УДК 681.513

АЛГОРИТМ УПРАВЛЕНИЯ ПРОЦЕССОМ КЛАСТЕРИЗАЦИИ ПО БЛИЖАЙШЕМУ РАССТОЯНИЮ

П.Н.Коваль

Международный научно-учебный центр информационных технологий и систем НАН и МОН Украины

dep175@irtc.org.ua

Предложен алгоритм управления процессом кластеризации по ближайшему расстоянию с использованием процедуры иерархической группировки.
Ключевые слова: кластеризация, иерархическая группировка.

There is proposed an algorithm of control for clustering process by nearest distance with the use of hierarchical grouping procedure.
Keywords: clustering, hierarchical grouping.

Запропоновано алгоритм керування процесом кластеризації за найближчою відстанню з використанням процедури ієрархічного групування.
Ключові слова: кластеризація, ієрархічне групування.

Вступление

Ранее [1,2] нами было предложено использовать кластеризацию для удаления неинформативных признаков из первичного набора данных при построении математических моделей сложных систем. При значительном количестве неинформативных параметров процесс кластеризации, основанный на процедуре иерархической группировки, завершается выделением большого количества непредставительных кластеров. В настоящей работе предложен метод управления процессом кластеризации для получения представительных кластеров.

Кластеризация по ближайшему расстоянию

Кластеризация осуществляется с помощью процедуры иерархической группировки по ближайшему расстоянию d в многомерном пространстве параметров. При длине выборки N и количестве параметров m ближайшее расстояние d_i для i – ой точки $X_i(x_{i1}, x_{i2}, x_{i3} \ldots x_{im})$ определяется следующим образом:

$$d_i = \min d_{ij}(X_i, X_j) \text{ при } r_i \geq r_j, \quad i \neq j, \quad j \in 1, N \tag{1}$$

где $d_{ij} = \sqrt{\sum_{k=1}^{m}(x_{ik} - x_{jk})^2}$; $r_i = \sqrt{\sum_{k=1}^{m}(\overline{x}_k - x_{ik})^2}$; $\overline{x}_k = \frac{1}{N}\sum_{l=1}^{N} x_{lk}$.

Использование функции плотности распределения для ближайшего расстояния d в виде:

$$p(x,d) = \frac{(m-1)x^{m-1}}{d^m} \exp\left\{ - \frac{\frac{m-1}{m} \cdot x^m}{d^m} \right\}, \qquad (2)$$

где d - наиболее вероятное значение ближайшего расстояния, позволило свести задачу кластеризации к задаче дихотомии, а именно, к отделению внутри кластерных ближайших расстояний от меж кластерных. Для вычисления порога θ может быть использовано одно из известных решающих правил: байесовское, Неймана-Пирсона, по МПУ (метод предельных упрощений [3]) или мини-максное.

С помощью функции плотности *p(x,d)* для каждого решающего правила может быть вычислена вероятность ошибки R , которая принята нами за оценку качества кластеризации, а именно, качество кластеризации тем выше, чем меньше значение вероятности ошибки. При использовании байесовского решающего правила для вероятности ошибки получено следующее выражение:

$$R = 0.5 * \left\{ 1 - \exp\left(- \frac{m \ln\left(\frac{D}{d}\right)}{\frac{D^m}{d^m} - 1} \right) + \exp\left(- \frac{m \ln\left(\frac{D}{d}\right)}{1 - \frac{d^m}{D^m}} \right) \right\}. \qquad (3)$$

где d - наиболее вероятное значение внутри кластерного ближайшего расстояния, усреднённое по кластерам;

D - наиболее вероятное значение меж кластерного ближайшего расстояния.

Как следует из (3), с ростом отношения D/d вероятность ошибки отделения внутрикластерных ближайших расстояний от межкластерных уменьшается. Кроме того, на ход функции R влияет мерность пространства параметров m . С увеличением мерности пространства значение R падает с ростом отношения D/d более резко.

Возможность использования процедуры кластеризации для устранения неинформативных параметров обусловлена тем, что их исключение из выборки уменьшает значение вероятности ошибки R, тогда как исключение информативного параметра эту вероятность увеличивает. Это позволило организовать итеративный процесс исключения из выборки данных неинформативных параметров.

Однако при наличии значительного количества неинформативных параметров в исходных данных процедура кластеризации оканчивается выделением большого количества слабо наполненных кластеров, либо объединением всех точек в один кластер.

Управление ходом процесса кластеризации

Термин кластеризация может быть расшифрован как автоматическая классификация при незаданном заранее числе классов. В одном из первых алгоритмов кластеризации «Форель» [4] вместо задания числа классов вводился максимальный радиус класса. При наличии «компактных» классов, т.е. ограниченных областей простой формы, равномерно заполненных точками, результаты классификации сохранялись при изменении заданного максимального радиуса в определенном интервале. Деформация классов (кластеров) сужала этот интервал.

Для управления процессом кластеризации, базирующемся на использовании процедуры иерархической группировки, нами предложено организовать кластеризацию в виде итеративного процесса, использующего решающее правило Неймана-Пирсона либо МПУ. В первом случае задаётся граничное значение ошибки второго рода:

$$\int_0^{\theta} p(x, D)dx \le \beta. \tag{4}$$

Это приводит к следующему выражению для решающего правила:

$$\theta = D \cdot \sqrt[m]{\frac{m}{m-1} \cdot \left(-\ln(1-\beta)\right)}, \tag{5}$$

где D – наиболее вероятное межкластерное расстояние. При использовании метода предельных упрощений [3] задаётся граничное значение вероятности ошибки первого рода α. Порог θ определяется при этом из условия:

$$\int_{\theta}^{\infty} p(x, d)dx \le \alpha. \tag{6}$$

Из этого условия получаем следующее выражение для порога:

$$\theta = d \cdot \sqrt[m]{-\frac{m}{m-1}\ln\alpha}, \tag{7}$$

где d – наиболее вероятное значение внутри кластерного расстояния. Управление кластеризацией осуществляется путём постепенного уменьшения значения величины ошибки первого рода α или второго рода β от 0.5 до значения, близкого к нулю. При значении α=0.5 число выделенных кластеров равняется числу точек выборки, а при α=0.0001 процесс кластеризации заканчивается объединением всех точек в один кластер.

Кластеризация при вычислении порога по величине ошибки второго рода β носит противоположный характер: при β=*0.5* получаем один кластер. Приемлемый вариант кластеризации находится повторением процедуры кластеризации с пошаговым уменьшением значения α или β.

Алгоритм управления процессом кластеризации

Таким образом, при значительном количестве неинформативных признаков в исходных данных, предложенный в [1] алгоритм следует дополнить итеративным процессом выбора окончательного варианта кластеризации. Опишем алгоритм управления кластеризацией на примере задания граничного значения величины ошибки первого рода α:

1. Принимаем $\alpha=0.25$ и проводим кластеризацию.

2. Если число кластеров близко к числу точек, то уменьшаем значение α вдвое и снова проводим кластеризацию.

3. При объединении всех точек в один кластер, увеличиваем значение α на половину величины верхнего интервала и снова проводим кластеризацию.

4. При выделении преемлемого числа кластеров l , т.е. при $l << N$, значение α считаем окончательным. По этому значению оцениваем качество кластеризации.

Решение тестового примера

Для подтверждения работоспособности описанного алгоритма было проведено его испытание на тестовом примере. Была смоделирована смешанная выборка из 20 точек по 6 параметров каждая. В плоскости первых двух параметров все данные группировались в 2 разнесённых кластера. и две отдельно отстоящие точки. В плоскости 3-го и 4-го параметров точки образовали 4 других по составу кластера. В плоскости 5-го и 6-го параметров все точки были равномерно рассеяны по всей плоскости, т.е. эти параметры неинформативные. При переходе в пространство 5-ти измерений исключение неинформативного параметра дало следующие значения вероятности ошибки: α=*0.0078125*. При исключении второго (информативного) параметра получено следующее значение вероятности ошибки: α=*0.046875*. Как видим, даже при таком соотношении информативны и неинформативных параметров может быть проведена кластеризация и получена оценка её качества, т.е. исключение информативного параметра позволяет провести кластеризацию при более высоком значении

вероятности ошибки первого рода, чем исключение неинформативного параметра. Исключение неинформативного параметра значительно снижает допустимую величину ошибки первого рода.

Выводы

Управление процессом кластеризации по ближайшему расстоянию с помощью процедуры иерархической группировки позволило выделить кластеры при значительном количестве неинформативных параметров в выборке исходных данных. При этом качество кластеризации оценивается непосредственно по значению вероятности ошибки (первого или второго рода) по отделению внутри кластерных расстояний от меж кластерных.

Литература

1. Коваль П.Н. Использование кластеризации при анализе данных // УСиМ. - 2010. - №6. - С. 32 – 34.

2. Коваль П.Н. Использование кластеризации при выборе структуры объекта управления // Індуктивне моделювання складних систем. Єбірник наукових праць / Відп. Редактор В.С.Степашко – Київ: Міжнар. наук.- навч. центр інформ. технологій та систем НАН та МОН України, 2010.- Вип. 2. - 280 с.

3. Васильев В.И., Шевченко А.И., Эш С.Н. Принцип редукции в задачах обнаружения закономерностей // Донецьк: ІПШІ «Наука і освіта», 2009. – 340 с.

4. Загоруйко Н.Г. Методы распознавания и их применение. – М: Сов. Радио, 1972. – 208с.