES

Para la elaboración del vídeo se han usado dos herramientas principalmente: Microsoft Power Point, para la elaboración de las diferentes pantallas incluyendo las animaciones de las figuras; Pinnacle Studio [1], que es un programa de edición de vídeo que nos ha permitido añadir las pistas de audio, así como controlar los tiempos de una manera más cómoda a Power Point.

II. LINK DEL VÍDEO

<https://youtu.be/ERDjUM78H1Q>

III. JUSTIFICACIÓN

El tema principal del vídeo es la validación de clustering. Se ha escogido esta técnica con el fin de acercar a una audiencia no especializada el clustering, una de las técnicas de análisis de datos más usada y nuclear en el aprendizaje no supervisado. Además, con el objetivo de profundizar en el tema, se da una visión de los índices de validación de clustering para motivar al espectador a plantearse algunas preguntas y cómo llegar a responderlas. A pesar del amplio uso del clustering, sus técnicas de validación son bastante desconocidas, si las comparamos, por ejemplo, a la amplia literatura sobre cómo acreditar la bondad de un clasificador.

La validación del clustering es una de las cuestiones más importantes en el análisis clustering. De hecho, se podría afirmar que medir la calidad de una solución de clustering es tan importante como el método de clustering en sí [2]. Para ello, en la literatura se definen índices de validación de clustering (CVI) que miden la calidad de la solución. Estas medidas podrían clasificarse en CVI internos y externos [3], [4]. Los CVI internos se basan en cómo se distribuyen los objetos en los distintos clusters utilizando sólo la distancia entre los objetos. Por otro lado, los índices externos se apoyan en posibles etiquetas de los objetos que puedan proporcionar

algún tipo de información externa para medir la calidad del clustering [5].

IV. RECURSOS

En primera instancia se quiso realizar el vídeo usando algún tipo de herramienta que facilitara la elaboración de material didáctico por medio de animaciones. Existen en el mercado servicios online como VideoScribe [6], Doodly [7] o Animaker [8], que a través de una sencilla e intuitiva interfaz permiten realizar vídeos con animaciones, figuras, iconos e incluso personajes. Todos estos servicios disponen de un periodo de prueba en el cual puedes probar la herramienta e incluso algunas de ellas permiten la exportación del vídeo a baja resolución. El inconveniente que encontramos al usar estas herramientas es que al tratarse de versiones de prueba, solo algunas permitían exportar el vídeo y lo hacían dejando una marca de agua.

Como se ha comentado en el apartado anterior, finalmente se ha usado Microsoft Power Point para elaboración de las diferentes pantallas que van apareciendo en el vídeo y Pinnacle Studio para añadir las pistas de audio y realizar el control de tiempos de cada pantalla.

V. UTILIDAD

Se ha elaborado un vídeo que permite a una audiencia no especializada introducirse en una de las técnicas de aprendizaje no supervisado más usadas de una manera fácil de entender. Además, se ha querido dar un paso más, presentando un nuevo concepto como puede ser la validación del clustering, y qué tipos de técnicas existen. Todo ello se muestra a través de sencillos ejemplos y se ilustra de manera gráfica qué es el clustering y cómo se puede validar la solución.

El vídeo ha sido visualizado por personas con desconocimiento total de la materia, y aunque ha habido diferentes valoraciones, por lo general han sido positivas, destacando la facilidad del lenguaje empleado y el acercamiento de algo complejo a través de ejemplos gráficos.

VI. CONCLUSIONES

Se ha elaborado un vídeo divulgativo de corta duración para acercar a una audiencia general, una técnica fundamental en el análisis de datos. El vídeo incluye una explicación



general del clustering además de dos tipos de técnicas de validación. A pesar de ser el clustering una técnica muy usada en múltiples ámbitos, no es muy abundante los trabajos divulgativos sobre su validación, a diferencia de las técnicas de aprendizaje supervisado, con mucha mayor presencia en ámbitos científicos.

REFERENCES

- [1] P. Studio, “Pinnacle Studio: software de edición de vídeo y grabación de pantalla,” <https://www.pinnaclesys.com/es/products/studio/>, 2018, [Online; accessed september-2018].
- [2] A. K. Jain, “Data clustering: 50 years beyond K-means,” *Pattern Recognition Letters*, vol. 31, no. 8, pp. 651–666, jun 2010. [Online]. Available: <http://linkinghub.elsevier.com/retrieve/pii/S0167865509002323>
- [3] O. Arbelaitz, I. Gurrutxaga, J. Muguerza, J. M. Pérez, and I. Perona, “An extensive comparative study of cluster validity indices,” *Pattern Recognition*, vol. 46, no. 1, pp. 243 – 256, 2013. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S003132031200338X>
- [4] M. Halkidi, Y. Batistakis, and M. Vazirgiannis, “On clustering validation techniques,” *Journal of Intelligent Information Systems*, vol. 17, no. 2, pp. 107–145, Dec 2001. [Online]. Available: <https://doi.org/10.1023/A:1012801612483>
- [5] Y. Lei, J. C. Bezdek, S. Romano, N. X. Vinh, J. Chan, and J. Bailey, “Ground truth bias in external cluster validity indices,” *Pattern Recognition*, vol. 65, pp. 58 – 70, 2017. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0031320316303910>
- [6] VideoScribe, “VideoScribe - Whiteboard Animation Software | Whiteboard Video Maker,” <https://www.videoscribe.co/>, 2018, [Online; accessed september-2018].
- [7] Doodly, “Doodly - Easily Create Whiteboard Doodle Videos In Minutes!” <https://www.doodly.com/>, 2018, [Online; accessed september-2018].
- [8] Animaker, “Animaker, Crea video-animaciones en la nube gratis,” <https://www.animaker.es>, 2018, [Online; accessed september-2018].