Integrating Fuzzy $c$-Means Clustering with PostgreSQL *

© Ruslan Miniakhmetov
South Ural State University
tavein@gmail.com
M.Sc. advisor Mikhail Zymbler

Abstract

Many data sets to be clustered are stored in relational databases. Having a clustering algorithm implemented in SQL provides easier clustering inside a relational DBMS than outside with some alternative tools. In this paper we propose Fuzzy $c$-Means clustering algorithm adapted for PostgreSQL open-source relational DBMS.

1 Introduction

Integrating clustering algorithms is a topic issue for database programmers [11]. Such an approach, on the one hand, encapsulates DBMS internal details from application programmer. On the other hand, it allows to avoid overhead connected with export data outside a relational DBMS. The Fuzzy $c$-Means (FCM) [9, 6, 2] clustering algorithm provides a fuzzy clustering of data. Currently this algorithm have many implementations on a high-level programming languages [5, 7]. For implementation the FCM algorithm in SQL we choose an open-source PostgreSQL DBMS [15].

The paper is organized as follows. Section 2 introduces basic definitions and an overview of the FCM algorithm. Section 3 proposes implementation of the FCM in SQL called pgFCM. Section 4 briefly discusses related work. Section 5 contains conclusion remarks and directions for future work.

2 The Fuzzy $c$-Means Algorithm

K-Means [10] is one of the most popular clustering algorithms, it is a simple and fairly fast [3]. The FCM algorithm generalizes K-Means to provide fuzzy clustering, where data vectors can belong to several partitions (clusters) at the same time with a given weight (membership degree). To describe FCM we use the following notation:

- $d \in \mathbb{N}$ — dimensionality of a data vectors (or data items) space;
- $l \in \mathbb{N}: 1 \leq l \leq d$ — subscript of the vector’s coordinate;
- $n \in \mathbb{N}$ — cardinal number of training set;
- $X \subseteq \mathbb{R}^d$ — training set for data vectors;
- $i \in \mathbb{N}: 1 \leq i \leq n$ — vector subscript in a training set;
- $x_i \in X$ — the $i$-th vector in the sample;
- $k \in \mathbb{N}$ — number of clusters;
- $j \in \mathbb{N}: 1 \leq j \leq k$ — cluster number;
- $C \subseteq \mathbb{R}^{k \times d}$ — matrix with clusters’ centers (centroids);
- $c_j \in \mathbb{R}^d$ — center of cluster $j$, $d$-dimensional vector;
- $x_{il}, c_{jl} \in \mathbb{R}$ — $l$-s coordinates of vectors $x_i$ and $c_j$ respectively;
- $U \subseteq \mathbb{R}^{n \times k}$ — matrix with membership degrees, where $u_{ij} \in \mathbb{R}$: $0 \leq u_{ij} \leq 1$ is a membership degree between vector $x_i$ and cluster $j$;
- $\rho(x_i, c_j)$ — distance function, defines a membership degree between vector $x_i$ and cluster $j$;
- $m \in \mathbb{R}: m > 1$ — the fuzzyification degree of objective function;
- $J_{FCM}$ — objective function.

The FCM is based on minimization of the objective function $J_{FCM}$:

$$J_{FCM}(X,k,m) = \sum_{i=1}^{n} \sum_{j=1}^{k} u_{ij}^m \rho^2(x_i, c_j)$$  \hspace{1cm} (1)

Fuzzy clustering is carried out through an iterative optimization of the objective function (1). Membership matrix $U$ and centroids $c_{jl}$ are updated using the following formulas:

$$u_{ij} = \sum_{l=1}^{k} \left( \frac{\rho(x_i, c_j)}{\rho(x_i, c_l)} \right)^{\frac{2}{m}}$$  \hspace{1cm} (2)

$$\forall j, l \hspace{0.2cm} c_{jl} = \frac{\sum_{i=1}^{n} u_{ij}^m \cdot x_{il}}{\sum_{i=1}^{n} u_{ij}^m}$$  \hspace{1cm} (3)

* This paper is supported by the Russian Foundation for Basic Research (grant No. 09-07-00241-a).
Let $s$ be a number of iteration, $u_{ij}^{(s)}$ and $u_{ij}^{(s+1)}$ are elements of matrix $U$ on steps $s$ and $s+1$ respectively, and $\varepsilon \in (0, 1) \subset \mathbb{R}$ is a termination criterion. Then the termination condition can be written as follows:

$$\max_{ij} \{|u_{ij}^{(s+1)} - u_{ij}^{(s)}|\} < \varepsilon \quad (4)$$

Objective function (1) converges to a local minimum (or a saddle point) [1].

**Algorithm 1 The Fuzzy c-Means Algorithm**

**Input:** $X, k, m, \varepsilon$

**Output:** $U$

1. $s := 0, U^{(0)} := (u_{ij})$ \{initialization\}
2. repeat
3. \{computation of new centroids’ coordinates\}
   Compute $C^{(s)} := (c_j)$ using formula (3)
   where $u_{ij} \in U^{(s)}$
4. \{update matrices values\}
   Compute $U^{(s)}$ and $U^{(s+1)}$ using formula (2)
5. $s := s + 1$
6. until $\max_{ij} \{|u_{ij}^{(s)} - u_{ij}^{(s-1)}|\} \geq \varepsilon$

Algorithm 1 shows the basic FCM. The input of algorithm receives a set of data vectors $X = (x_1, x_2, \ldots, x_n)$, number of clusters $k$, fuzzyfication degree $m$, and termination criterion $\varepsilon$. The output is a matrix of membership degrees $U$.

### 3 Implementation of Fuzzy c-Means Algorithm using PostgreSQL

In this section we suggest pgFCM algorithm as a way to integrate FCM algorithm with PostgreSQL DBMS.

#### 3.1 General Definitions

To integrate FCM algorithm with a relational DBMS it is necessary to perform matrixes $U$ and $X$ as relational tables. Subscripts for identification elements of relational tables are presented in Table 1 (numbers $n, k, d$ a defined above in a section 2).

As a function of distance $\rho(x_i, c_j)$, without loss of generality, we use the Euclidean metric:

$$\rho(x_i, c_j) = \sqrt{\sum_{t=1}^{d} (x_{it} - c_{jt})^2} \quad (5)$$

#### 3.2 Database Scheme

Table 2 summarizes database scheme of pgFCM algorithm (underlined columns are primary keys).

In order to store sample of a data vectors from set $X$ it is necessary to define table $SH(i, x_1, x_2, \ldots, xd)$. Each row of sample stores vector of data with dimension $d$ and subscript $i$. Table $SH$ has $n$ rows and column $i$ as a primary key.

FCM steps demand aggregation of vector coordinates (sum, maximum, etc.) from set $X$. However, because of its definition, table $SH$ does not allow using SQL aggregation functions. To avoid this obstacle we define a table $SV(i, l, val)$, which contains $n \cdot d$ rows and has a composite primary key $(i, l)$. Table $SV$ represents a data sample from table $SH$ and supports SQL aggregation functions max and sum.

Due to store coordinates of cluster centroids temporary table $C(j, l, val)$ is defined. Table $C$ has $k \cdot d$ rows and the composite primary key $(j, l)$. Like the table $SV$, structure of table $C$ allows to use aggregation functions.

In order to store distances $\rho(x_i, c_j)$ table $SD(i, j, dist)$ is used. This table has $n \cdot k$ rows and the composite primary key $(i, j)$.

Table $U(i, j, val)$ stores membership degrees, calculated on $s$-th step. To store membership degrees on $s+1$ step similar table $UT(i, j, val)$ is used. Both tables have $n \cdot k$ rows and the composite primary key $(i, j)$.

Finally, table $P(d, k, n, s, delta)$ stores iteration number $s$ and the result of computation function (6) $delta$ for this iteration number. Number of rows in table $P$ depends on the number of iterations.

#### 3.3 The pgFCM Algorithm

The pgFCM algorithm is implemented by means of a stored function in PL/pgSQL language. Algorithm 2 shows the main steps of the pgFCM.

---

**Table 2: Relational Tables of pgFCM Algorithm**

<table>
<thead>
<tr>
<th>No.</th>
<th>Table</th>
<th>Semantics</th>
<th>Columns</th>
<th>Number of rows</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$SH$</td>
<td>training set for data vectors (horizontal form)</td>
<td>$i, x_1, x_2, \ldots, x_d$</td>
<td>$n$</td>
</tr>
<tr>
<td>2</td>
<td>$SV$</td>
<td>training set for data vectors (vertical form)</td>
<td>$i, l, val$</td>
<td>$n \cdot d$</td>
</tr>
<tr>
<td>3</td>
<td>$C$</td>
<td>centroids’ coordinates</td>
<td>$j, l, val$</td>
<td>$k \cdot d$</td>
</tr>
<tr>
<td>4</td>
<td>$SD$</td>
<td>distances between $x_i$ and $c_j$</td>
<td>$i, j, dist$</td>
<td>$n \cdot k$</td>
</tr>
<tr>
<td>5</td>
<td>$U$</td>
<td>degree of membership vector $x_i$ to a cluster $c_j$ on step $s$</td>
<td>$i, j, val$</td>
<td>$n \cdot k$</td>
</tr>
<tr>
<td>6</td>
<td>$UT$</td>
<td>degree of membership vector $x_i$ to a cluster $c_j$ on step $s+1$</td>
<td>$i, j, val$</td>
<td>$n \cdot k$</td>
</tr>
<tr>
<td>7</td>
<td>$P$</td>
<td>result of computation function $\delta$ (6) on step $s$</td>
<td>$d, k, n, s, \text{delta}$</td>
<td>$s$</td>
</tr>
</tbody>
</table>

**Table 1: Data Elements Subscripts**

<table>
<thead>
<tr>
<th>Subscript</th>
<th>Range</th>
<th>Semantics</th>
</tr>
</thead>
<tbody>
<tr>
<td>$i$</td>
<td>$\tilde{1}$</td>
<td>vector subscript</td>
</tr>
<tr>
<td>$j$</td>
<td>$\tilde{1}, k$</td>
<td>cluster subscript</td>
</tr>
<tr>
<td>$l$</td>
<td>$\tilde{1}, d$</td>
<td>vector’s coordinate subscript</td>
</tr>
</tbody>
</table>

To compute the termination criterion 4 we introduce the function $\delta$ as follows:

$$\delta = \max_{ij} \{|u_{ij}^{(s+1)} - u_{ij}^{(s)}|\} \quad (6)$$
The pgFCM Algorithm

**Input:** SH, k, m, eps
**Output:** U

1: {initialization}
   Create and initialize temporary tables (U, P, SV, etc.)
2: repeat
3: {computations}
4: Compute centroids coordinates. Update table C.
5: Compute distances \( p(x_i, c_j) \). Update table SD.
6: Compute membership degrees \( UT = (u_{ij}) \). Update table UT.
7: {update}
   Update tables P and U.
8: {check for termination}
9: until \( P.\delta \geq \epsilon \)

The input set of data vectors \( X \) stored in table SH. Fuzzyfication degree \( m \), termination criterion \( \epsilon \), and number of clusters \( k \) are function parameters. The table \( U \) contains a result of pgFCM work.

### 3.4 Initialization

Initialization of tables SV, U, and P provided by SQL-code I1, I2, and I3 respectively. Table SV is formed by sampling records from the table SH.

**I1:**
```
INSERT INTO SV
SELECT SH.i, 1, x1 FROM SH;
...
INSERT INTO SV
SELECT SH.i, d, xd FROM SH;
```

For table U a membership degree between data vector \( x_i \) and cluster \( j \) takes 1 for all \( i = j \).

**I2:**
```
INSERT INTO U (i, j, val)
VALUES (1, 1, 0);
...
INSERT INTO U (i, j, val)
VALUES (j, j, 1);
...
INSERT INTO U (i, j, val)
VALUES (n, k, 0);
```

In other words, as a start coordinates of centroids, first \( d \) data vectors from sample \( X \) are used.

\[ \forall i = j \quad u_{ij} = 1 \Rightarrow c_j = x_i \]

When initializing the table \( P \), the number of points \( k \) is taken as a parameter of the function pgFCM. A data vectors space dimensionality \( d \) and a cardinal number of the training set \( n \) also provided by function pgFCM parameters. The iteration number \( s \) and \( \delta \) initializes as zeros.

**I3:**
```
INSERT INTO P(d, k, n, s, \( \delta \))
VALUES (d, k, n, 0, 0);
```

### 3.5 Computations

According to Algorithm 2, the computation stage is split to the following three sub-steps: computation coordinates of centroids, computation of distances, and computation membership degrees, marked as C1, C2, and C3 respectively.

**C1:**
```
SELECT R1.j, R1.l, R1.s1 / R2.s2 AS val
FROM (SELECT j, 1,
      sum(U.val^m * SV.val)
      AS s1
FROM U, SV
WHERE U.i = SV.i
GROUP BY j, l) AS R1,
(SELECT j, sum(U.val^m) AS s2
FROM U
GROUP BY j) AS R2
WHERE R1.j = R2.j;
```

**C2:**
```
SELECT i, j,
    sqrt(sum((SV.val - C.val)^2)))
AS dist
FROM SV, C
WHERE SV.l = C.l
GROUP BY i, j;
```

Through the FCM, computations of the distances provide by formula (2). In formula (3) the fraction’s numerator does not depend on \( t \), then we can rewrite this formula as follows:

\[ u_{ij} = \rho^{\frac{m}{m-1}}(x_i, c_j) \cdot \left( \sum_{t=1}^{k} \rho^{\frac{m}{m-1}}(x_i, c_t) \right)^{-1} \]  \( (7) \)

Thus, the computation of membership degrees can be written as follows:

**C3:**
```
SELECT i, j,
    SD.dist^(2.0^(1.0-m)) *
    SD1.den AS val
FROM (SELECT i,
      1.0 /
      sum(dist^(2.0^(m-1.0))))
    AS den
FROM SD
GROUP BY i) AS SD1, SD
WHERE SD.i = SD1.i;
```

### 3.6 Update

Update stage of the pgFCM modifies \( P \) and \( U \) tables as shown below in U1 and U2 respectively.

**U1:**
```
INSERT INTO P
SELECT L.d, L.k, L.n, L.s + 1 AS s,
      E.\( \delta \)
FROM (SELECT i, j,
      max(abs(UT.val - U.val))
      AS \( \delta \)
FROM U, UT
GROUP BY i, j) AS E,
(SELECT d, k, n, max(s)
FROM P
GROUP BY d, k, n) AS L
GROUP BY d, k, n, s
```

Table \( UT \) stores temporary membership degrees to be inserted into table \( U \). To provide the rapid removal all the table \( U \) rows, obtained at the previous iteration, we use the truncate operator.
3.7 Check

This stage is the final for the algorithm pgFCM. On each iteration the termination condition (4) must be checked.

To implement the check, the result $\delta$ of the function (6) from table $P$ is stored in the temporary variable $tmp$.

CH1: \[
\begin{aligned}
&\text{SELECT } \delta \text{ INTO } tmp \\
&\text{FROM } P, \left( \text{SELECT } d, k, n, \\
&m_{\text{max}}(s) \text{ AS } \text{max}_s \right) \text{ AS } L \\
&\text{WHERE } P.s = L.\text{max}_s \text{ AND } P.d = L.d \\
&\text{AND } P.k = L.k \text{ AND } P.n = L.n;
\end{aligned}
\]

After selecting the $\delta$, we need to check the condition $\delta < \varepsilon$. Then if this condition is true we should stop, otherwise, work will be continued.

CH2: \[
\begin{aligned}
&\text{IF } (\text{tmp } < \text{eps}) \text{ THEN} \\
&\text{RETURN;} \\
&\text{END IF;}
\end{aligned}
\]

The final result of the algorithm pgFCM will be stored in table $U$.

4 Related Work

Research on integrating data mining algorithms with relational DBMS includes the following. Association rules mining is explored in [13]. General data mining primitives are suggested in [4]. Primitives for decision trees mining are proposed in [8].

Our research was inspired by papers [11, 12], where integrating K-Means clustering algorithm with relational DBMS, was carried out. The way we exploit is similar to mentioned above. The main contribution of the paper is an extension of results presented in [11, 12] for the case where data vectors may belong to several clusters. Such a case is very important in problems connected with medicine data analysis [14, 16]. To the best of our knowledge there are no papers devoted to implementing fuzzy clustering with relational DBMS.

5 Conclusion

In this paper we have proposed the pgFCM algorithm. pgFCM implements Fuzzy c-Means clustering algorithm and processes data stored in relational tables using PostgreSQL open-source DBMS. There are following issues to continue our research. Firstly, we plan to investigate pgFCM scalability using both synthetic and real data sets. The second direction of our research is developing a parallel version of pgFCM for distribution memory multiprocessors.

References


