

Aprendizaje no supervisado combinando autoencoders y agrupamiento

Lucas Cochella¹ [0009-0007-8333-8080], Laura Lanzarini^{1,2} [0000-0001-7027-7564],
and Patricia Jimbo Santana³ [0000-0001-7432-1622]

Facultad de Informática. Universidad Nacional de La Plata

lucas.cochella@econo.unlp.edu.ar

Instituto de Investigación en Informática LIDI (Centro CICPBA)

laural@lidi.info.unlp.edu.ar

Facultad de Ciencias Administrativas Universidad Central del Ecuador

prjimbo@uce.edu.ec

Resumen Las técnicas de agrupamiento son las más utilizadas a la hora de resolver tareas descriptivas. Su entrenamiento no supervisado requiere del uso de alguna medida de distancia para identificar similitudes entre los datos disponibles. La cantidad de características utilizadas suele tener un impacto directo en la calidad del agrupamiento por lo que contar con una representación adecuada de los datos es de sumo interés. Un área de investigación reciente en el aprendizaje no supervisado es la combinación del aprendizaje de representación con redes neuronales profundas y la agrupación de datos. Este artículo presenta una línea de investigación relacionada con el estudio, diseño e implementación de un modelo que combine autoencoders con distintas técnicas de agrupamiento para analizar datos estructurados con alta dimensionalidad. Si bien los autoencoders han demostrado ser efectivos para aprender representaciones adecuadas a la hora de procesar imágenes, aún quedan aspectos por resolver cuando se trata de trabajar con datos estructurados. El ámbito de aplicación son los datos académicos de los estudiantes de la Universidad Central del Ecuador entre 2018 y 2022. Estos datos pueden contener información valiosa sobre el desempeño y la evolución de los estudiantes, así como posibles casos de error, fraude o corrupción que afecten a la calidad educativa. El agrupamiento basado en la representación aprendida puede ayudar a identificar estos casos, así como a descubrir patrones y tendencias que permitan mejorar el sistema educativo.

Keywords: Autoencoders · Agrupamiento · Reducción de Dimensionalidad · Deep Learning.

1. Introducción

El aprendizaje no supervisado es una modalidad de aprendizaje automático que busca extraer conocimiento de los datos sin contar con etiquetas o clases previas. Una de sus aplicaciones más frecuentes es el agrupamiento de datos, que consiste en asignar los datos a grupos o clústeres según algún criterio de similitud o distancia.

Para realizar el agrupamiento se requiere una representación adecuada de los datos que capture las características relevantes para la tarea sin incrementar excesivamente la dimensión del espacio de entrada. Una forma de obtener esta representación es mediante el uso de redes neuronales profundas, que son modelos computacionales inspirados en

el funcionamiento del cerebro humano, capaces de aprender representaciones complejas y no lineales a partir de los datos.

Dentro de las redes neuronales profundas, los autoencoders son redes que se entrenan para reconstruir los datos de entrada a partir de una representación interna más compacta y significativa. Esta representación se puede utilizar luego para realizar el agrupamiento con distintas técnicas.

Este trabajo se enmarca en un proyecto de Tesis de culminación de la Maestría en Inteligencia de Datos orientada a Big Data de la Facultad de Informática en la Universidad Nacional de La Plata y tiene por objetivo analizar el potencial de la combinación de métodos de agrupamiento no supervisados con técnicas de deep learning como los diferentes tipos de autoencoders. Como resultado de esta investigación se espera aportar una solución al problema de agrupar datos estructurados con alta dimensionalidad mediante la reducción que ofrecen los modelos de autoencoders. A continuación se presenta una revisión bibliográfica sobre los temas a analizar y una breve descripción del caso de estudio.

2. Trabajos relacionados

El aprendizaje profundo es una técnica basada en redes neuronales artificiales capaz de aprender representaciones complejas y no lineales a partir de los datos. En este trabajo, se propone combinar el aprendizaje de representación con autoencoders y las técnicas de agrupamiento para analizar el progreso académico de los estudiantes universitarios a partir de sus registros personales.

En relación a la aplicación de algoritmos de agrupamiento, que son métodos no supervisados que tienen como objetivo dividir un conjunto de datos en grupos o clústeres según algún criterio de similitud o distancia, encontramos a Ahmad Fikri Mohamed Nafuri et al (2022) [1] que analiza los datos comparando la performance de K-Means, BIRCH y DBSCAN realizando una selección de los atributos preponderantes de los datos con el fin de evitar las complicaciones que trae la multidimensionalidad al análisis. Valarmathy & Krishnaveni [2] por su parte van un paso más allá y realizan agrupamiento de datos comparando el rendimiento de 8 técnicas como Expectation Maximization, CLOPE, DBSCAN, entre otras.

Por otra parte, los autoencoders son un tipo especial de redes neuronales profundas que pueden aprender una representación comprimida y reconstruir los datos originales a partir de ella. Dentro de los trabajos relacionados con esta propuesta podemos mencionar a Tschannen et al. (2018) [3] quienes realizan una revisión en profundidad de los avances recientes en el aprendizaje de representación con autoencoders, y analizan el aprendizaje de representación desde la perspectiva de la teoría de la tasa-distorsión. Huang et al. (2023) [4] propone un método de agrupamiento profundo basado en un autoencoder embebido, que incorpora un autoencoder en la unidad codificadora y decodificadora del autoencoder prototipo, experimentan con múltiples conjuntos de datos de imágenes y muestran que su método tiene un rendimiento superior al de otros métodos existentes. Lu & Li (2022) [5] presentan un marco generalizado de aprendizaje de representación basado en autoencoders para el

agrupamiento de datos, y analizan diferentes tipos de datos. Muestran que su enfoque mejora el rendimiento del algoritmo K-Means.

El trabajo de Kanazawa et al. (2019) [6] se basa en un estudio sobre el análisis de agrupamiento de características basado en la representación de la capa oculta de un autoencoder. Aplican su método a datos sintéticos y reales, y evalúan la calidad del agrupamiento mediante el índice Davies-Bouldin. En el trabajo de Jimbo et al (2023) [7] se realiza una revisión sistemática de la literatura, considerando las investigaciones que se han desarrollado utilizando técnicas de inteligencia artificial para analizar el rendimiento académico en instituciones de educación superior. Aquí se identifican las principales técnicas utilizadas, como minería de datos, aprendizaje automático, sistemas expertos o redes neuronales, así como las variables y los indicadores empleados para medir el rendimiento académico.

Finalmente, Florent Forest, et al (2021) [8], combina la aplicación de la técnica de aprendizaje no supervisado Self Organizing Map con la aplicación de autoencoders para reducir la dimensionalidad de los datos obteniendo resultados alentadores mediante la implementación de DESOM en datasets compuestos por imágenes y texto.

3. Caso de estudio

Como caso de estudio, se propone analizar los datos de los estudiantes de la Universidad Central del Ecuador (UCE) entre 2018 y 2022.

La UCE es la universidad más antigua y la segunda más grande por número de estudiantes de la República del Ecuador; cuenta con 21 facultades, 66 carreras, 11 institutos y 4 centros, lo que implica una gran diversidad y complejidad en sus procesos académicos. Además, la UCE tiene como misión contribuir al desarrollo científico, tecnológico, cultural y social del país, mediante la formación integral de profesionales competentes, críticos y comprometidos con la sociedad. Por lo tanto, el análisis de los datos de los estudiantes de la UCE puede aportar información valiosa para mejorar la gestión académica, detectar posibles casos de error, fraude o corrupción, y garantizar el cumplimiento de los objetivos institucionales.

Las investigaciones realizadas en relación al aprendizaje de representación con autoencoders y las técnicas de agrupamiento serán aplicadas al análisis del progreso académico de los estudiantes universitarios a partir de sus registros personales.

El objetivo es aprender una transformación de los datos originales a un espacio de menor dimensión, que preserve o resalte las características relevantes y luego aplicar diferentes técnicas de agrupamiento para obtener grupos significativos y coherentes de estudiantes. El análisis de los datos de los estudiantes puede servir para identificar error, fraude y corrupción en los procesos del vicerrectorado académico en las instituciones de educación superior, ya que puede revelar patrones anómalos, inconsistencias o irregularidades en los procesos académicos y administrativos que afectan a la calidad y la equidad de la educación.

Referencias

1. Mohamed Nafari, A.F.; Sani, N.S.; Zainudin, N.F.A.; Rahman, A.H.A.; Aliff, M. Clustering Analysis for Classifying Student Academic Performance in Higher Education. *Appl. Sci.* 2022, 12, 9467.
2. Valarmathy N., Krishnaveni, S. (2019) Performance Evaluation and Comparison of Clustering Algorithms used in Educational Data Mining. *International Journal of Recent Technology and Engineering (IJRTE)*, 7 (655). ISSN 2277-3878.
3. Tschannen, M., Bachem, O., Lucic, M. (2018). Recent Advances in Autoencoder-Based Representation Learning. arXiv preprint arXiv:1812.05069.
4. Huang, X., Hu, Z., Lin, L. (2023). Deep clustering based on embedded auto-encoder. *Soft Computing*, 27, 1075–1090.
5. Lu, S., Li, R. (2022). DAC–Deep Autoencoder-Based Clustering: A General Deep Learning Framework of Representation Learning. In: Arai, K. (eds) *Intelligent Systems and Applications. IntelliSys 2021. Lecture Notes in Networks and Systems*, vol 294. Springer, Cham.
6. Kanazawa, S., Sugiyama, Y., Yang, T., Goto, M. (2019). A Study of Feature Clustering Analysis based on the Hidden Layer Representation of an Autoencoder. *Total Quality Science*, 5(1), pp.1-10.
7. Jimbo-Santana, P., Lanzarini, L., Jimbo-Santana, M., Morales-Morales, M. (2023). Inteligencia artificial para analizar el rendimiento académico en instituciones de educación superior. Una revisión sistemática de la literatura. *Cátedra*, 6(2), pp.30–50.
8. Forest F., Lebbah M., Azzag H., Lacaille J. (2021). Deep embedded self-organizing maps for joint representation learning and topology-preserving Clustering. *Neural Computing and Applications*, 33(24), pp.1-31