A Comparative Study of clustering algorithms
Using weka tools

Bharat Chaudhari¹, Manan Parikh²

¹²MECSE, KITRC KALOL

ABSTRACT

Data clustering is a process of putting similar data into groups. A clustering algorithm partitions a data set into several groups based on the principle of maximizing the intra-class similarity and minimizing the inter-class similarity. This paper analyzes the three major clustering algorithms: K-Means, Hierarchical clustering and Density based clustering algorithm and compare the performance of these three major clustering algorithms on the aspect of correctly class wise cluster building ability of algorithm. Performance of the 3 techniques are presented and compared using a clustering tool WEKA.

Keywords: Data mining algorithms, Weka tools, K-means algorithms, Hierarchical clustering, Density based clustering algorithm etc.

1. INTRODUCTION

Knowledge discovery process consists of an iterative sequence of steps such as data cleaning, data integration, data selection, data transformation, data mining, pattern evaluation and knowledge presentation. Data mining functionalities are characterization and discrimination, mining frequent patterns, association, correlation, classification and prediction, cluster analysis, outlier analysis and evolution analysis [1]. Three of the major data mining techniques are regression, classification and clustering. CLUSTERING is a data mining technique to group the similar data into a cluster and dissimilar data into different clusters. I am using Weka data mining tools for clustering.

2. WEKA

Weka is a landmark system in the history of the data mining and machine learning research communities, because it is the only toolkit that has gained such widespread adoption and survived for an extended period of time[2]. The Weka or woodhen (Gallirallus australis) is an endemic bird of New Zealand. It provides many different algorithms for data mining and machine learning. Weka is open source and freely available. It is also platform-independent.

Figure 1 View of weka tool
The GUI Chooser consists of four buttons:
- Explorer: An environment for exploring data with WEKA.
- Experimenter: An environment for performing experiments and conducting statistical tests between learning schemes.
- Knowledge Flow: This environment supports essentially the same functions as the Explorer but with a drag-and-drop interface. One advantage is that it supports incremental learning.
- Simple CLI: Provides a simple command-line interface that allows direct execution of WEKA commands for operating systems that do not provide their own command line interface.

3. CLUSTERING ALGORITHMS AND TECHNIQUES

Many algorithms exist for clustering. Following figures showing three major clustering methods and their approach for clustering.

3.1 K-means Clustering

The term “k-means" was first used by James MacQueen in 1967 [3], though the idea goes back to 1957 [4]. The standard algorithm was first proposed by Stuart Lloyd in 1957 as a technique for pulse-code modulation, though it wasn’t published until 1982. K-means is a widely used partitional clustering method in the industries. The K-means algorithm is the most commonly used partitional clustering algorithm because it can be easily implemented and is the most efficient one in terms of the execution time.

Here’s how the algorithm works [5]:

**K-Means Algorithm**: The algorithm for partitioning, where each cluster’s center is represented by mean value of objects in the cluster.

**Input**: k: the number of clusters. D: a data set containing n objects.

**Output**: A set of k clusters.

**Method**:
1. Arbitrarily choose k objects from D as the initial cluster centers.
2. Repeat.
3. (re)assign each object to the cluster to which the object is most similar using Eq. 1, based on the mean value of the objects in the cluster.
4. Update the cluster means, i.e. calculate the mean value of the objects for each cluster.
5. until no change.
This figure shows the result of k-means clustering methods using WEKA tool. After that we saved the result, the result will be saved in the ARFF file format. We also open this file in MS Excel.

### 3.2 Hierarchical clustering

HIERARCHICAL CLUSTERING builds a cluster hierarchy or, in other words, a tree of clusters, also known as a dendrogram[6].

**Agglomerative (bottom up)**
1. Start with 1 point (singleton).
2. Recursively add two or more appropriate clusters.
3. Stop when k number of clusters is achieved.

**Divisive (top down)**
1. Start with a big cluster.
2. Recursively divide into smaller clusters.
3. Stop when k number of clusters is achieved.

This figure shows that the result of Hierarchical clustering methods with single-linkage [7] between data points using WEKA tool.
3.3 Density based clustering

Density-based clustering algorithms try to find clusters based on density of data points in a region. The key idea of density-based clustering is that for each instance of a cluster the neighborhood of a given radius (Eps) has to contain at least a minimum number of instances (MinPts). One of the most well known density-based clustering algorithms is the DBSCAN [7].

DBSCAN separates data points into three classes:

- Core points: These are points that are at the interior of a cluster.
- Border points: A border point is a point that is not a core point, but it falls within the neighborhood of a core point.
- Noise points: A noise point is any point that is not a core point or a border point.

To find a cluster, DBSCAN starts with an arbitrary instance (p) in data set (D) and retrieves all instances of D with respect to Eps and MinPts. The algorithm makes use of a spatial data structure(R*tree) to locate points within Eps distance from the core points of the clusters [8]. Another density based algorithm OPTICS is introduced in [9], which is an interactive clustering algorithm, works by creating an ordering of the data set representing its density-based clustering structure.

![Figure 5 Result of Density based Clustering.](image)

Above figure showing the result of density-based clustering methods using WEKA tool.

4. COMPARISON

Above section involves the study of each of the three techniques introduced previously using Weka Clustering Tool on a set of banking data consists of 11 attributes and 600 entries [10]. Clustering of the data set is done with each of the clustering algorithm using Weka tool and the results are:

<table>
<thead>
<tr>
<th>Name</th>
<th>No. of clusters</th>
<th>Cluster Instances</th>
<th>No. of iterations</th>
<th>Within clusters sum of squared errors</th>
<th>Time taken to build model</th>
<th>Log likelihood</th>
<th>Unclassified Instances</th>
</tr>
</thead>
<tbody>
<tr>
<td>K-Means</td>
<td>2</td>
<td>0: 247(41%)</td>
<td>3</td>
<td>1737.279199193</td>
<td>0.23 seconds</td>
<td>-22.19241</td>
<td>0</td>
</tr>
<tr>
<td>Hierarchical</td>
<td>2</td>
<td>0: 599 (100%)</td>
<td>3</td>
<td>0.04</td>
<td>4.67 seconds</td>
<td>-22.19241</td>
<td>0</td>
</tr>
<tr>
<td>density-based</td>
<td>2</td>
<td>0.230 (38%)</td>
<td>3</td>
<td>0.05</td>
<td>4.67 seconds</td>
<td>-22.19241</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 2 Comparison result of algorithms using weka tool
5. CONCLUSION

After analyzing the results of testing the algorithms we can obtain the following conclusions:

- The performance of K-Means algorithm is better than Hierarchical Clustering algorithm.
- All the algorithms have some ambiguity in some (noisy) data when clustered.
- Density based clustering algorithm is not suitable for data with high variance in density.
- K-Means algorithm is produces quality clusters when using huge dataset.
- Hierarchical clustering algorithm is more sensitive for noisy data.

REFERENCES


AUTHOR

Mr. Bharat Chaudhari doing ME in Department of Computer Science and Engg from Kalol Institute of Technology & Research Centre, Kalol-382721,Gujarat Technological University, published his paper titled A Comparative Study of clustering algorithms using weka tools. In this paper we are trying to compare major clustering algorithms using weka tools.

Mr. Manan Parikh doing ME in Department of Computer Science and Engg from Kalol Institute of Technology & Research Centre, Kalol-382721, Gujarat Technological University, published his paper titled A Comparative Study of clustering algorithms using weka tools.