УДК 681.513

ИСПОЛЬЗОВАНИЕ КЛАСТЕРИЗАЦИИ ПРИ ВЫБОРЕ СТРУКТУРЫ ОБЪЕКТА УПРАВЛЕНИЯ

П.Н.Коваль

Международный научно-учебный центр информационных технологий и систем НАН и МОН Украины
dep175@irtc.org.ua

Запропоновано використати кластеризацію за величиною найближчої відстані в багатовимірному просторі параметрів на етапі попередньої обробки експериментальних даних. На основі виразу для ймовірності похибки побудовано алгоритм для виключення з повного набору параметрів неінформативних ознак.
Ключеві слова: кластеризація, ієрархічне групування, якість кластеризації.

The stage of preliminary analysis of dates is considered. There is proposed to use the clustering over short distance in multi measure space of the parameters. The probability formula of error is shown. The method of strike off non information signs from parameters set is created.
Keywords: clustering, hierarchical grouping, quality of clustering

Предложено использовать кластеризацию по величине ближайшего расстояния в многомерном пространстве параметров на этапе предварительной обработки экспериментальных данных. На основе выражения для вероятности ошибки построен алгоритм исключения из полного набора параметров неинформативных признаков.
Ключевые слова: кластеризация, иерархическая группировка, качество кластеризации.

Вступление

В работе предлагается использовать кластеризацию для отбора так называемых неинформативных признаков, то-есть тех параметров, которые могут быть исключены из математической модели объекта управления.

При построении математических моделей объектов управления в форме уравнений регрессии в качестве входных переменных зачастую используются все доступные для измерения параметры объекта. Однако увеличение количества входных переменных приводит к тому, что знаки при некоторых переменных в построенной модели могут не соответствовать известному из опыта характеру влияния этих параметров на объект управления. Поэтому весьма важным является вопрос отбора параметров на роль входных переменных.

Если выборка содержит результаты измерения параметров при работе технического объекта в некоторых фиксированных режимах, то с помощью кластеризации эти режимы могут быть выделены как отдельные кластеры в многомерном пространстве параметров. Присутствие в данных неинформативных параметров будет отражаться на «компактности» кластеров, т.е. будут размываться границы между кластерами.

Авторы [1] предлагают использовать кластеризацию для исключения неинформативных параметров как таких, что ухудшают качество выделенных кластеров. Качество кластеризации может быть оценено с помощью различных внешних критериев, в частности, по величине вероятности ошибки при отделении внутрикластерных расстояний от межкластерных [2].

Кластеризация по ближайшему расстоянию

Пусть исходные данные о работе технологического объекта представлены в виде смешанной выборки, составленной из N измерений, каждое измерение представлено значениями m параметров. Это позволяет представить выборку как множество точек в многомерном пространстве параметров. Наличие каких-либо закономерностей в данных будет отражено в характере расположения точек в этом пространстве. Статические закономерности или связи приведут к наличию локальных областей с повышенной плотностью точек, т.е. кластеров.

Будем рассматривать случай, когда имеет место следующие статистические модели порождения кластеров: кластеры представляют собой ограниченные области, равномерно заполненные точками или кластеры представляют собой центры рассеяния. Такие кластеры могут быть охарактеризованы координатами своих центров. Для центров кластеров примем следующую статистическую модель порождения – центры кластеров равномерно рассеяны по всему пространству параметров. Предполагается также, что плотность точек в пространстве параметров невысока.

Используя евклидово расстояние между точками в многомерном пространстве параметров для случая равномерного распределения точек в кластере (первая модель порождения кластеров) нами была получена функция плотности по скалярной величине – ближайшему расстоянию[2]:

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

где d – наиболее вероятное значение ближайшего расстояния,

m – мерность пространства параметров.

Ближайшее расстояние для данной точки выборки определяется как наименьшее расстояние до точки, менее удаленной от центра выборки, чем заданная. Эта плотность распределения может быть применима как для точек в кластере, так и для центров кластеров с параметром D - наиболее вероятным значением ближайшего межкластерного расстояния.

Для случая статистической модели кластеров как центров рассеяния (вторая модель) в качестве функции плотности по расстояниию r к центру взята функция плотности многомерного хи-квадрат распределения:

$$p(r,\Sigma) = \frac{|r|^{m-1}}{2^{\frac{m}{2}-1} \cdot \Gamma\left(\frac{m}{2}\right) \cdot |\Sigma|^{\frac{1}{2}}} \cdot e^{-\frac{r^t \Sigma^{-1} r}{2}}, \qquad (2)$$

где r – радиус-вектор произвольной точки кластера в системе координат, помещённой в центр кластера,

Σ - ковариационная матрица кластера,

$\Gamma(t)$ – Гамма-функция $\Gamma(x) = \int_0^\infty t^{x-1} \cdot e^{-t} \cdot dt$,

m - мерность пространства параметров.

Для проведения самой кластеризации используется процедура иерархической группировки [2], когда на каждом шаге последовательно в один кластер объединяются две точки или промежуточные кластеры с минимальным расстоянием и при условии, что это расстояние не превышает порогового значения. При этом весь массив ближайших расстояний, полученный по смешанной выборке исходных данных, с помощью порога разделяется на внутрикластерные и межкластерные расстояния.

Наличие $p(x,d)$, $p(r,\Sigma)$ и $p(x,D)$ позволяет получать разные решающие правила и вычислять значение порога. Из условия минимума среднего риска может быть получено байесовское решающее правило. В случае первой модели пороговое значение θ для отделения внутрикластерных ближайших расстояний от межкластерных вычисляется по формуле:

$$\theta = d \sqrt[m]{\frac{\frac{m^2}{m-1} \ln\left(\frac{D}{d}\right)}{1 - \frac{d^m}{D^m}}}. \qquad (3)$$

Для второй модели порождения кластеров и при упрощающих предположениях о единичных ковариационных матрицах и нормировке данных на дисперсию получим следующее выражение для порогового значения радиуса кластера ρ (усреднённого по составляющим многомерного пространства):

$$\rho = \frac{\sigma}{\sqrt{1-\sigma^2}} \cdot \sqrt{-2m \cdot \ln \sigma}, \qquad (4)$$

где σ - усреднннное по составляющим среднеквадратичное отклонение точек в кластере,

m - мерность пространства параметров.

Также можно построить решающее правило, основанное на задании граничного значения вероятности ошибки первого рода α, т.е. определить порог θ из условия:

$$\int_{\vartheta}^{\infty} p(x,d)dx \le \alpha . \tag{5}$$

Для первой модели получаем следующее:

$$\theta = d \cdot \sqrt[m]{-\frac{m}{m-1} \cdot \ln \alpha} . \tag{6}$$

При небольшом числе неинформативных параметров выделение кластеров процедурой иерархической группировки может быть проведено на основе байесовского решающего правила, то-есть с вычислением порога θ по (3). В противном случае следует использовать значение порога θ, вычисленного по формуле (6). Изменяя значение допустимой ошибки первого рода α от 0.01 до 0.5 можно выделить представительные кластеры даже при большом количестве неинформативных признаков.

Оценка качества кластеризации

После выделения всех кластеров с помощью процедуры иерархической группировки (для первой модели порождения кластеров) и с использованием байесовского решающего правила может быть вычислен средний риск или вероятность ошибки:

$$R = 0.5 * \left\{ 1 - \exp\left(-\frac{m \ln\left(\frac{D}{d}\right)}{\frac{D^m}{d^m} - 1} \right) + \exp\left(-\frac{m \ln\left(\frac{D}{d}\right)}{1 - \frac{d^m}{D^m}} \right) \right\}. \tag{7}$$

Как следует из (7), 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.

Эта вероятность ошибки *R* принята за критерий качества кластеризации: чем меньше *R*, тем выше качество. При использовании решающего правила по формуле (5) качество кластеризации определяется значением величины α, при котором были получены представительные кластеры.

Алгоритм исключения неинформативных признаков

Полученное значение α и вычисленное по формуле (7) значение R позволяют решить задачу исключения из выборки данных неинформативных параметров. При вычислении евклидового расстояния между точками в многомерном пространстве, каждый параметр вносит свой вклад в значение расстояния. Если допустить, что неинформативный параметр вносит в среднем одинаковый вклад как во внутрикластерное, так и в межкластерное ближайшее расстояние, то при условии, что $D/d>1$, наличие этого признака приводит к уменьшению отношения D/d. Следовательно, исключение такого параметра увеличивает отношение D/d и уменьшает вероятность ошибки R. Информативный параметр дает разный вклад в эти расстояния, а именно: больший в межкластерные и меньший во внутрикластерные. Поэтому исключение информативного параметра приведёт к меньшему, чем в предыдущем случае изменению отношения D/d, т.е. вероятность ошибки будет больше.

Следовательно, на роль неинформативного признака будет претендовать параметр, исключение которого приводит к наибольшему уменьшению вероятности ошибки R. Это позволяет построить следующую процедуру отбора неинформативных параметров на основе использования кластеризации при незначительном количестве неинформативных параметров:

1. Исключая поочерёдно каждый параметр, т.е. сократив мерность пространства параметров на единицу, проводим кластеризацию с помощью процедуры иерархической группировки.

2. Вычисляем вероятность ошибки R по формуле (7) для каждого случая.

3. Параметр, исключение которого даёт наименьшее значение вероятности ошибки R , исключаем как неинформативный.

4. Процесс исключения параметров повторяем до тех пор, пока не получим возрастание вероятности ошибки R или уменьшение R станет незначительным. Тогда весь оставшийся набор параметров считаем информативными признаками.

При значительном количестве неинформативных признаков следует использовать решающее правило в соответствии с формулой (5) с вычислением порога по формуле (6). В этом случае после исключения одного параметра процедуру кластеризации проводим многократно, изменяя значение α с шагом 0.05 в диапазоне от 0.05 до 0.5, пока не будут выделены представительные кластеры, т.е. кластеры, содержащие 25 – 50% точек всей выборки. В качестве критерия качества в этом случае выступает сама величина α.

Для подтверждения работоспособности описанного алгоритма было проведено его испытание на тестовом примере. Была смоделирована смешанная выборка из 20 точек по 4 параметра каждая. В плоскости первых двух

параметров все данные группировались в 2 разнесённых кластера и две отдельно отстоящие точки. В плоскости следующих двух параметров все точки были равномерно рассеяны по всей плоскости, т.е. эти параметры неинформативные. После проведения кластеризации в четырёхмерном пространстве параметров значение вероятности ошибки R было равным 0.1. При переходе в пространство трёх измерений исключение каждого неинформативного параметра давало следующие значения вероятности ошибки: R=0.04 и R=0.06. При исключении информативных параметров получены следующие значения вероятности ошибки: R=0.16 и R=0.2. Как видим, исключение информативных параметров ухудшает качество кластеризации, тогда как исключение неинформативных параметров качество кластеризации улучшает. Исключение параметра, давшего наименьшее значение R, и переход в пространство двух измерений дало следующие значения вероятности ошибки: неинформативный параметр – R=0.06, информативные - R=0.7 и R=0.8. Исключив неинформативный параметр, окончательно получим двумерное пространство из информативных параметров со значением вероятности ошибки R=0.06.

Выводы

Таким образом, если предположительно в области изменения входных переменных объекта могут существовать локальные области повышенной плотности, то с использованием критерия качества кластеризации в форме вероятности ошибки согласно формулам (5) или (7), можно исключить из дальнейшего анализа те параметры, которые ухудшают качество кластеризации. Так как вероятность ошибки R зависит от отношения D/d экспотенциально, то её использование в качестве критерия предпочтительнее простой оценки качества по отношению D/d. Значение R является достаточно надёжным индикатором для процедуры исключения неинформативных параметров.

Литература

1. Васильев В.И., Шевченко А.И. Искусственный интеллект: формирование и опознавание образов. Издание второе, дополненное и переработанное. – Донецк. ДонГИИИ, 2000. - 360 с.

2. Коваль П.Н. Кластеризация на основе скалярной оценки ближайшего расстояния // Міжнародний семінар з індуктивного моделювання МСИМ-2005. Збірник праць / Відповідальний редактор д.т.н. Степашко В.С. – Київ: Міжнародний науково-навчальний центр інформаційних технологій та систем НАН та МОН України, 2005. – 370 с.