International Advance Journal of Engineering, Science and Management (IAJESM)
ISSN -2393-8048, January-June 2019, Submitted in January 2019, iajesm2014@gmail.com

Study On the Clustering in Novel Techniques with Related to Its Classification

Basavaraj U, Assistant Professor, Department of Computer Science, Government First Grade College and PG Centre
Thenkinidiyur Udupi, Karnataka India, Email- ubkottur@gmail.com

Abstract

Feature extraction is essential in bioinformatics because it transforms genome sequences into the feature vectors required for data mining activities such as classification and clustering. The data mining activities enable us to classify or cluster the newly sequenced genome to the known families. Nowadays, a variety of feature extraction strategies are available for genome data. Nevertheless, several existing algorithms do not extract context-sensitive key properties, also some approaches extract features, which are unable to distinguish between two non-similar sequences. In addition, the efficacy of existing feature extraction techniques is evaluated on either supervised or unsupervised learning models, but not on both. Thus, an efficient feature extraction technique that extracts significantly relevant features from genome sequences is required. In this paper, a novel feature extraction method is proposed that extracts features based on the length of the sequence, the frequency of nucleotide bases, the modified positional sum of nucleotide bases, the distribution of nucleotide bases, and the entropy of the sequence to generate a 14-dimensional fixed-length numeric vector to describe each genome sequence uniquely. By applying extracted features to both supervised and unsupervised machine learning approaches, the performance of the proposed feature extraction method is assessed. The experimental results show that the proposed strategy for clustering and classifying novel genome sequences into recognized genome classes is highly effective and efficient. The same is proven by comparing the proposed method to the standard state-of-the-art method. Data Distribution can be obtained by clustering data. In this work we observed the characteristics of selected cluster, and make a further study on particular clusters. Also, cluster analysis generally acts as the preprocessing of other data mining operations. Consequently, cluster analysis has become a very active research topic in data mining. Data mining is a new technology, developing with database as well as artificial intelligence. It is a processing procedure of extracting credible and effective novel techniques and understandable patterns from the database. Cluster analysis can be important data mining method used to figure out the data segmentation and pattern information. The development of data mining methods, different types of clustering techniques establish. The study of clustering method from the perception of statistics, based on the statistical theory, The review of this paper make an effort to combine statistical method with the machine learning algorithm technique as well as introduce the existing best r-statistical softwares, including factor, correspondence and analysis of functional data into data mining. The present study is undertaken to develop a Data Mining workflow using clustering and classification of data, solving clustering problem as well as extracting association rules. Use the suitable proximity measure in addition to that to select the optimal clustering model to solve clustering problems. Develop a Data Mining workflow to extract association rules.

Keywords: ISODATA, SRIDHCR, r-solution, K-means

Introduction: Cluster analysis divides data into meaningful or useful groups (clusters). If meaningful clusters are our objective, then the resulting clusters should capture the “natural” structure of the data. Cluster analysis is only a useful starting point for other purposes, e.g., data compression or efficiently finding the nearest neighbors of points. Whether for understanding or utility, cluster analysis has long been used in a wide variety of fields: psychology and other social sciences, biology, statistics, pattern recognition, information retrieval, machine learning, and data mining. In this chapter we provide a short introduction to cluster analysis. We present a brief view recent technique, which uses a conceptbased approach. In this case, the approach to clustering high dimensional data must deal with the “curse of dimensionality”. Clustering technique is a method used to group similar pixels of an object together while disregarding dissimilar pixels. It involves creating multiple clusters, which can then be used to identify different types of objects. In the field of computer science, a popular algorithm used for clustering is the k-means clustering algorithm. It classifies pixels on a leaf into clusters based on their colors, such as background, foreground, and diseased parts. The algorithm relies on initially selected cluster centroids, and it iteratively computes the cluster centers until no more changes occur. The resulting cluster pixels can then be reshaped into images for further analysis.

In the data-mining world, clustering and classification are two types of methods. Both these methods can be used to characterize the objects into groups which are having one or more features. The difference between classification and clustering is an unsupervised learning technique that can be used for similar group based on features, while classification is a supervised learning technique. The business. Abundant data carries abundant problems. Data mining involves various techniques, also data mining plays a very important in today and tomorrow's scenarios. This becomes possible with the usage of clustering and classification of data to make the data more robust, meaningful and usable with least wastage. The major problems we see with huge date are; a) Data mining algorithms need to be more efficient and scalable to extract the information from the huge amount of data in databases. b) Dealing with datasets that require distributed approaches. c) Mining information from heterogeneous databases and global information systems. d) Processing of large, complex and unstructured data into a structured format. Data clustering is an exploratory as well as descriptive data analysis technique that has gained a lot of attention, e.g., in statistics, data mining, pattern recognition, etc. It is an explorative way to investigate multivariate data sets that contain possibly various type of data. This kind of data sets differs from each other's in size concerning some objects as well as dimensions, or they contain different data types. certainly, the data clustering related to the core technique of data mining, in that one focuses on large data sets with unknown underlying structure. The intention of this report is to be an introduction to specific parts of this methodology called cluster analysis. Partitioning based clustering techniques may be flexible methods for iterative relocation of data points between clusters. The quality of the solutions can be measured by a clustering criterion. Each iteration the iterative relocation algorithms reduce the value of the criterion function, until convergence mean while changing the clustering criterion, it is possible to construct robust clustering method which can be more insensitive to incorrect and missing.

LITERATURE SURVEY

Marwah A. Helaly et al. [4] proposed a deep learning strategy for the taxonomy classification of bacterial sequence. To describe the genome data, they employed a variety of representations, including one-hot encoding, inter-encoding, and k-mers-based representation. They evaluated their strategy on the 16S rRNA dataset utilizing a deeper convolution neural network (CNN) and obtained an accuracy of 91.7% with a more representative representation and 90.0% with a less figurative representation. The problem with this method is that it can only be used with labeled data. After that, Jasbir Dhaliwal and John Wanger [5] made a new way to extract features for SNPs that are highly expressed. As features, they employed k-mer to describe the SNP sequence. According to them, ideal k-mer and feature size may vary between research problems. They assessed their technique using a multinomial naive bayes on 49 human tissues and obtained optimal k-mer of size 3. One of the biggest problems with using k-mer is that storing an SNP sequence with large-sized k-mers takes a lot of memory. Later on, Preeti Jha et al. [3] presented a feature extraction method named 12d-FV for the SNP sequencing analysis of unlabeled real-world plant genome datasets. To describe an SNP sequence, they employed three sorts of features: frequency, total distance, and nucleotide distribution. They used kernelized scalable random sampling with iterative fuzzy c-means (KSRSIO-FCM) to test their method and evaluated the results using the silhouette index [?]. The disadvantage of this strategy is that the total distance and the distribution for each nucleotide may be the same for dissimilar sequences. This technique may be incapable of differentiating between the sequences as a result of this incapability. In addition, the sequence length, which differs between organisms, has not been used in this method. In 2018, Bonidia et al. [6] proposed a new package named MathFeature, for extracting the numerical features from the ribonucleic acid (RNA), DNA and protein sequences. In this package they used 20 numerical feature extraction descriptors based on the numerous numeric mappings, chaos game theory, gemonic signal processing, complex networks and entropy for converting the biological sequences into numerical values. They evaluated their method on eight benchmark datasets and found that MathFeature outperformed competing methods. According to the aforementioned literature, the majority of feature extraction algorithms are either available for labeled data or unlabeled data, but not for both. In addition, several algorithms do not extract context-based properties. In contrast, some algorithms fail to extract essential features such as length and entropy. To overcome the limitations identified in this study, a novel 14-dimensional feature extraction technique is proposed, which extracts features based on length, frequency, modified position sum, distribution, and entropy to characterize the genome sequence uniquely. In the further section we will brief about the concept-byconcept analysis of the proposed method, i.e., the five distinct categories of features and their extraction procedure, illus trated with an example and the implementation of proposed approach.

What is Clustering ?

The task of grouping data points based on their similarity with each other is called Clustering or Cluster Analysis. This method is defined under the branch of Unsupervised Learning, which aims at gaining insights from unlabelled data points, that is, unlike supervised learning we don’t have a target variable.

Clustering aims at forming groups of homogeneous data points from a heterogeneous dataset. It evaluates the similarity based on a metric like Euclidean distance, Cosine similarity, Manhattan distance, etc. and then group the points with highest similarity score together.

For Example, In the graph given below, we can clearly see that there are 3 circular clusters forming on the basis of distance.

Uses of Clustering

Now before we begin with types of clustering algorithms, we will go through the use cases of Clustering algorithms. Clustering algorithms are majorly used for:

- Market Segmentation – Businesses use clustering to group their customers and use targeted advertisements to attract more audience.
- Market Basket Analysis – Shop owners analyze their sales and figure out which items are majorly bought together by the customers. For example, In USA, according to a study diapers and beers were usually bought together by fathers.
- Social Network Analysis – Social media sites use your data to understand your browsing behaviour and provide you with targeted friend recommendations or content recommendations.
- Medical Imaging – Doctors use Clustering to find out diseased areas in diagnostic images like X-rays.
- Anomaly Detection – To find outliers in a stream of real-time dataset or forecasting fraudulent transactions we can use clustering to identify them.
- Simplify working with large datasets – Each cluster is given a cluster ID after clustering is complete. Now, you may reduce a feature set’s whole feature set into its cluster ID. Clustering is effective when it can represent a complicated case with a straightforward cluster ID. Using the same principle, clustering data can make complex datasets simpler.

There are many more use cases for clustering but there are some of the major and common use cases of clustering. Moving forward we will be discussing Clustering Algorithms that will help you perform the above tasks.

Types of Clustering Algorithms

At the surface level, clustering helps in the analysis of unstructured data. Graphing, the shortest distance, and the density of the data points are a few of the elements that influence cluster formation. Clustering is the process of determining how related the objects are based on a metric called the similarity measure. Similarity metrics are easier to locate in smaller sets of features. It gets harder to create similarity measures as the number of features increases. Depending on the type of clustering algorithm being utilized in data mining, several techniques are employed to group the data from the datasets. In this part, the clustering techniques are described. Various types of clustering algorithms are:

1. Centroid-based Clustering (Partitioning methods)
2. Density-based Clustering (Model-based methods)
3. Connectivity-based Clustering (Hierarchical clustering)
4. Distribution-based Clustering

We will be going through each of these types in brief.

1. Centroid-based Clustering (Partitioning methods)

Partitioning methods are the most easiest clustering algorithms. They group data points on the basis of their closeness. Generally, the similarity measure chosen for these algorithms are Euclidian distance, Manhattan Distance or Minkowski Distance. The datasets are separated into a predetermined number of clusters, and each cluster is referenced by a vector of values. When compared to the vector value, the input data variable shows no difference and joins the cluster.

The primary drawback for these algorithms is the requirement that we establish the number of clusters, “k,” either intuitively or scientifically (using the Elbow Method) before any clustering machine learning system starts allocating the data points. Despite this, it is still the most popular type of clustering. K-means and K-medoids clustering are some examples of this type clustering.

2. Density-based Clustering (Model-based methods)

Density-based clustering, a model-based method, finds groups based on the density of data points. Contrary to centroid-based clustering, which requires that the number of clusters be predefined and is sensitive to initialization, density-based clustering determines the number of clusters automatically and is less susceptible to beginning positions. They are great at handling clusters of different sizes and forms, making them ideally suited for datasets with irregularly shaped or overlapping clusters. These methods manage both dense and sparse data regions by focusing on local density and can distinguish clusters with a variety of morphologies.

In contrast, centroid-based grouping, like k-means, has trouble finding arbitrary shaped clusters. Due to its preset number of cluster requirements and extreme sensitivity to the initial positioning of centroids, the outcomes can vary. Furthermore, the tendency of centroid-based approaches to produce spherical or convex clusters restricts their capacity to handle complicated or irregularly shaped clusters. In conclusion, density-based clustering overcomes the drawbacks of centroid-based techniques by autonomously choosing cluster sizes, being resilient to initialization, and successfully capturing clusters of various sizes and forms. The most popular density-based clustering algorithm is DBSCAN.

3. Connectivity-based Clustering (Hierarchical clustering)

A method for assembling related data points into hierarchical clusters is called hierarchical clustering. Each data point is initially taken into account as a separate cluster, which is subsequently combined with the clusters that are the most similar to form one large cluster that contains all of the data points.

Think about how you may arrange a collection of items based on how similar they are. Each object begins as its own cluster at the base of the tree when using hierarchical clustering, which creates a dendrogram, a tree-like structure. The closest pairings of clusters are then combined into larger clusters after the algorithm examines how similar the objects are to one another. When every object is in one cluster at the top of the tree, the merging process has finished. Exploring various granularity levels is one of the fun things about hierarchical clustering. To obtain a given number of clusters, you can select to cut the dendrogram at a particular height. The more similar two objects are within a cluster, the closer they are. It’s comparable to classifying items according to their family trees, where the nearest relatives are clustered together and the wider branches signify more general connections. There are 2 approaches for Hierarchical clustering:

- **Divisive Clustering:** It follows a top-down approach, here we consider all data points to be part one big cluster and then this cluster is divide into smaller groups.

- **Agglomerative Clustering:** It follows a bottom-up approach, here we consider all data points to be part of individual clusters and then these clusters are clubbed together to make one big cluster with all data points.

4. Distribution-based Clustering

Using distribution-based clustering, data points are generated and organized according to their propensity to fall into the same probability distribution (such as a Gaussian, binomial, or other) within the data. The data elements are grouped using a probability-based distribution that is based on statistical distributions. Included are data objects that have a higher likelihood of being in the cluster. A data point is less likely to be included in a cluster the further it is from the cluster's central point, which exists in every cluster.

A notable drawback of density and boundary-based approaches is the need to specify the clusters a priori for some algorithms, and primarily the definition of the cluster form for the bulk of algorithms. There must be at least one tuning or hyper-parameter selected, and while doing so should be simple, getting it wrong could have unanticipated repercussions. Distribution-based clustering has a definite advantage over proximity and centroid-based clustering approaches in terms of flexibility, accuracy, and cluster structure. The key issue is that, in order to avoid overfitting, many clustering methods only work with simulated or manufactured data, or when the bulk of the data points certainly belong to a preset distribution. The most popular distribution-based clustering algorithm is Gaussian Mixture Model.

Applications of Clustering in different fields:

1. **Marketing:** It can be used to characterize & discover customer segments for marketing purposes.
2. **Biology:** It can be used for classification among different species of plants and animals.
3. **Libraries:** It is used in clustering different books on the basis of topics and information.
4. **Insurance:** It is used to acknowledge the customers, their policies and identifying the frauds.
5. **City Planning:** It is used to make groups of houses and to study their values based on their geographical locations and other factors present.
6. **Earthquake studies:** By learning the earthquake-affected areas we can determine the dangerous zones.
7. **Image Processing**: Clustering can be used to group similar images together, classify images based on content, and identify patterns in image data.
8. **Genetics:** Clustering is used to group genes that have similar expression patterns and identify gene networks that work together in biological processes.
9. **Finance:** Clustering is used to identify market segments based on customer behavior, identify patterns in stock market data, and analyze risk in investment portfolios.
10. **Customer Service:** Clustering is used to group customer inquiries and complaints into categories, identify common issues, and develop targeted solutions.
11. **Manufacturing**: Clustering is used to group similar products together, optimize production processes, and identify defects in manufacturing processes.
12. **Medical diagnosis:** Clustering is used to group patients with similar symptoms or diseases, which helps in making accurate diagnoses and identifying effective treatments.
13. **Fraud detection:** Clustering is used to identify suspicious patterns or anomalies in financial transactions, which can help in detecting fraud or other financial crimes.
14. **Traffic analysis:** Clustering is used to group similar patterns of traffic data, such as peak hours, routes, and speeds, which can help in improving transportation planning and infrastructure.
15. **Social network analysis:** Clustering is used to identify communities or groups within social networks, which can help in understanding social behavior, influence, and trends.
16. **Cybersecurity:** Clustering is used to group similar patterns of network traffic or system behavior, which can help in detecting and preventing cyberattacks.

17. **Climate analysis:** Clustering is used to group similar patterns of climate data, such as temperature, precipitation, and wind, which can help in understanding climate change and its impact on the environment.
18. **Sports analysis:** Clustering is used to group similar patterns of player or team performance data, which can help in analyzing player or team strengths and weaknesses and making strategic decisions.
19. **Crime analysis:** Clustering is used to group similar patterns of crime data, such as location, time, and type, which can help in identifying crime hotspots, predicting future crime trends, and improving crime prevention strategies.

Conclusion

In this article we discussed Clustering, it's types, and it's applications in the real world. There is much more to be covered in unsupervised learning and Cluster Analysis is just the first step. This article can help you get started with Clustering algorithms and help you get a new project that can be added to your portfolio. In this paper, we proposed a novel 14-dimensional feature extraction method, abbreviated as "14d-FET" for genome data made up of the four nucleotide bases A, T, G, and C. Utilizing a novel power method, the proposed 14d-FET tackled the problem of positional sum and distribution features, which can sometimes be the same for dissimilar sequences. In addition, we extracted the most essential features, such as sequence length and entropy, to improve the effectiveness of the proposed feature extraction approach. Moreover, experimental results demonstrate that the proposed method generates the generalized strategy for feature extraction regardless of the evaluation method employed. When the labeled dataset was evaluated on six different classifiers, the proposed method performed better than the existing method. It increased the precision of validation across all labeled datasets. In the case of unlabeled datasets, the proposed method likewise performed well and yielded an enhanced SI compared to existing strategy. Therefore, we can conclude that the proposed method performs well in both supervised and unsupervised learning. In the future, additional feature extraction criteria can be added in proposed feature extraction method to enhance the efficacy of feature extraction method.

References:

1. Formaggio, A. R., Vieira, M. A., & Rennó, C. D. (2012, July). Object Based Image Analysis (OBIA) and Data Mining (DM) in Landsat time series for mapping soybean in intensive agricultural regions. In 2012 IEEE International Geoscience and Remote Sensing Symposium (pp. 2257-2260). IEEE.

2. Saggar, M., Agrawal, A. K., & Lad, A. (2004, October). Optimization of association rule mining using improved genetic algorithms. In 2004 IEEE International Conference on Systems, Man and Cybernetics (IEEE Cat. No. 04CH37583) (Vol. 4, pp. 3725-3729). IEEE.

3. Hu, Y., Guo, Z., Wen, J., & Han, J. (2015, June). Research on knowledge mining for agricultural machinery maintenance based on association rules. In 2015 IEEE 10th Conference on Industrial Electronics and Applications (ICIEA) (pp. 885-890). IEEE.

4. Bharara, Sanyam, A. SaiSabitha, and AbhayBansal. "A review on knowledge extraction for business operations using data mining." In 2017 7th International Conference on Cloud Computing, Data Science & EngineeringConfluence, pp. 512-518. IEEE, 2017.

5. Narander Kumar, SabitaKLhatri, Department of Computer, 3rd IEEE International Conference on Computational Intelligence and Communication Technology (IEEE-CICT 2017) " Implementing WEKA for medical data classification and early disease prediction 978-1-50, 2017

6. Thiyagaraj, M., & Suseendran, G. Research of Chronic Kidney Disease based on Data Mining Techniques.

7. Kirubha, V., & Priya, S. M. (2016). Survey on data mining algorithms in disease prediction. Int J Comput Trends Tech, 38(3), 24-128