КЛАСТЕРИЗАЦИЯ ОБЪЕКТОВ НА ОСНОВЕ НЕЧЕТКОЙ ЛОГИКИ

Арышева К.С., Аксенов С.В.
Томский политехнический университет
varks@tpu.ru

Введение

Развитие науки в настоящее время, наблюдения и эксперименты, рост количества получаемой информации и необходимость ее обработки требуют создания высокопроизводительных вычислительных систем для кластеризации или классификации огромного неструктурированного множества данных. В данной работе рассматривается задача кластеризации астрономических объектов с использованием методов нечеткой логики.

Нечеткая логика представляет собой обобщение традиционной логики и теории множеств, базирующееся на понятии нечеткого множества. Данное понятие было введено в 1965 году математиком Л.Заде, которое расширяет определение классического множества, допуская значение функции принадлежности множеству в интервале [0;1]. Это означает, что объект может принадлежать множеству с некоторой степенью. Такой тип принадлежности позволяет описывать более естественные задачи кластеризации объектов. Кластеризация – объединение объектов в группы – кластеры – на основе их схожести по качественным, количественным и другим признакам. Для определения кластера используются логические выражения вида: если $x_1 = a$, $x_2 = b$, … $x_n = n$, то $y \in I$, где y – объект кластера I, имеющий $\{x_1;x_n\}$ параметров [1].

Основные положения

Для кластеризованного множества объектов с определенным набором признаков возможно выделение кластеризующих признаков на основе правил вывода нечеткой логики. Рассмотрим пример. Пусть имеется N объектов x с k признаками, k = [1 ..3]. Пусть на данном множестве выделено два кластера. Рассмотрим выделенные кластеры относительно первого признака и второго признака на Рис. 1.

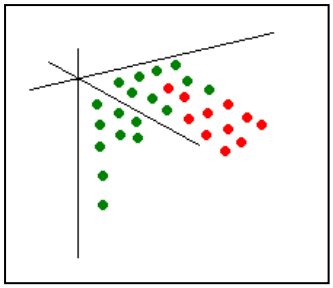

Рис. 1. 2 кластера

Как можно заметить, по первому признаку объекты двух разных кластеров идентичны. Тем не менее, второй признак является определяющим, так как не содержит пересечения кластеров. Таким образом, можно заключить, что второй признак является определяющим для первого и второго кластеров. Теперь если к данному множеству добавить объекты третьего кластера, можно заметить, что второй признак перестал быть определяющим. Для второго и третьего кластера определяющим признаком будет первый параметр (Рис. 2).

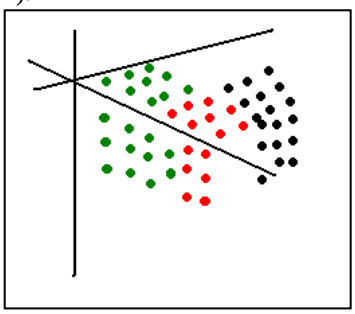

Рис. 2. 3 кластера

Т.е. для множества объектов достаточно двух признаков для кластеризации и вывода правил вида: если xk1 = pk1, xk2 = pk2, то x принадлежит кластеру k. На основе данных правил можно выявить ошибки кластеризации или подтвердить ее правильность.

Реализация системы

Приложение на данном этапе позволяет проводить кластеризацию объектов на основе методов нечеткой логики и горной кластеризации. Результаты данной работы станут основой для вывода правил кластеризации, рассмотренных в теоретической части.

Алгоритм заключается в поиске точек, которые потенциально могут быть центрами кластеров. Потенциальными центрами кластеров должно быть конечное число точек, определенное несколькими способами. Первый способ – принять количество кластеров количеству объектов исследования. Другой способ выбора центров кластеров использует дискретизацию пространства входных признаков. Диапазоны изменения входных данных разбиваются на интервалы, а через точки разбиения проводятся прямые параллельные координатным осям. Произведение количества точек пересечения для всех признаков равно начальному количеству кластеров. Потенциал каждой точки рассчитывается по формуле (1), где Z_h – потенциальные центры кластеров, X_k – признак кластеризуемого объекта, M - количество объектов, α – положительная константа.

$$P(Z_h) = \sum_{k=1}^{M} \exp\left(-\alpha * D(Z_h, X_k)\right),$$
$$D(Z_h, X_k) = \sqrt{\|Z_h - X_k\|^2} \quad (1)$$

В качестве центра первого кластера выбирают точку с наибольшим потенциалом. После для всех остальных потенциальных центров пересчитывается их потенциал с учетом влияния только что

выделенного центра кластера по формуле (2), где P – потенциал первого выделенного кластера, V – центры первого кластера, β – положительная константа.

$$P'(Z_h) = P(Z_h) - P(V) * \exp\left(-\beta * D(Z_h, V)\right) \quad (2)$$

Данная процедура повторяется, пока потенциал следующего центра не окажется меньше какого-либо установленного значения [2][3].

В приложении реализованы два метода кластеризации. В первом способе в качестве потенциальных центов выбираются объекты, обладающие параметрами с максимальными или минимальными значениями.

```
for (int i = 0; i < 4; i++){
initCenters.Add(new Center(spaceObjects.Find(sObject =>sObject.parameters[i].value == spaceObjects.Max(sObj => sObj.parameters[i].value)), 0));
initCenters.Add(new Center(spaceObjects.Find(sObject =>sObject.parameters[i].value == spaceObjects.Min(sObj => sObj.parameters[i].value)), 0));}
```

Второй способ в качестве потенциальных центров отбирает элементы, обладающие наиболее распространенными значениями параметров – объекты, дающие большую плотность.

```
var commonParameterValue = GetDensity();
for (int i = 0; i < commonParameterValue.Count; i++) {
for (int j = 0; j < maxParameterValue[i].Count; j++)
{initCenters.Add(newCenter( spaceObjects.Find(sObject => sObject.parameters[i].value == commonParameterValue[i][j]), 0)); } }
```

Первый способ дает следующий результат кластеризации (Рис. 3).

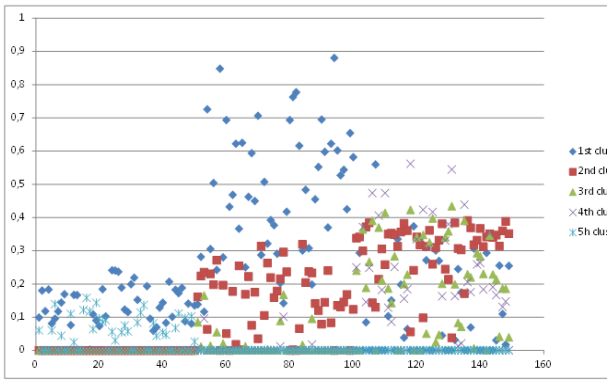

Рис. 3. Результат кластеризации для первого способа

Результаты для второго способа (Рис. 4):

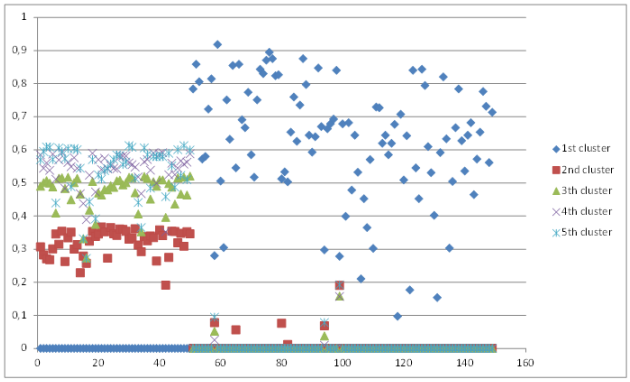

Рис. 4. Результат кластеризации для второго способа

Заключение

Как видно из представленных результатов, предложенные методы кластеризации достаточно эффективны и выделенные кластеры на графиках имеют достаточно определенные границы. Конечно, данные методы еще возможно улучшить, но уже в настоящий момент на основе полученных кластеров можно выделить правила кластеризации следующего вида:

$$Если\ p_1 = x_1, p_2 = x_2, \dots, p_n = x_n тогда\ объект \in Кластеру$$

В ближайшее время на основе полученных правил кластеризации планируется реализация анализа кластеризованного множества для выявления достаточного количества определяющих признаков. Таким образом, полученный кластеризатор можно буте использовать не только для кластеризации множества объектов, но и для проверки уже кластеризованного множества.

Список литературы

1. Joseph M. Barone, Dimitar P. Filev, Ronald R. Yager, “Moimtain method-based fuzzy clustering: methodological considerations”, International Journal of General Systems,vol 23:4, 281-305,(1995).
2. Khaled Hammouda, “A Comparative Study of Daya Clustaring Techiques”, University of Waterloo, Ontario, Canada, Volume 13, Issues 2-3, pp. 149-159,(November 1997).
3. Kuhu Pal, Nikhil R. Pal, James M. Keller, James C. Bezdek, “Relational Mountain (Density) Clustering Method and Web Log Analysis”, International Journal of Intelligent Systems, vol. 20, 375–392 (2005).