International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 04 | April - 2024 SJIF Rating: 8.448 ISSN: 2582-3930

Football Player Classification

Idhika Suri, Aayush Sharma, Battula Vinay Kumar

Abstract

Clustering is an essential tool in data mining research and applications. It is the subject of active research in many fields of study, such as computer science, data science, statistics, pattern recognition, artificial intelligence, and machine learning. Several clustering techniques have been proposed and implemented, and most of them successfully find excellent quality or optimal clustering results in the domains mentioned earlier. However, there has been a gradual shift in the choice of clustering methods among domain experts and practitioners alike, which is precipitated by the fact that most traditional clustering algorithms still depend on the number of clusters provided a priori. These conventional clustering algorithms cannot effectively handle real-world data clustering analysis problems where the number of clusters in data objects cannot be easily identified. Also, they cannot effectively manage problems where the optimal number of clusters for a high-dimensional dataset cannot be easily determined. Therefore, there is a need for improved, flexible, and efficient clustering techniques. Recently, a variety of efficient clustering algorithms have been proposed in the literature, and these algorithms produced good results when evaluated on real-world clustering problems. This study presents an up-to-date systematic and comprehensive review of traditional and state-of-the-art clustering techniques for different domains. This survey considers clustering from a more practical perspective. It shows the outstanding role of clustering in various disciplines, such as education, marketing, medicine, biology, and bioinformatics. It also discusses the application of clustering to different fields attracting intensive efforts among the scientific community, such as big data, artificial intelligence, and robotics. This survey paper will be beneficial for both practitioners and researchers. It will serve as a good reference point for researchers and practitioners to design improved and efficient state-of-the-art clustering algorithms.

Introduction

Clustering (an aspect of data mining) is considered an active method of grouping data into many collections or clusters according to the similarities of data points features and characteristics (Jain, 2010, Abualigah, 2019). Over the past years, dozens of data clustering techniques have been proposed and implemented to solve data clustering problems (Zhou et al., 2019, Abualigah et al., 2018a, Abualigah et al., 2018b). In general, clustering analysis techniques can be divided into two main groups: hierarchical and partitional (Tan, 2018). Although methods in these two groups have proved to be very effective and efficient, they generally depend on providing prior knowledge or information of the exact number of clusters for each dataset to be clustered and analyzed (Chang et al., 2010). More so, when dealing with real-world datasets, it is normal not to expect or have any prior information regarding the number of naturally occurring groups in the data objects (Liu et al., 2011). Therefore, the concept of automatic data clustering algorithms is introduced to address this limitation. Automatic clustering algorithms refer to any clustering techniques used to automatically determine the number of clusters without having any prior information of the dataset features and attributes (Ezugwu, 2020a). Many automatic data clustering algorithms have been proposed in the literature, and several of them are nature-inspired. The current survey presents a systematic study of traditional and recently proposed clustering techniques applied in different fields.

Many surveys on clustering techniques exist in the literature (Xu and Wunsch, 2005, Xu and Tian, 2015, Benabdellah et al., 2019, Adil et al., 2014, Dafir et al., 2021; Saxena et al., 2017, Nagpal, 2013, Oyelade et al., 2016, Bindra and Mishra, 2017, Singh and Srivastava, 2020, Djouzi and Beghdad-Bey, 2019, Ezugwu, 2020a). Xu and Tian (2015) explained the basic elements involved in the clustering process and broadly categorized existing clustering algorithms into two major perspectives: the traditional and modern ones. Xu and Wunsch (2005) reviewed major clustering algorithms for datasets appearing in Statistics, Computer Science, and Machine learning. Benabdellah et al. (2019) categorized clustering algorithms using the three V's properties of Big Data: Volume, Variety, and Velocity. These three properties were used to explore the various categories of clustering algorithms. Adil et al. (2014) gave a concise survey of existing clustering algorithms and conducted extensive experiments to highlight the best-performing clustering algorithm for Big data analysis. Berkhin et al. (2001) reviewed clustering techniques in data mining, emphasizing object attribute type, large dataset scalability, handling high dimensional data, and finding irregularly shaped clusters.

Dafir et al. (2021)'s work was on parallel clustering algorithms, classifying and summarizing them. He discussed the framework for each kind of parallel clustering algorithm. Saxena et al. (2017) presented a taxonomy of existing clustering algorithms, debating each algorithm's various measures of similarity and evaluation criteria. Nagpal (2013) carried out a comparative analysis of the different clustering algorithms concerning both the mixed and categorical datasets with the observation that no clustering algorithm can be adjudged as best for handling a large dataset of either the mixed or categorical dataset. Oyelade et al. (2016) examined various clustering algorithms and their suitability for gene expression data to discover and provide helpful knowledge that will guarantee stability and

a high degree of accuracy in the area. Jain (2010) summarized well-known clustering methods with a discussion on critical issues and challenges in the design of clustering algorithms. Jain et al. (1999) discussed emerging techniques for non-numeric constraints and large sets of patterns. Ezugwu et al. (2020a) presented an in-depth and systematic review of nature-inspired metaheuristic algorithms used for automatic clustering analysis focusing on the metaheuristic algorithms that have been employed to solve clustering problems over the last three decades.

Obviously, from the literature, there has been a considerable growth of interdisciplinary interests and dynamics in the application of clustering analysis to different research domains indicating that without a doubt, much has been achieved regarding clustering with new emerging research directions in automatic clustering algorithms (Ezugwu et al., 2020a). However, despite the decades of reported research on clustering methods and algorithms, the existing literature is remarkably segmented. Moreover, applied researchers find it challenging to acquire systematic information on research progress and advancement on the subject (Ezugwu, 2020a). Therefore, there is the need for a comprehensive systematic survey of literature on both the traditional and recently proposed clustering techniques that have been applied in different fields. Hence, the following main research question for this study has been formulated as follow:

"*What are the various state-of-the-art clustering methods and algorithms discussed in the literature, and in what research domains have they been applied?*".

Towards realizing the answer for the main research question, the following sub-research questions are formulated:

- (a)
  What are the various *traditional and recently proposed clustering techniques and algorithms* in existence today?
- (b)
  What research has been conducted using both the traditional and recently proposed clustering techniques to address *identified challenges of clustering*?
- (c)
  In what domains have both the traditional and recently proposed clustering techniques been applied in solving clustering problems?
- (d)
  How have various similarity measures been employed in traditional and *recently proposed clustering techniques*?
- (e)
  What are the characteristic differences between the traditional and recently proposed clustering techniques that have been applied in different fields?

What *are other challenges of clustering problems* yet to be explored by researchers in this research area? This survey aims to provide an up-to-date comprehensive review of the different clustering techniques applied to many data mining-related fields. Retrospectively, we also highlight novel and most recent practical applications areas of clustering. This survey is intended to provide a convenient research path for new researchers, furnishing them with a comprehensive study on the various data clustering techniques and research progression over the years in clustering techniques. This survey will also help experts develop new algorithms for emerging challenges in the research area. The main contribution of this survey study is as follows:

- •

  Provides an up-to-date comprehensive systematic review of the traditional and recently proposed clustering techniques that have been applied in different fields.

- •

  Provides a concise presentation of concepts, architecture, and taxonomy of clustering algorithms.

- •

  Presents a discussion on open recent research issues relating to clustering problems

- •

  Defined possible future research trends and directions regarding the implementation and application of clustering algorithms in different research domains.

Section snippets

Methodology

This section presents the procedure used in selecting and reviewing the various clustering methods considered in this survey. In this comprehensive review process and methodology, the standard approach for systematic literature review was adopted and followed to ensure that the topic of interest is sufficiently covered and reduce bias on the review work. In this study, the literature review procedure proposed by (Weidt and Silva, 2016) was used in this paper. Moreover, the work (Thilakaratne et

Comparison with existing survey works

This section presents and discusses the main difference between the already published review papers and this survey paper. Although there have been several attempts in the literature to comprehensively present a systematic review of different clustering techniques with their trending applications areas, this effort has become almost impossible

because of the evolving nature of the study area and its relevance to different theoretical and practical fields of study relative to data mining and

Taxonomy of clustering algorithms

Several clustering algorithms have been identified and broadly classified under two categories, namely: the Hierarchical Clustering Algorithm and the Partitional Clustering Algorithm (Xu and Wunsch, 2005, Xu and Tian, 2015, Benabdellah et al., 2019, Adil et al., 2014, Dafir et al., 2021; Saxena et al., 2017, Nagpal, 2013; Oyelade et al., 2016, Bindra and Mishra, 2017, Singh and Srivastava, 2020, Djouzi and Beghdad-Bey, 2019, Ezugwu, 2020a). A hierarchical clustering algorithm is further

Recent work on clustering methods

Cluster validity evaluation is a major problem in the clustering algorithm. Li et al. (2020) addressed this problem by designing a cluster validity evaluation technique based on the Ratio of Deviation of Sum-of-squares and Euclid distance. The technique was evaluated on both artificial and real-world datasets, and the results show that it can dynamically obtain the near-optimal number of clusters. Chowdhury et al. (2020) introduced another technique for calculating the optimal number of

Discussion and open challenges

Many clustering-based algorithms have been proposed in the literature, and some of them performed remarkably well. This section presents a discussion on various issues in clustering analysis. The discussion is divided into three sub-sections. The first section presents a discussion on the performance of existing clustering algorithms. The second section presents some open issues in clustering algorithms. The third subsection presents some validation and similarity measures used in both

Trending application areas of clustering algorithms

Clustering algorithms can be applied to different domains. This section provides diverse areas that cluster analysis that has been successfully utilized. Specifically, we present the applicability of the clustering algorithms reviewed in Section 4 to the field of medicine, financial sector, artificial intelligence, aviation sector, marketing and sales sector, industries and manufacturing context, urban development, privacy protection, and robotics. Fig. 10 summarizes these fields by presenting

Concluding remark

Clustering is a powerful data mining and analysis tool used in many fields, including machine learning, bioinformatics, robotics, pattern recognition, and image analysis. Identifying the number of clusters apriori is the

most fundamental problem in cluster analysis. Specifying the correct number of clusters apriori can help obtain optimal solutions to many clustering problems. Because of this, automatic clustering algorithms are taking over traditional clustering algorithms. Automated

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.