Investigating Clustering Techniques through WEKA

Maham Hameed
Department of Computer Science
Riphah International University
Lahore, Pakistan
mahamhameed120@gmail.com

Aqsa Abbas
Department of Computer Science
Riphah International University
Lahore, Pakistan
aqsaabbas081@gmail.com

Amara Javed
Department of Computer Science
Riphah International University
Lahore, Pakistan
amarajaved706@gmail.com

***Abstract*—Data Mining is the process of extracting useful information and discovering patterns from large datasets to generate some useful information and helps in getting knowledge-based information. One of the main purposes of data mining is to identify which algorithm works best for the given problem or situation. In this paper, we will use Clustering techniques and explore its different algorithms on a Pima Indian diabetes dataset. The algorithms which we will be using are Simple K-means, Farthest First, EM and Filtered cluster. Implementation will be done on a data mining tool called WEKA to determine which of these algorithms give better optimized results.**

***Keywords*—Data Mining, Clustering, WEKA, algorithm, Simple K-means, Farthest First, EM, Filtered cluster.**

I. Introduction

Nowadays, as the technology is increasing rapidly and lots of data is present, much of the data is useful. Hence, data mining plays an important role here. Data mining extracts the useful data from huge raw data. Data mining procedure follows six steps which are Data collection, data storage, data cleaning, data analysis, visualization and decision. It provides us with many benefits such as: it helps to predict future trends, tells us about the customers habit, helps in decision making, fraud detection is quickly identified, increasing the revenue of the companies, increases customer satisfaction, targets the advertising campaigns and much more. Therefore, data mining is beneficial in every aspect. Classification, clustering, regression, etc. are the techniques in data mining which are widely used. Among these, we will be using clustering and its different algorithms which are discussed in this paper.

Clustering basically deals with uncategorized data. It is unsupervised learning which means the data which is UN labelled and you do not need to supervise the model. One does not have any knowledge of the given data.

The model works on its own and discovers information. It is unpredictable. It is the process of Finding patterns from uncategorized data and finding the similarities between data, based on its features. Similar items are grouped together. There is no built-in class. We assemble the data into set of classes. Some of the different clustering methods are: Partitioning methods, Hierarchical clustering, Fuzzy clustering, Density based clustering, Model based clustering, K means clustering, EM, Farthest First, Simple K means, Filtered Cluster, etc. Deciding which clustering method is best for a dataset is really important. High quality clusters can be achieved by using good Clustering techniques. These techniques acknowledge different patterns of data. It gives us to the point summaries of data.

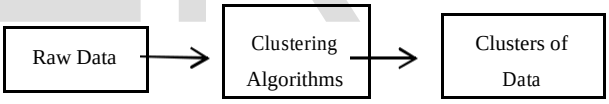


Figure 1. Clustering patterns

Clustering in used in different fields such as image analysis, bioinformatics, pattern recognition, Marketing- characterizing and discovering customer segments, Libraries-on the basis of topics and information in different books, Biology, Planning, Identifying Fake News or criminal activity, etc.

II. Literature Review

In [19], V.Mahalakshmi, M. G. proposed the Performance Analysis of Clustering Algorithms for Diabetes Data. They analyzed three algorithms on the diabetes dataset which are k-means, EM and DBS can and concluded that k-mean has taken smallest time to build. In [11], Pradeep Ra, S. S. proposed the Survey of clustering Techniques which describes the review of all the data mining clustering techniques.

They have discussed about the classification of clusters which are grid-based algorithms, partitioning methods, density-based algorithms and hierarchal methods.

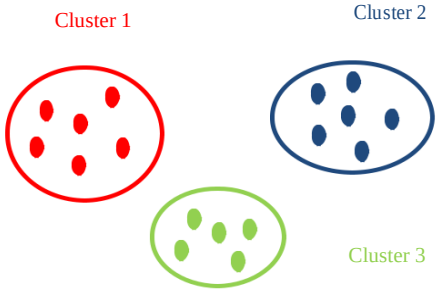
In [14], Rupali Patil, S. D., K Rajeswari proposed the Analysis of Simple K-Means with Multiple Dimensions using WEKA. They performed the experiment on irrigation census dataset and concluded that Simple Kmeans clustering works best on the numerical data. They also concluded that simple K means works more effectively on smaller dataset which have numerical attributes. In [2], Anju Parmar, D. C. a. D. K. L. B. proposed the Performance evaluation of weka clustering algorithms on large datasets. In this paper, different datasets have been taken to evaluate clustering algorithms. Nine different clustering algorithms have been analyzed with respect to the time taken to build the model and the number of the clusters formed. They performed the experiment on weka and stated that weka works well with small datasets and cannot handle large datasets as it supports sequential single node. They said that hadoop based technologies are best for large datasets. In [18],T. Sajana, C. M. S. R. a. K. V. N proposed a Survey on Clustering Techniques for Big Data Mining. Their aim was to detect the outliers in large datasets and they suggested that Birch, Clique and Orclus algorithms should be used on large datasets.

By analyzing different clustering techniques, they concluded which algorithms should be used on categorical data, numerical data and spatial data. Similar work is been done on analyzing and comparing different clustering algorithms or large and small datasets. As our technology is increasing, new ideas and new approaches for doing things also get modified and updated. Hence, researches are doing much work on demonstrating the data mining algorithms.

III. Methodology

A. Clustering

Clustering is to isolate or sort out group with similar characteristics and assign them into clusters.



The aim of this paper is to analyze four different clustering methods which are

- Farthest First
- Simple K means
- EM
- Filtered cluster

We will see all these algorithms in detail and their implementation in this paper.

B. Dataset

The dataset used for analyzing of the clustering algorithms is Pima Indian diabetes and is taken from kaggle (an online community of data scientists. It has huge bank of datasets easily available). This dataset contains 769 rows and 9 columns which determines risk factors that lead to diabetes. Diabetes mellitus is a disease in which blood sugar which is also called as glucose levels is unusually high as the body does not produce enough insulin. In this type of diabetes people lose lot of weight. Patients become progressively thirsty and hungry.

Polyuria is experienced because of high blood sugar. Diabetes destructs the nerves and causes problems with sensation as the patients do not get required amount of glucose in their bodies which are extremely important for human health.

C. Weka

Weka is the Waikato Environment for knowledge analysis, developed at University of Waikato, New Zealand. A tool for data mining that performs task such as data pre-processing, Classification, regression, association and visualization. You upload the dataset and perform the desired task by using any of these algorithms. Weka is an excellent platform for Machine Learning and Data Mining. In weka, you can run experiments and analyze results. It is widely used by researchers, industries and in academic purposes.

D. Farthest First

Farthest First is proposed by Hochbarm and Shmoys in 1985. It is the variant of K means. It deals with K-center problem. Each cluster center is placed at the point furthest from the existing centers and must be within the area of data. It chooses centroids and clusters are assigned in objects with maximum distance.

This type of clustering algorithm is suitable for large datasets but it creates non-uniform clusters. It is also called as the greedy algorithm. A brief overview of this algorithm is that it selects the center randomly. The second center is selected greedily i.e. the point which is farthest or maximum from the second point. In this way, the remaining point are also selected (distance farthest from the first). For each of the remaining points, distance is calculated to each cluster center. Put in the cluster with the most less or minimum distance. This algorithm produces threshold values.

E. EM

EM stands for Expectation Maximization. E is the expectation of log likelihood whereas M maximizes the log likelihood found from E. To some extent, this algorithm is similar to K-means. It performs the maximum likelihood in the existence of inactive variables. It does this by estimating the values for the inactive variables, the model is been optimized and until convergence these two steps will be repeated.
For density estimation in clustering algorithms, this is the most general approach. It decides creation of clustering by cross validation.

F. Filtered Cluster

In mathematics, filter is an important and unique subset of a relatively lined set. There is a need for filtering, to probe into the importance of filtering, preliminary to cluster analysis. This clustering algorithm establishes on saving the multidimensional data point in kd-tree which is somehow similar to a binary tree, representing a hierarchical segmentation of the data point set's bounding box utilizing their axis followed by splitting which is line up by the hyper planes. Every node of the kd-tree is connected with cell which is the bounding box of the point in dataset. It is proclaimed to be a leaf if the cell has at most one point. The discovery points present in the cell are divided to one side or the other side of the hyper plane. The sub cells which resulted are the children of the original cell which then leads to binary tree structure. Filtering enhances the knowledge of the principles component plot considerably.

G. Simple K means

It is one of the important algorithms in Machine learning. It is an unsupervised machine learning algorithm and uses Euclidean distance measure. K-means computes distance between the clusters and instances.

IV. Results

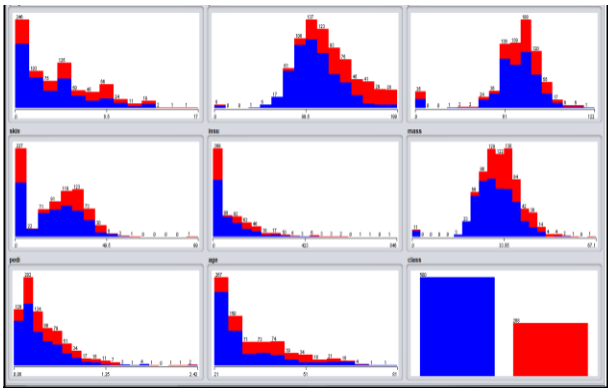


Figure 2. Visualization

Figure 2, shows the visualization of the histogram of each attribute present in diabetes dataset.

Name	No. of clusters	Cluster instances	Time taken (sec)
Farthest First	2	615 (80%) 153 (20%)	0.11s
EM	3	228 (30%) 203 (26%) 337 (44%)	8.53
Filtered Cluster	2	500 (65%) 268 (35%)	0.06
Simple K means	2	500 (65%) 268 (35%)	0.01

Figure 3. Experimental results

By performing all these algorithms one by one on the Pima Indian diabetes dataset, we have seen the number of clusters formed and their instances. Simple K means takes less time to build which is 0.01 seconds as compared to other algorithms. More accurate clusters are also formed in k means. K means and filtered cluster, both have made the same cluster instances, only there is a difference in time taken to build the model. Expectation Minimization EM has taken the most time to build the model which is 8.53 seconds.

V. Conclusion

The significance of data mining has been discussed in this paper i.e. it extracts useful information from huge sets of data. The aim of the paper was to analyze different techniques of clustering. Clustering is grouping of similar items so that in a group they are similar to each other. We have analyzed four different algorithms of clustering which are Farthest First, EM-Expectation Maximization, Simple K means and Filtered clustering on a pima Indian diabetes dataset to check which algorithms works more efficiently. However, the experiment was conducted and among all the four algorithms simple k means was the best and gave more accurate results. Experiment was carried out on data mining tool, WEKA. In future, we decide to work on large datasets and explore different algorithms of clustering.

Refrences

[1]. Sathyendranath Malli, D. N. H. R., Dr. H G Joshi (September 2014). "A Study on Rural Health care Data sets using Clustering Algorithms." from https://www.researchgate.net/publication/290111927_A_Study_on_Rural_Health_care_Data_sets_using_Clustering_Algorithms.

[2]. Anju Parmar, D. C. a. D. K. L. B. (June 2017). "PERFORMANCE EVALUATION OF WEKA CLUSTERING ALGORITHMS ON LARGE DATASETS." from http://www.journalijar.com/article/18041/performance-evaluation-of-weka-clustering-algorithms-on-large-datasets/.

[3]. Berkhin, P. (2006). "A Survey of Clustering Data Mining Techniques." from https://link.springer.com/chapter/10.1007/3-540-28349-8_2.

[4]. Dhara Patel, R. M., Ketan Sarvakar (october 2014). "A Comparative Study of Clustering Data Mining: Techniques and Research Challenges." from https://www.researchgate.net/publication/267763228_A_Comparative_Study_of_Clustering_Data_Mining_Techniques_and_Researc h_Challenges.

[5]. Dongre, D. P. a. Y. (June 2015). "A CLUSTERING TECHNIQUE FOR EMAIL CONTENT MINING." from https://pdfs.semanticscholar.org/be6e/85e325250af1aebb8943b2b6bde8a772d99d.pdf.

[6]. Godwin Ogbuabor; Ugwoke, F. N. (30 April 2108). "Clustering Algorithm for a Healthcare Dataset Using Silhouette Score Value." from https://zenodo.org/record/1248795#.XrXTB mgzbIU.

[7]. harmila, M. K. (2013). "An Optimized Farthest First Clustering Algorithm." from https://www.researchgate.net/publication/262979934_An_optimized_farthest_first_clustering_algorithm.

[8]. Kashish Ara Shakil, S. A., Mansaf Alam (18 February 2015). "Dengue disease prediction using weka data mining tool." from https://arxiv.org/abs/1502.05167.

[9]. Litoriya, R. (may 2012). "Comparison of the various clustering algorithms of weka tools." from https://www.researchgate.net/publication/293173843_Comparison_of_the_various_clustering_algorithms_of_weka_tools.

[10]. N. G. J. Dias, R. P. T. H. G., M. C. Wijegunasekara (december 2014). "Comparison of Major Clustering Algorithms Using Weka Tool." from https://www.researchgate.net/publication/272623527_Comparison_of_Major_Clustering_Algorithms_Using_Weka_Tool.

[11]. Pradeep Ra, S. S. (October 2010). "A Survey of Clustering Techniques." from https://pdfs.semanticscholar.org/e9c0/1ff0a823114473ad773cf18e6f0e91a1adf2.pdf.

[12]. Priya Kakkar, A. P. (2014). "Comparison of Different Clustering Algorithmsusing WEKA Tool." from https://www.academia.edu/8298134/Comparison_of_Different_Clustering_Algorithms_using_WEKA_Tool.

[13]. Quan Zou, K. Q., Yamei Luo, Dehui Yin, Ying Ju and Hua Tang (6 November 2018). "Predicting Diabetes Mellitus With Machine Learning Techniques." from https://www.frontiersin.org/articles/10.3389/fgene.2018.00515/full.

[14]. Rupali Patil, S. D., K Rajeswari (January 2015). "Analysis of Simple K-Means with Multiple Dimensions using WEKA." from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.695.7313&rep=rep1&type=pdf.

[15]. Sondakh, D. E. (January 2016). "Performance Analysis of Machine Learning Algorithms for Multi-class Document Using WEKA." from https://jurnal.unai.edu/index.php/jiscse/article/view/310.

[16]. Srivastava, S. (January 2014)."Weka: A Tool for Data preprocessing, Classification, Ensemble, Clustering and Association Rule Mining." from https://www.researchgate.net/publication/262984959_Weka_A_Tool_for_Data_preprocessing_Classification_Ensemble_Clustering_and_Association_Rule_Mining.

[17]. Swasti Singhal, M. J. (May 2013). "A Study on WEKA Tool for Data Preprocessing, Classification and Clustering." from http://citeseerx.ist.psu.edu/viewdoc/downloa d?doi=10.1.1.687.799&rep=rep1&type=pdf.

[18]. T. Sajana, C. M. S. R. a. K. V. N. (february 2016). "A Survey on Clustering Techniques for Big Data Mining." from https://www.researchgate.net/publication/29 8082409_A_Survey_on_Clustering_Techniq ues_for_Big_Data_Mining.

[19]. T. Sajana, C. M. S. R. a. K. V. N. (february 2016). "A Survey on Clustering Techniques for Big Data Mining." from https://www.researchgate.net/publication/29 8082409_A_Survey_on_Clustering_Techniq ues_for_Big_Data_Mining.

[20]. Umatejaswi, P. S. K. a. V. (june 2017). "Diagnosing Diabetes using Data Mining Techniques." from http://www.ijsrp.org/research-paper-0617/ijsrp-p6689.pdf.

[21]. V.Mahalakshmi, M. G. (16 november 2015). "Performance Analysis of Clustering Algorithms for Diabetes Data." from https://www.researchgate.net/publication/28 3124447_Performance_analysis_of_clusteri ng_algorithms_for_diabetes_data.

[22]. Mining Clustering Techniques in Academia.". from https://www.researchgate.net/publication/22 0708919_Data_Mining_Clustering_Techniq ues_in_Academia.

[23]. Sondakh, D. E. (January 2016). "Performance Analysis of Machine Learning Algorithms for Multi-class Document Using WEKA." from https://jurnal.unai.edu/index.php/jiscse/articl e/view/310.

[24]. V.Mahalakshmi, M. G. (16 november 2015). "Performance Analysis of Clustering Algorithms for Diabetes Data." from https://www.researchgate.net/publication/28 3124447_Performance_analysis_of_clusteri ng_algorithms_for_diabetes_data.