New Trends in Mathematical Sciences
http://dx.doi.org/10.20852/ntmsci.2022.465

Data clustering and its applications in medicine

Mantas Lukauskas and Tomas Ruzgas

Kaunas University of Technology, Kaunas, Lithuania

Received: 25 December 2021 , Accepted: 31 January 2022
Published online: 13 March 2022

Abstract: Artificial intelligence was first mentioned back in 1956, but the biggest leap in its use has been seen in the last two decades. It goes without saying that with the increasing availability of artificial intelligence, one of the most important areas in which it must be applied is medicine. The purpose of this short article is to provide an overview of one of the groups of machine learning, by reviewing clustering algorithms and also by reviewing the use in medicine. For an excellent interpretation, this can usually be done by finding certain groups in your data that are much easier to interpret than individual observations. This task can be solved using a clustering. In medicine, the application of clustering allows one to distinguish various groups of patients and to summarize these groups to provide much more precise recommendations. Different clustering techniques are used to identify breast cancer, Parkinson's disease, migraine, various psychological and psychiatric disorders, heart and diabetes diseases, Huntington's disease, and Alzheimer's disease, among others. Looking at all the examples, it can be seen that even when evaluating only one area, clustering is applied particularly widely. Every year, more and more different research is carried out, which allows us to apply these algorithms in medicine and helps doctors assess the situation and start treatment faster or even more accurately. It can be assumed that only a greater use of AI and clustering in medicine will be observed in the coming years.

Keywords: Machine learning, artificial intelligence, clustering, medicine.

1 Introduction

Artificial intelligence was first mentioned back in 1956, but the biggest leap in its use has been seen in the last two decades. Such high availability, development, and application of artificial intelligence have been driven by the ever-increasing power of possible computations. Artificial intelligence has found use in a variety of fields: manufacturing, marketing, education, and medicine. It goes without saying that with the increasing availability of artificial intelligence, one of the most important areas in which it must be applied is medicine. Using these methods makes it much easier to analyze large amounts of data, automate various processes, and provide treatment recommendations. These technologies can quickly learn, analyze, predict, and present final conclusions, often without human intervention. The use of artificial intelligence in medicine is particularly wide and includes areas such as planning, imaging, natural language recognition, and others. The purpose of this short article is to provide an overview of one of the groups of artificial intelligence / machine learning, clustering, by reviewing clustering algorithms and also by reviewing the use of clustering in medicine.

2 Machine learning algorithm

This section describes the types of machine learning and the algorithms of one of the groups of machine learning, clustering.

* Corresponding author e-mail: mantas.lukauskas@ktu.lt

2.1 Types of machine learning

Machine learning is a field of computer science which aims to "teach" a computer to recognize certain patterns and to find connections between individual data. Sometimes, after viewing the data, we cannot interpret the extract information from the data. In the case of classical programming, rules and data are provided to the computer to obtain a final answer. Compared to classical programming in the case of machine learning, it is not the answer that is obtained but certain rules that can be applied to new data. There are usually four main groups of machine learning methods: unsupervised, stimulating, supervised learning, and semi-supervised learning. Reinforcement learning is an area of machine learning in

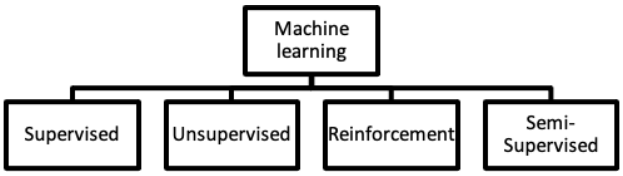

Fig. 1: Types of machine learning

which an agent receives information about the environment and learns to choose actions that maximize the function of the goal. This learning is quite rare in solving real tasks, but it is stated that in the future this teaching method may find a very wide application: autonomous cars, logistics solutions, etc. [1]. Supervised learning is an area of machine learning where the model is created from available past / learning sample data. The learning data set contains data pairs X (variables / observation properties) and Y (class value) [2]. Semi-supervised learning is the branch of machine learning that deals with the use of labeled and unlabeled data to perform certain learning tasks [3]. Finally, the most important class of machine learning discussed in this short article is unsupervised learning. This training is done with data whose class is not known in advance, and these methods are used to try to discover the "hidden" links between the data. It can be observed that due to the previously unknown classes of data, it is not possible to estimate the accuracy of the obtained model (in contrast to the case of supervised learning) [4]. Unattended machine learning methods can be divided into three smaller groups of methods: clustering, dimension reduction, or association rules, but this paper focuses only on clustering algorithms and their application in medicine.

2.2 Clustering algorithms

It is often very important to find hidden information in the data that would be much easier to interpret in practice. For an excellent interpretation, this can usually be done by finding certain groups in your data that are much easier to interpret than individual observations. This task can be solved using an unsupervised machine learning method, clustering. According to Jain [5], the main goal of clustering is to per-form accurate data grouping using observations, points, or objects. Furthermore, the purpose of clustering is to reveal subgroups within heterogeneous data such that each individual group has greater homogeneity than the whole [6]. To measure how close individual observations are to each other, different distance measures and techniques are used to assess the similarity of different observations. Observations between which the distance is the smallest are recognizable as like each other and vice versa. Many different clustering methods are used in practice, so it is important to separate these clustering algorithms into groups according to their working principle. According to Fraley and Raftery [7], data clustering can be distinguished into two main groups: hierarchical clustering and divisional clustering. Han and Kamber [8] propose to classify data clustering differently and to determine the following groups: density-based methods, model-based methods, and grid-based methods. By analyzing scientific articles, it is observed that k-means clustering algorithms are still the most widely used [8,9]. Clustering of K means is used to divide the data into k groups of presets between which a 'hidden' relationship is found that a person may not notice. But this method does not always work properly. This method is usually suitable for distinguishing only observations that are

spherical in shape. There are also many other clustering algorithms that are used in scientific research: UNIC [10], k-Medoids (PAM) [11], Gaussian mixture [11], TCLUST [11], Trimmed k-means [12], Spectral Clustering, Density-Based Spatial Clustering, MULIC, DENCLUE, SOMs (Neural-Net), SVM, HIERDENC, Deep embedded clustering [13] etc. However, all these methods are often applied to specific tasks and are not universal; for this reason, the aim of later section is to introduce different clustering algorithms applications in medicine and how these algorithms can be used in the future.

3 Data clustering application in medicine

This application is comprehensive when evaluating clustering algorithms, as clustering algorithms do not require prior data classes. In medicine, data clustering can also be applied and is applied quite widely. In medicine, certain treatments are often applied to all individuals with the same disease. In this case, the application of clustering allows one to distinguish various groups of patients and to summarize these groups to provide much more precise recommendations [14]. Studies show that clustering algorithms can be applied to identify different diseases [15,16,17,18,19]. For example, different clustering techniques are used to identify breast cancer [20], Parkinson's disease [21,22], migraine [23], various psychological and psychiatric disorders [24], heart and diabetes diseases [25], Huntington's disease [26], and Alzheimer's disease [27], among many others. Given that clustering is an area of unsupervised learning, its application to physicians allows for the observation of certain exceptional cases and their subsequent analysis in much more detail. Mention may also be made of the application of clustering in medicine, where unstructured data is used. In this case, one of the possibilities of applying clustering algorithms is the analysis of various text documents. First, using various clustering algorithms, it is possible to divide various large-volume documents into certain groups. Then, these groups of documents can be summarized, thus avoiding high reading of the documents [28]. Another important example that can be particularly widely applied in medicine is image analysis. Image analysis can allow the application of clustering in a variety of different ways, but the main one is the segmentation of various X-rays, MRIs, and other images, allowing some changes to be observed in these images [29,30,31].

4 Conclusion

The main purpose of this short article was to review what machine learning is, one of its areas is clustering, what clustering algorithms are currently most commonly found in the scientific literature, and how these methods are used in medicine. Clustering can be seen to be only one area of machine learning, so the potential for even greater use of AI arises from the inclusion of other areas as well. Looking at all the examples, it can be seen that even when evaluating only one area, clustering is applied particularly widely. Every year, more and more different research is carried out, which allows us to apply these algorithms in medicine and helps doctors assess the situation and start treatment faster or even more accurately. It can be assumed that only a greater use of AI and clustering in medicine will be observed in the coming years.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

All authors have contributed to all parts of the article. All authors read and approved the final manuscript.

References

[1] Chollet, F. "Deep learning with Python, Vol. 1." Greenwich, CT: Manning Publica-tions CO (2017).

[2] Mohri M., Rostamizadeh A., Talwalkar A. Foundations of Machine Learning // The MIT Press, 2012, ISBN 9780262018258.
[3] Van Engelen, Jesper E., and Holger H. Hoos. "A survey on semi-supervised learning." Machine Learning 109.2 (2020): 373-440.
[4] Duda, Richard O., Peter E. Hart and David G. Stork. "Unsupervised learning and clus-tering." Pattern classification (2001): 517-601.
[5] Jain, Anil K. "Data clustering: 50 years beyond K-means." Pattern recognition letters 31.8 (2010): 651-666.
[6] Fraley, Chris and Adrian E. Raftery. "How many clusters? Which clustering method? Answers via model-based cluster analysis." The Computer Journal 41.8 (1998): 578-588.
[7] Han, Jiawei, Jian Pei, and Micheline Kamber. Data mining: concepts and techniques. Elsevier, 2011.
[8] W.Guogiu. Robust self-tuning spectral clustering. Neurocomputing, 1 (391):243-248, 2020.
[9] K. Sinaga and M. Yang. Unsupervised clustering algorithm of K-means. IEEE Access, 8 (8):80716-80727, 2020.
[10] N.Leopold, O.Rose. A fast nonparametric clustering. Pattern Recognition, 100 (100):107-117, 2020.
[11] A. E. Attar, R. Khatoun, B. Birregah, and M. Lemercier. Article in proceedings. In: Ro-bust clustering methods for detecting smartphone abnormal behavior, IEEE 2014 Wireless Communications and Networking Conference (WCNC), 2014, 2552-2557. (Istanbul, Turkey)
[12] Cuesta-Albertos, Juan Antonio, Alfonso Gordaliza, and Carlos Matrán. "Trimmed *k*-means: An attempt to robustify quantizers." The Annals of Statistics 25.2 (1997): 553-576.
[13] R. Yazhou et al. Semi-supervised deep embedded clustering. Neurocomputing, 325 (325):121-130, 2019
[14] Nezhad et al. "SUBIC: A supervised bi-clustering approach for precision medicine." 2017 IEEE 16th International Conference on Machine Learning and Applications (ICMLA). IEEE, 2017.
[15] Nugent, R., Meila, M. (2010). An overview of clustering applied to molecular biology, in Statistical Methods in Molecular Biology, eds Bang H. and Zhou X. K., van Epps H. L., Mazumdar M. (Springer;), 369-404.
[16] Li X., Zhu F. (2013). On clustering algorithms for biological data. Engineering 5:549 10.4236/eng.2013.510B113
[17] Nithya N., Duraiswamy K., Gomathy P. (2013). A survey on clustering techniques in medical diagnosis. Int. J. Comput. Sci. Trends Technol. 1, 17-23.
[18] Wiwie C., Baumbach J., Rottger R. (2015). Comparison of the performance of biomed-ical clustering methods. Nat. Methods 12:1033. 10.1038/nmeth.3583
[19] Chen, C.-H. (2014). A hybrid intelligent model of analysing clinical breast cancer data using clustering techniques with feature selection. Appl. Soft. Comput. 20, 4-14. 10.1016/j.asoc.2013.10.024
[20] Polat K. (2012). Classification of Parkinson's disease using feature weighting method on the basis of fuzzy 366 c-means clustering. Int. J. Syst. Sci. 43, 597-609. 10.1080/00207721.2011.581395
[21] Nilashi M., Ibrahim O., Ahani A. (2016). Accuracy improvement for predicting Parkinson's disease progression. Sci. Rep. 6:34181. 10.1038/srep34181
[22] Wu Y., Duan H., Du S. (2015). Multiple fuzzy c-means clustering algorithm in medical diagnosis. Technol. Health Care 23, S519-S527. 10.3233/THC-150989
[23] Trevithick, L., Painter, J., Keown, P. (2015). Mental health clustering and diagnosis in psychiatric inpatients. BJPsych Bull. 39, 119-123. 10.1192/pb.bp.114.047043
[24] Yilmaz N., Inan O., Uzer M. S. (2014). A new data preparation method based on clus-tering algorithms for diagnosis systems of heart and diabetes diseases. J. Med. Syst. 38:48. 10.1007/s10916-014-0048-7
[25] Nikas J. B., Low W. C. (2011). Application of clustering analyses to the diagnosis of Huntington's disease in mice and other diseases with well-defined group boundaries. Comput. Methods Programs Biomed. 104, e133-e147. 10.1016/j.cmpb.2011.03.004
[26] Alashwal et al. 'The Application of Unsupervised Clustering Methods to Alzheimer's Disease.' Frontiers in computational neuroscience vol. 13 31. 24 May. 2019, doi:10.3389/fncom.2019.00031.
[27] Alashwal, Hany, et al. "The application of unsupervised clustering methods to Alz-heimer's disease." Frontiers in computational neuroscience 13 (2019): 31.
[28] Renganathan, Vinaitheerthan. "Text mining in biomedical domain with emphasis on document clustering." Healthcare informatics research 23.3 (2017): 141-146.
[29] Suetens, Paul, et al. "Image segmentation: methods and applications in diagnostic ra-diology and nuclear medicine." European journal of radiology 17.1 (1993): 14-21.
[30] Boudraa, Abdel-Ouahab, and Habib Zaidi. "Image segmentation techniques in nuclear medicine imaging." Quantitative analysis in nuclear medicine imaging. Springer, Bos-ton, MA, 2006. 308-357.
[31] Qu, Panling, et al. "Automatic tongue image segmentation for traditional chinese med-icine using deep neural network." International Conference on Intelligent Computing. Springer, Cham, 2017.