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 clusterization algorithm implemented in SQL provides easier clusterization 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;
- $l \in \mathbb{N} : 1 \leqslant l \leqslant d$ — subscript of the vector's coordinate;
- $n \in \mathbb{N}$ — cardinal number of training set;
- $X \subset \mathbb{R}^d$ — training set for data vectors;
- $i \in \mathbb{N} : 1 \leqslant i \leqslant n$ — vector subscript in a training set;
- $x_i \in X$ — the i-th vector in the sample;
- $k \in \mathbb{N}$ — number of clusters;
- $j \in \mathbb{N} : 1 \leqslant j \leqslant k$ — cluster number;
- $C \subset \mathbb{R}^{k \times d}$ — 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;
- $U \subset \mathbb{R}^{n \times k}$ — matrix with membership degrees, where $u_{ij} \in \mathbb{R}$: $0 \leqslant u_{ij} \leqslant 1$ 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 fuzzyfication degree of objective function;
- J_{FCM} — objective function.

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

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

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

$$u_{ij} = \sum_{t=1}^{k} \left(\frac{\rho(x_i, c_j)}{\rho(x_i, c_t)} \right)^{\frac{2}{1-m}} \qquad (2)$$

$$\forall j, l \quad c_{jl} = \frac{\sum\limits_{i=1}^{n} u_{ij}^m \cdot x_{il}}{\sum\limits_{i=1}^{n} u_{ij}^m} \qquad (3)$$

* This paper is supported by the Russian Foundation for Basic Research (grant No. 09-07-00241-a).

Proceedings of the Spring Researcher's Colloquium on Database and Information Systems, Moscow, Russia, 2011

Table 2: Relational Tables of pgFCM Algorithm

No.	Table	Semantics	Columns	Number of rows
1	SH	training set for data vectors (horizontal form)	$\underline{i}, x1, x2, \ldots, xd$	n
2	SV	training set for data vectors (vertical form)	$\underline{i,l}, val$	$n{\cdot}d$
3	C	centroids' coordinates	$\underline{j,l}, val$	$k{\cdot}d$
4	SD	distances between x_i and c_j	$\underline{i,j}, dist$	$n{\cdot}k$
5	U	degree of membership vector x_i to a cluster c_j on step s	$\underline{i,j}, val$	$n{\cdot}k$
6	UT	degree of membership vector x_i to a cluster c_j on step $s{+}1$	$\underline{i,j}, val$	$n{\cdot}k$
7	P	result of computation function δ (6) on step s	$\underline{d,k,n}, s, delta$	s

Let s is 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:

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

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

Algorithm 1 The Fuzzy c-Means Algorithm

Input: X, k, m, ε
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 matrixes values}
Compute $U^{(s)}$ and $U^{(s+1)}$ using formula (2)
5: $s := s + 1$
6: **until** $\max_{ij}\{|u_{ij}^{(s)} - u_{ij}^{(s-1)}|\} \geqslant \varepsilon$

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 ε. 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:

$$\rho(x_i, c_j) = \sqrt{\sum_{l=1}^{d} (x_{il} - c_{jl})^2} \tag{5}$$

Table 1: Data Elements Subscripts

Subscript	Range	Semantics
i	$\overline{1,n}$	vector subscript
j	$\overline{1,k}$	cluster subscript
l	$\overline{1,d}$	vector's coordinate subscript

To compute the termination criterion 4 we introduce the function δ as follows:

$$\delta = \max_{ij}\{|u_{ij}^{(s+1)} - u_{ij}^{(s)}|\} \tag{6}$$

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 $SH(\underline{i}, x1, x2, \ldots, xd)$. 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 $SV(\underline{i,l}, val)$, which contains $n{\cdot}d$ rows and have a composite primary key (i, l). Table SV represents a data sample from table SH ans supports SQL aggregation functions max and sum.

Due to store coordinates of cluster centroids temporary table $C(\underline{j,l}, val)$ 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 $SD(\underline{i,j}, dist)$ is used. This table has $n{\cdot}k$ rows and the composite primary key (i, j).

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

Finally, table $P(\underline{d,k,n}, s, delta)$ 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.

Algorithm 2 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 $\rho(x_i, c_j)$. Update table SD.
6: Compute membership degrees $UT = (ut_{ij})$.
   Update table UT.
7: {update}
   Update tables P and U.
8: {check for termination}
9: **until** $P.delta \geqslant eps$

The input set of data vectors X stored in table SH. Fuzzyfication degree m, termination criterion ε, 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 splitted 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: INSERT INTO C
      SELECT R1.j, R1.l,
             R1.s1 / R2.s2 AS val
      FROM (SELECT j, l,
                   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: INSERT INTO SD
      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:

$$u_{ij} = \rho^{\frac{2}{1-m}}(x_i, c_j) \cdot \left(\sum_{t=1}^{k} \rho^{\frac{2}{m-1}}(x_i, c_t) \right)^{-1} \qquad (7)$$

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

```
C3: INSERT INTO UT
      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) AS R
```

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.

```
U2: TRUNCATE U;
    INSERT INTO U
      SELECT * FROM UT;
```

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: SELECT delta INTO tmp
     FROM P, (SELECT d, k, n,
                     max(s) AS max_s
              FROM P
              GROUP BY d, k, n) AS L
     WHERE P.s = L.max_s AND P.d = L.d
     AND P.k = L.k AND P.n = L.n;
```

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: IF (tmp < eps) THEN
       RETURN;
     END IF;
```

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 synthetical and real data sets. The second direction of our research is developing a parallel version of pgFCM for distribution memory multiprocessors.

References

[1] J. Bezdek, R. Hathaway, M. Sobin, and W. Tucker. Convergence Theory for Fuzzy c-means: Counterexamples and Repairs. *IEEE Trans. Syst. Man Cybern.*, 17:873–877, October 1987.

[2] J. C. Bezdek. *Pattern Recognition with Fuzzy Objective Function Algorithms*. Kluwer Academic Publishers, Norwell, MA, USA, 1981.

[3] P. S. Bradley, U. M. Fayyad, and C. Reina. Scaling Clustering Algorithms to Large Databases. In *KDD*, pages 9–15, 1998.

[4] J. Clear, D. Dunn, B. Harvey, M. Heytens, P. Lohman, A. Mehta, M. Melton, L. Rohrberg, A. Savasere, R. Wehrmeister, and M. Xu. Non-Stop SQL/MX primitives for knowledge discovery. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '99, pages 425–429, New York, NY, USA, 1999. ACM.

[5] E. Dimitriadou, K. Hornik, F. Leisch, D. Meyer, and Weingessel A. Machine Learning Open-Source Package 'r-cran-e1071', 2010. http://cran.r-project.org/web/packages/e1071/index.html.

[6] J. C. Dunn. A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters. *Journal of Cybernetics*, 3:32–57, 1973.

[7] Apache Software Foundation, I. Drost, T. Dunning, J. Eastman, O. Gospodnetic, G. Ingersoll, J. Mannix, S. Owen, and K. Wettin. Apache Mahout, 2010. https://cwiki.apache.org/confluence/display/MAHOUT/Fuzzy+K-Means.

[8] G. Graefe, U. M. Fayyad, and S. Chaudhuri. On the Efficient Gathering of Sufficient Statistics for Classification from Large SQL Databases. In *KDD*, pages 204–208, 1998.

[9] A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: a review. *ACM Comput. Surv.*, 31:264–323, September 1999.

[10] J. B. MacQueen. Some Methods for Classification and Analysis of MultiVariate Observations. In L. M. Le Cam and J. Neyman, editors, *Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297. University of California Press, 1967.

[11] C. Ordonez. Programming the K-means clustering algorithm in SQL. In W. Kim, R. Kohavi, J. Gehrke, and W. DuMouchel, editors, *KDD*, pages 823–828. ACM, 2004.

[12] C. Ordonez. Integrating K-Means Clustering with a Relational DBMS Using SQL. *IEEE Trans. Knowl. Data Eng.*, 18(2):188–201, 2006.

[13] S. Sarawagi, S. Thomas, and R. Agrawal. Integrating association rule mining with relational database

systems: alternatives and implications. In *Proceedings of the 1998 ACM SIGMOD international conference on Management of data*, SIGMOD '98, pages 343–354, New York, NY, USA, 1998. ACM.

[14] A. I. Shihab. *Fuzzy Clustering Algorithms and their Applications to Medical Image Analysis*. PhD thesis, University of London, 2000.

[15] M. Stonebraker, L. A. Rowe, and M. Hirohama. The Implementation of POSTGRES. *IEEE Trans. on Knowl. and Data Eng.*, 2:125–142, March 1990.

[16] D. Zhang and S. Chen. A Novel Kernelized Fuzzy c-Means Algorithm with Application in Medical Image Segmentation. *Artificial Intelligence in Medicine*, 32:37–50, 2004.