Review and Comparison between Clustering Algorithms with Duplicate Entities Detection

Purpose

Maryam Bakhshi, Mohammad-Reza Feizi-Derakhshi, Elnaz Zafarani

Department of Computer, Zanjan Branch, Islamic Azad University, Zanjan-Iran

Computer Department, University of Tabriz Tabriz-Iran

Department of Computer Engineering, Tabriz Branch, Islamic Azad University Tabriz-Iran

Abstract —the issue of identifying iterative records issue is one of the challenging issues in the field of databases. As a result, finding appropriate algorithms in this field helps significantly to organize information and extract the correct answer from different queries of database. One of steps of duplicate detection is clustering. Clustering is a classification process of existing data sets into different clusters so that, the similarity among data within each cluster is maximum and similarity among the data of different clusters is at least. The aim of this paper is to find appropriate clustering algorithms for Iteration Detection issues on existing data set. In this paper, 4 algorithms, K-Means, Single-Linkage, DBSCAN and Self-Organizing Maps have been implemented and compared. F1 measure was used in order to measure accuracy and quality of clustering, that according to the obtained results, SOM algorithm obtained high accuracy. F1 measure was used in order to evaluate precision and quality of clustering that by studying the obtained results, the SOM algorithm obtained high F1 measure. Also a comparison between 2 methods, mapping to two dimensional space and statistical average, performed, that according to the results, mapping method is better than average method.

Keywords— Clustering, K-Means, Single-Linkage, DBSCAN, Self-Organizing Maps, F1 Measure.

1. Introduction

Databases play an important role in the information age. Industries and systems in their operations depend on precision of databases. Therefore, the quality of data (or lack thereof) stored in the database have an important impact on the cost of information based systems. Often the data are not carefully controlled of quality and also they have not been defined compatible with various sources of data. Therefore, data quality is often compared with several factors such as input errors (e.g. Microsoft instead of Microsoft), elimination of integrity constraints (e.g. permitting inputs such as age = 476) and various formats for data storage (e.g. St. A, Street A). In worse conditions in the databases which are managed independently, rather than values, structure, concepts and assumptions about data may be different from each other. Consequently, in order to manage better and extract the correct answer from different queries of database, the problem of iteration detection is important. Iteration detection procedure includes three follow steps:

- Field matching
- Record matching
- Clustering

That this research is emphasized on the third stage, clustering, and its ultimate goal is to find appropriate algorithms for iteration detection problem. Method is the following, after the first and two stages and get the degree of similarity between the records, clustering is done to similar records lie in a cluster. The ultimate goal of this research is to find the degree of the most suitable algorithms for existing data set. Used dataset include property information. There are different categories for clustering methods that the following is the most comprehensive:

- Partitioning Clustering
- Hierarchical Clustering
- Density Based Clustering
- Grid Based Clustering
- Model Based Clustering
- Fuzzy Clustering

In this paper, four algorithms of Partitioning, Hierarchical, Density-based and Model-based categories were selected and compared.
2. Related Works

Some researchers have been improved clustering algorithms. Some were presented new algorithms. And some other studied and compared clustering algorithms. In this section, we will review previous studies that presented influence of different factors on efficiency of a number of clustering algorithms and results were compared.

Madjid Khalilian [6] with using of divide and conquer method improves the K-Means algorithm for use in high-dimension data sets. Zhexue [4] extend K-Means algorithm and used it for combined values. K-Modes algorithm used of adaptation heterogeneity measure to deal with combined data that use average instead of mode. They reduce cost function with using of this method, that is a repetition based method and modes update in clustering process has cut the cost function. HE Ling[9] provided a detailed survey of current clustering algorithms in data mining at first, then it makes a comparison among them, presented their scores (merits), and identified the problems to be solved and the new directions in the future according to the application requirements in multimedia domain. Rui [7] presented the survey of clustering algorithms for data sets including in statistics, computer science, and machine learning, and explained their applications in some benchmark data sets, the traveling salesman problem, and bioinformatics and also subjects like adjacent measures and evaluating clustering were discussed. Several tightly related topics, proximity measure, and cluster validation, are also discussed. Treshansky [10] presented a survey of clustering algorithms and paid particular attention to those algorithms that require less amount of knowledge about the domain being clustered. Ossama Abu Abbas [3] studied and compared various clustering algorithms. Algorithms have been compared based on factors: Dataset size, number of clusters, type of dataset and used software.

3. Reviewed Algorithms

3.1 K-Means Algorithm

Algorithm K-Means [1] is one of the most popular iteration based clustering methods. This algorithm has application in some cases in which any data belongs only to one class. This algorithm is an unsupervised algorithm and has iteration in which the data set has been divided into K clusters and data points are randomly assigned to the clusters. Then for each point, the distance of point to the center of cluster has been calculated and target point is assigned to the closest cluster. These steps will be repeated until no point is shifted longer. Characteristics of this algorithm are as follows:

- Always there is k clusters.
- Always at least one point in each cluster is available.
- Clusters are not as hierarchical and do not overlap with each other.
- Each member of a cluster compared with other clusters has the lowest distance from cluster center.

Implementation steps of K-Means algorithm are expressed as follows:

1 - Establish primary centers of clusters with random selection of C point among all data points.

2 - Calculate the membership matrix U using the equation (1):

\[ u_{ij} = \begin{cases} 1 & \text{if } \left\| x_i - c_j \right\|^2 \leq \left\| x_i - c_k \right\|^2, \text{for each } k \neq i, \\ 0 & \text{otherwise} \end{cases} \]  

3 - Calculate membership function using the following equation:

\[ J = \sum_{i=1}^{c} J_i = \sum_{i=1}^{c} \left( \sum_{k \in G_i} \left\| x_k - c_i \right\|^2 \right) \]  

4 - Calculate the new cluster centers using the following equation:

\[ c_i = \frac{1}{|G_i|} \sum_{k \in G_i} x_k \]  

And then return to Step 2.

It is noteworthy that the algorithm performance depends on the initial location of center of clusters.

Therefore there is not any guarantee to reach the expected response by this algorithm.

Advantages of K-Means clustering algorithm

- In the case of large number of variables, this algorithm has higher computing rate than hierarchical approach (If k is small).
- K-Means algorithm produces more dense clusters than the hierarchical method especially when clusters are spherical.

Limitations of K-Means clustering algorithm:

- Difficulty of qualitative comparison of produced clusters
- Fixedness of the number of clusters that makes prediction of k value difficult
- It is not so suitable for non-spherical clusters
• It is sensitive to noisy data

Difference of the number of primary clusters leads to difference of final clusters. So it is better to run algorithm for different values and compare results with each other.

3.3 Single-Linkage Algorithm

This is one of the oldest and simplest methods of clustering methods, and can be considered as a member of hierarchical and exclusive clustering methods. This clustering also called nearest neighbor technique (Nearest Neighbor). In this method for calculating the similarity between cluster A and B are used of the following criteria:

\[ d_{AB} = d_{ij} \leq \min_{p \in A, q \in B} \text{dist}(p, q) \]  

(4)

![Figure 1 Single-Linkage Algorithm](image)

That i is a sample belonged to cluster A and j is a sample belong to a cluster B. In fact, the similarity between two clustering methods is the minimum distance between a member of one than from one another.

3.4 DBSCAN Algorithm

Density based spatial clustering of applications with noise, DBSCAN; rely on a density-based notion of clusters, which is designed to discover clusters of arbitrary shape and also have ability to handle noise. The main task of this algorithm is class identification, i.e. the grouping of objects into meaningful subclasses.

Two global parameters of DBSCAN algorithms are:

• Eps: Maximum radius of the neighbourhood.

• MinPts: Minimum number of points in an Eps-neighbourhood of that point.

Core Object: Object with at least MinPts objects within a radius 'Eps- neighbourhood.

Border Object: Object that on the border of a cluster

NEps (p): \{ q belongs to D | dist (p, q) <= Eps \}

Directly Density-Reachable: A point p is directly density-reachable from a point q w.r.t Eps, MinPts If p belongs to NEps (q)

\[ |\text{NEps}(q)| \geq \text{MinPts} \]

Density-Reachable: A point p is density-reachable from a point q w.r.t Eps, MinPts if there is a chain of points \{p_1, \ldots, p_n\}, \ p_1 = q, \ p_n = p \ such that \ p_{i+1} \ is directly density-reachable from \ p_i.

Density-Connected: A point p is density-connected to a point q w.r.t Eps, MinPts if there is a point ‘o’ such that both, p and q are density-reachable from ‘o’ w.r.t Eps and MinPts.

The algorithm of DBSCAN is as follows:

• Arbitrary selection of a point p

• Retrieve all points density-reachable from p w.r.t Eps and MinPts.

• If p is a core point, then a cluster is formed.

• If p is a border point, no points are density-reachable from p and DBSCAN visits the next point of the database.

• Continue the process until all of the points have been processed.

3.5 Self-Organizing Algorithm

In 1975 Teuvo Kohonen introduced new type of neural network that uses competitive, unsupervised learning. This approach is based on WTA (Winner Takes All) and WTM (Winner Takes Most) algorithms. Therefore, these algorithms will be explained here briefly. The most basic competitive learning algorithm is WTA. When input vector (a pattern) is presented, a distance to each neuron's synaptic weights is calculated. The neuron whose weights are most correlated to current input vector is the winner. Correlation is equal to scalar product of input vector and considered synaptic weights. Only the winning neuron modifies it's synaptic weights to the point presented by input pattern. Synaptic weights of other neurons do not change. The learning process can be described by the following equation:

\[ \| x - w_c \| = \min_{i} \{ \| x - w_i \| \} \]  

(5)

\[ w_c(t + 1) = w_c(t) + \alpha(t) [x(t) - w_c(t)] \]  

(6)

where \( i \) \([0, \text{num of neurons}] \), \( W_c \) represents all synaptic weights of the winning neuron, \( \alpha(t) \) is learning rate in the interval \([0, 1]\) that linearly proportional with t inverse reduced and shows total weights attached to the winning cell and \( x \) stands for current input vector.
In this section WTM strategy describe that is a extension of WTA strategy. The difference between those two algorithms is that many neurons in WTM strategy adapt their synaptic weights in one learning iteration. In this case not only the winner, but also its neighbourhood adapts. The further the neighbouring neuron is from the winner, the smaller the modification which is applied to its weights. This adaptation process can be described as:

\[ w_{i,t+1} = w_i + \eta(t) \times K(x, x_i) \times (x - w_i) \]

For all neurons \( i \) that belong to winner's neighbourhood. \( w_i \) stands for synaptic weights of neuron \( i \) and \( x \) is current input vector. \( \eta \) stands for learning rate and \( N(i, x) \) is a function that defines neighbourhood. Where shows the weights attached to the cells and cells located in the neighbourhood of winning. \( x \) vector is input pattern and \( \eta \) learning rate that have a positive value smaller than the unit. \( K(i, x) \) is Neighbourhood function that is Gaussian kernel that was a descent function and with a way of win cell and time decreases. And thus the cell in farthest neighbourhood will have low change in weights. Neighbourhood function can be described as:

\[ N(i, x) = \begin{cases} \exp \left(-\frac{||x - w_i||^2}{x^2} \right), & w_i \in \Lambda(i, x) \\ 0, & wi \in \Lambda(i, x) \end{cases} \quad (8) \]

In order to train SOM network the Euclidean distance between input vector and weight vectors of all cells should be computed. The cell which has the lowest distance with input vector, in other words the cell which has the most similarity to input pattern is selected as a winner and its adjoined weights change in order to approach input pattern. In addition, adjacent cells are selected and according to their distance to winner cell their weights are modified in the same orientation. The movement of cells and the number of mobile cells is high in the beginning of algorithm and they reach their minimum value due to reducing the rate of learning and neighbour radius. This algorithm draws input vector on one line (in two-dimensional topological state). Figure (2) shows one two dimensional SOM neural network.

Input patterns that are similar to each other, that have minimum Euclidean distance from each other, are also after mapped are placed together. In 1-D network each cell has 2 neighbours, a neighbour on the left and the other cell in the right placed. Two-dimensional network in each cell has four neighbours, which is on the left, right, top and bottom cell are placed.

SOM algorithm can be summarized as follows:

1. Choose weight of all the cells randomly.
2. Apply input pattern to network.
3. Find win cells.
4. Select the neighbour cells.
5. Correct weights attached to the cells and the winner of the neighbour cells according to their Euclidean distance, learning rate and neighbourhood radius.
6. Repeat stages 2 to 5 for the number of distinct and pre-determined periods.

4. Methods and Tests

4.1 Data Set

Applied data set includes data relating to a property information system. This database has high volumes of data. Due to being in different parts, the system has high volumes of redundancy. The effective fields in iteration detection includes: address, owner name, identification code, area, etc. Initially we have used owner name and address fields.

4.2 Evaluation Measure

The goodness measure of one cluster is based on similarities of interior and exterior class. The similarity of within class should be high and this is when then similarity of out of class is so low. In other words, the data in one cluster should be identical and however they should be different from other cluster data. In order to evaluate the quality of clustering we need one qualitative measure of clustering. There are several measures for clustering and classifying that in this article we have used F-Measure.

4.3.1 F1-Measure

There are two states “belongs” and “not belongs” in the problems of binary classification in order to assign data to one particular category. Data which belong to one category are called positive and data which do not belong to one category are called negative. Consider the variables of table I., if ‘1’ was the dataset which is really positive, and ‘J’ is a dataset which was predicted as positive by clustering algorithm, according to table I., variables a, b, c and d are defined as follows:

![Figure 2 SOM with the neighbourhood of two-dimensional input vectors](image-url)
\[ a = I \cap J \quad b = I \cap \overline{J} \]
\[ c = \overline{I} \cap J \quad d = \overline{I} \cap \overline{J} \]  \hspace{1cm} (9)

**Table 1 Variables**

<table>
<thead>
<tr>
<th></th>
<th>Positive Predicated</th>
<th>Negative Predicated</th>
</tr>
</thead>
<tbody>
<tr>
<td>Real Positive</td>
<td>a</td>
<td>b</td>
</tr>
<tr>
<td>Real Negative</td>
<td>c</td>
<td>d</td>
</tr>
</tbody>
</table>

In this section two variables are defined that are not effective separately. Variable \( P \) (Precision) that is defined as follows shows that how many data which have been predicted as positive are really positive.

\[ \frac{a}{a + b} \]  \hspace{1cm} (10)

And \( R \) (Recall) which is defined as follows shows that how many of the true positives have been predicted accurately.

\[ \frac{c}{c + d} \]  \hspace{1cm} (11)

According to the mentioned information, parameter \( P \) is obtained by dividing the number of common clusters into the number of computed clusters with clustering algorithm (system) and parameter \( R \) is obtained by dividing the number of common clusters into the number of accurate clusters. A good clustering algorithm should consider both parameters \( P \) and \( R \). These two measures are combined and form F-Measure.

\[ P = \frac{\text{number of common clusters}}{\text{number of system calculated clusters}} = \frac{|I \cap J|}{|I|} \]  \hspace{1cm} (12)

\[ R = \frac{\text{number of common clusters}}{\text{number of revnet target clusters}} = \frac{|I \cap J|}{|J|} \]  \hspace{1cm} (13)

\[ F_P = \frac{2 \times P \times R}{P + R} \]  \hspace{1cm} (14)

Often, \( P \) replaced with 1 as follow:

\[ \beta = 1 - \frac{1}{1 + \frac{2 \times P \times R}{P + R}} \]  \hspace{1cm} (15)

**5. Methods**

As was mentioned in the first section, duplicate detection includes three steps. The first step is field matching. At this stage to detect similarities between the fields, distance-based algorithms are used. Implemented algorithms in this phase are jaro and jaro winkler. That has achieved similar results almost. Basically jaro algorithm to compare names will do better. For this purpose, similarity between owner name and address fields achieved with using of jaro algorithm. And the degree of similarity obtained from the first stage, used as input of record matching. To perform record matching of the two follow methods used:

**Mapping to two dimensional spaces:** Owner name and address fields are considered as dimensions. Methods are as follows that a record be considered as origin record, and other records according to similarity degree to origin record, lie in that space. In other words, each record according to degree of similarity assigns a coordinate of space.

**Use Statistical average:** By accounting mean of field similarity degree of records. Similarity degree of records obtained.

**5.1 Comparison of record Matching Methods**

A comparison performed between the results of two methods is mapped to two-dimensional space and the statistical average. As Figure 3 shows mapping method has better results than the statistical average. Vector ‘x’ in the graph relates to precision field and vector ‘y’ relates to the F1 measure field of table II.
6. Results

Table II. Shows the results of the implementation of mentioned algorithms on existing data set. Using criteria defined in section IV, P, R and F1 measure for each algorithm computed and results were compared. As the results are observed, SOM (WMA) algorithm includes high F1-measure. In other words, this algorithm has a high degree of accuracy is compared with other algorithms. The result is The WMA is a better convergence than that of the WTA.

Table 2 Evaluation Results

<table>
<thead>
<tr>
<th>Method</th>
<th>Used Alg.</th>
<th>Num. of System Clusters</th>
<th>Num. of correct clusters</th>
<th>Num. of common clusters</th>
<th>precision</th>
<th>recall</th>
<th>F1 measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mapping to two dimensional space</td>
<td>K-Means</td>
<td>27</td>
<td>38</td>
<td>19</td>
<td>0.7</td>
<td>0.58</td>
<td>0.58</td>
</tr>
<tr>
<td></td>
<td>DBSCAN (eps: 0.005, minpt= 1)</td>
<td>23</td>
<td>38</td>
<td>10</td>
<td>0.43</td>
<td>0.26</td>
<td>0.33</td>
</tr>
<tr>
<td></td>
<td>Single-Linkage</td>
<td>25</td>
<td>38</td>
<td>20</td>
<td>0.8</td>
<td>0.52</td>
<td>0.62</td>
</tr>
<tr>
<td></td>
<td>SOM(WTA)</td>
<td>28</td>
<td>38</td>
<td>22</td>
<td>0.785</td>
<td>0.578</td>
<td>0.66</td>
</tr>
<tr>
<td></td>
<td>SOM(WMA)</td>
<td>31</td>
<td>38</td>
<td>27</td>
<td>0.87</td>
<td>0.71</td>
<td>0.78</td>
</tr>
<tr>
<td>Use average</td>
<td>K-Means</td>
<td>17</td>
<td>38</td>
<td>12</td>
<td>0.44</td>
<td>0.31</td>
<td>0.36</td>
</tr>
<tr>
<td></td>
<td>DBSCAN (eps: 0.005, minpt= 1)</td>
<td>26</td>
<td>38</td>
<td>8</td>
<td>0.23</td>
<td>0.15</td>
<td>0.19</td>
</tr>
<tr>
<td></td>
<td>Single-Linkage</td>
<td>18</td>
<td>38</td>
<td>10</td>
<td>0.55</td>
<td>0.26</td>
<td>0.35</td>
</tr>
<tr>
<td></td>
<td>SOM(WTA)</td>
<td>22</td>
<td>38</td>
<td>12</td>
<td>0.54</td>
<td>0.31</td>
<td>0.39</td>
</tr>
<tr>
<td></td>
<td>SOM(WMA)</td>
<td>25</td>
<td>38</td>
<td>17</td>
<td>0.68</td>
<td>0.447</td>
<td>0.53</td>
</tr>
</tbody>
</table>
7. Conclusion

The purpose of this study to obtain the appropriate algorithms to detect duplicate records in the data set is available. Algorithms K-Means, Single-Linkage, DBSCAN, and Self-Organizing Maps were implemented and compared. According to the results, SOM(WMA) algorithm achieve to higher F1 measure. Also done a comparison between the two methods, mapping to 2-D space and statistical average, that as figure 3. shows mapping technique has better results than average method.

References