

Южно-Уральский государственный университет
Факультет Вычислительной математики и информатики
Кафедра системного программирования
Выпускная квалификационная работа

Интеграция алгоритма кластеризации Fuzzy c-Means в СУБД PostgreSQL

Р.М. Миниахметов

Рецензент:

кандидат физ.-мат. наук
Т.Ю. Лымарь

Научный руководитель:

кандидат физ.-мат. наук, доцент
М.Л. Цымблер

25 июня 2011 г.

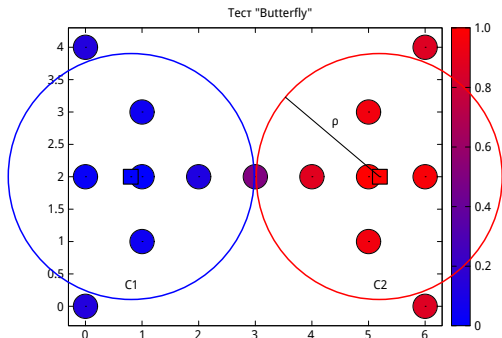
- Сверхбольшие объемы данных
- Программное обеспечение с открытым исходным кодом
- Интеллектуальный анализ данных (Data Mining) и реляционные СУБД
- Нечеткая кластеризация в медицинских исследованиях
 - Количество различных параметров более 10
 - Частота дискретизации 10 мс
 - Объем необработанных данных по пациенту за день исследований 20 МБайт

Цель работы — интегрировать алгоритм нечеткой кластеризации данных в СУБД PostgreSQL.

Задачи:

- 1 Проектирование алгоритма нечеткой кластеризации данных на языке реляционных баз данных SQL
- 2 Реализация алгоритма, адаптированного для СУБД PostgreSQL
- 3 Проведение тестирования разработанного алгоритма
- 4 Проведение вычислительных экспериментов по исследованию производительности разработанного алгоритма

Нечеткая кластеризация



- k — количество кластеров;
- N — количество векторов;
- m — степень нечеткости целевой функции;
- $x_i \in X$ — i -й вектор данных входного множества X , $|X| = N$;
- $c_j \in C$ — центр кластера j , вектор размерности d (центроид);
- C — множество центроидов, $|C| = k$;
- u_{ij} — функция принадлежности;
- $\rho(x_i, c_j)$ — функция расстояния.

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

Алгоритм Fuzzy c-Means

Вход: X — входное множество, k — количество кластеров,
 m — степень нечеткости, ε — точность кластеризации

Выход: U — матрица степеней принадлежности

// Инициализация

$s := 0, U^{(0)} := random(0..1)$

повторять

// Вычисление новых координат центроидов

$$C^{(s)} := (c_j), \text{ где } c_{jl} = \frac{\sum_{i=1}^n u_{ij}^m \cdot x_{il}}{\sum_{i=1}^n u_{ij}^m}$$

// Обновление матрицы степеней принадлежности

$$U^{(s+1)} := (u_{ij}), \text{ где } u_{ij} = \sum_{t=1}^k \left(\frac{\rho(x_i, c_j)}{\rho(x_i, c_t)} \right)^{\frac{2}{1-m}}$$

пока $\max_{ij} \{|u_{ij}^{(s)} - u_{ij}^{(s-1)}|\} > \varepsilon$

Входное множество данных

	x_1	\dots	x_d
1	1.0	\dots	2.1
\vdots	\vdots	\ddots	\vdots
n	3.4	\dots	2.9

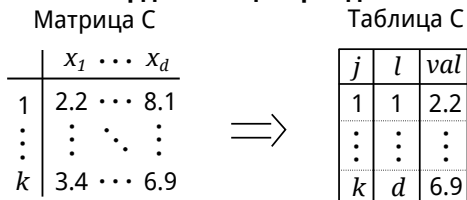
 \Rightarrow

i	x_1	\dots	x_d
1	1.0	\dots	2.1
\vdots	\vdots	\ddots	\vdots
n	3.4	\dots	2.9

i	l	val
1	1	1.0
\vdots	\vdots	\vdots
n	d	2.9

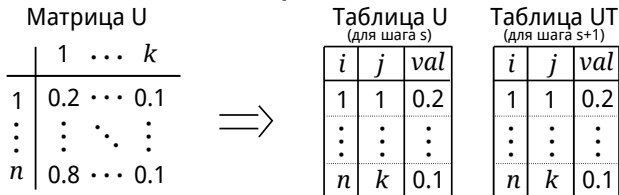
№	Таблица	Семантика	Атрибуты	Записи
1	SH	Выборка векторов данных	$\underline{i}, x_1, x_2, \dots, x_d$	n
2	SV	Выборка векторов данных	\underline{i}, l, val	$n \cdot d$

Координаты центроидов



№	Таблица	Семантика	Атрибуты	Записи
3	C	Координаты центроидов	\underline{j}, l, val	$k \cdot d$

Степени принадлежности



№	Таблица	Семантика	Атрибуты	Записи
4	U	Степени принадлежности вектора x_i кластеру j на шаге s	\underline{i}, j, val	$n \cdot k$
5	UT	Степени принадлежности вектора x_i кластеру j на шаге $s+1$	\underline{i}, j, val	$n \cdot k$

Вспомогательные таблицы

№	Таблица	Семантика	Атрибуты	Записи
6	<i>SD</i>	Расстояния между x_i и c_j	$\underline{i, j}, dist$	$n \cdot k$
7	<i>P</i>	Значение функции $\delta = \max_{ij} \{ u_{ij}^{(s+1)} - u_{ij}^{(s)} \}$ на текущей итерации	$\underline{d, k, n, s, delta}$	s

Вход: SH — выборка векторов, k — количество кластеров,
 m — степень нечеткости, eps — точность кластеризации

Выход: U — таблица степеней принадлежности

- - Инициализация

Создание и инициализация таблиц U, P, SV

повторять

- - Вычисления

Вычислить координаты центроидов (таблица C)

Вычислить расстояния (таблица SD)

Вычислить степени принадлежности $UT = (ut_{ij})$ (таблица UT)

- - Обновление

Обновить таблицы P и U

- - Проверка завершения

пока $P.delta > eps$

Алгоритм pgFCM

Интерфейс функции pgFCM

- Нечеткая кластеризация таблицы SH.
- Вход: d - количество координат, k - количество кластеров,
-- m - степень нечеткости, eps - точность кластеризации.
- Выход: -.
- Результаты кластеризации хранятся в таблицах U и C.

```
CREATE OR REPLACE FUNCTION pgfcm(d INTEGER, k INTEGER,  
    m NUMERIC, eps NUMERIC)  
RETURNS VOID  
LANGUAGE plpgsql
```

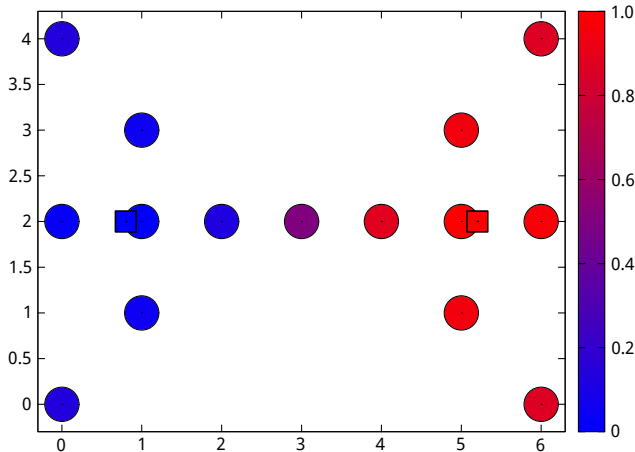
```
C1: INSERT INTO C
      SELECT R.j, SV.l,
             sum(R.s * SV.val) / sum(R.s) AS val
      FROM (SELECT i, j, U.val^m AS s
            FROM U) AS R, SV
      WHERE R.i = SV.i
      GROUP BY j, l;

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;
```

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

Алгоритм pgFCM

Тест на наборе данных "Butterfly"



Параметры алгоритма: $d = 2$, $k = 2$, $m = 2.0$, $eps = 0.01$.

Таблица U

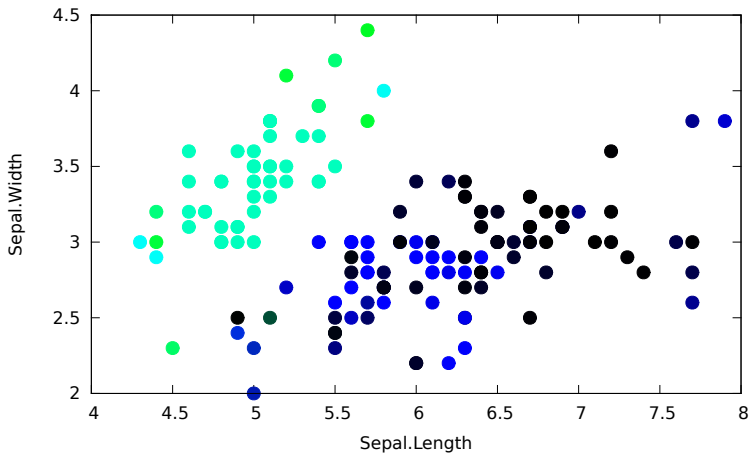
i	j	val
1	1	0.86
1	2	0.13
2	1	0.97
2	2	0.02
...
8	1	0.49
8	2	0.50
...
15	1	0.13
15	2	0.86

Таблица C

j	l	val
1	1	0.79
1	2	2.0
2	1	5.2
2	2	1.99

Алгоритм pgFCM

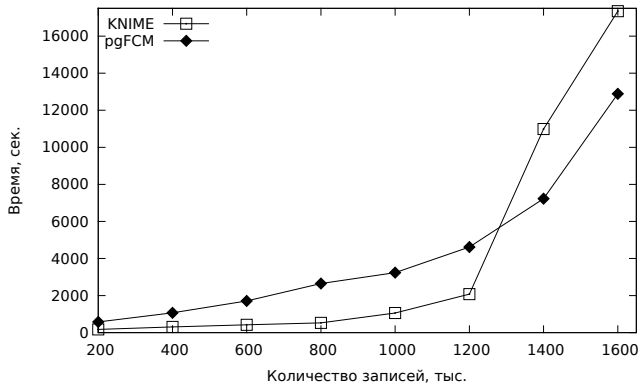
Тест на наборе данных "Iris"



Параметры алгоритма: $d = 5$, $k = 3$, $m = 2.0$, $eps = 0.01$.

Алгоритм pgFCM

Эксперименты по исследованию производительности



Параметры алгоритма: $d = 5$, $k = 3$, $m = 2.0$, $eps = 0.01$, $n = \overline{200\ 000, 1\ 600\ 000}$.

Апробация

Международная конференция SYRCoDIS'2011 "The Seventh Spring Researchers Colloquium on Databases and Information Systems" (June 2-3, 2011, Moscow, Russia).

Публикация

Miniakhmetov R. Integrating Fuzzy c-Means Clustering with PostgreSQL // Proceedings of the Seventh Spring Researchers' Colloquium on Databases and Information Systems (SYRCoDIS'2011). Moscow: Moscow State University, 2011. P. 6-10.

- Выполнено проектирование алгоритма нечеткой кластеризации на языке запросов SQL и схемы соответствующей реляционной базы данных.
- Выполнена реализация разработанного алгоритма для реляционной СУБД с открытым исходным кодом PostgreSQL.
- Выполнено тестирование на стандартных наборах данных Butterfly и Iris.
- Проведены эксперименты для исследования эффективности разработанного алгоритма на различных наборах данных.

Дальнейшие исследования могут быть направлены на улучшение текущей реализации алгоритма pgFCM, а также на разработку параллельной версии алгоритма.

Алгоритм pgFCM

Эксперименты по исследованию производительности

N, тыс.	pgFCM, сек	KNIME, сек	JDBC1, сек	JDBC2, сек
200	578	174	3	25
400	1067	310	9	49
600	1711	423	13	75
800	2648	529	28	100
1000	3238	1061	95	125
1200	4620	2078	123	152
1400	7229	10989	161	178
1600	12888	17347	216	223

Параметры алгоритма: $d = 5$, $k = 3$, $m = 2.0$, $eps = 0.01$, $n = \overline{200\ 000, 1\ 600\ 000}$.