УДК 681.513

П.Н. Коваль

Использование кластеризации при анализе данных

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.

It is suggested to use the clastering over the short distance in a multi-measure space of the parameters at the stage of the preliminary processing of experimental data. On the basis of the expression for the probability of an error an algorithm of the elimination of non-informing signs from a complete set of parameters is constructed.

Запропоновано використання кластеризації 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.

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

Пусть исходные данные представлены в виде смешанной выборки из N объектов, каждый из которых описан m параметрами. Это позволяет представить выборку как множество точек в многомерном пространстве параметров. Наличие каких-либо закономерностей в данных отражено в характере расположения точек в этом пространстве. Статические закономерности или связи приведут к наличию локальных областей с повышенной плотностью точек, т.е. кластеров. Присутствие в данных неинформативных параметров отражается на «компактности» кластеров [1], т.е. будут размываться границы между кластерами. Для оценки этого влияния в [2] предложен критерий качества кластеризации в виде выражения для вероятности ошибки. Это позволяет использовать кластеризацию для исключения неинформативных параметров как таких, которые ухудшают качество кластеризации.

Статистическая модель порождения кластеров

Рассмотрим случай, когда имеет место следующая статистическая модель порождения кластеров, а именно, кластеры представляют собой ограниченные области, равномерно заполненные точками. Такие кластеры могут быть охарактеризованы координатами своих центров. Для центров кластеров примем такую же 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 при указанных предположениях, получена функция плотности по скалярной величине – ближайшему расстоянию [2]:

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

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.

Кластеризация на основе ближайшего расстояния

Наличие $p(x,d)$ и $p(x,D)$ позволяет разделить весь смешанный массив ближайших расстояний, полученный по смешанной выборке исходных данных, на внутрикластерные и межкластерные расстояния. При этом могут быть выбраны разные решающие правила. Используя байесовское решающее правило, получим следующее пороговое значение θ для отделения внутрикластерных ближайших расстояний от межкластерных:

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

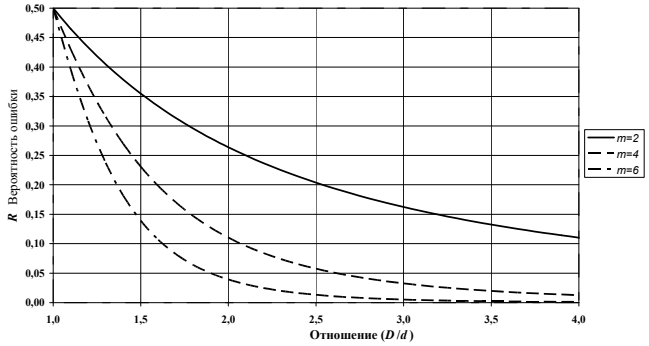
Для кластеризации использовалась процедура иерархической группировки, когда на каждом шаге последовательно в один кластер объединяются две точки (или два промежуточных кластера) с минимальным расстоянием и при условии, что это расстояние не превышает порогового значения θ.

Оценка качества кластеризации

После выделения всех кластеров, например с помощью процедуры иерархической группировки [2], может быть вычислен средний риск или вероятность ошибки:

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

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. На рисунке показан ход R при росте отношения D/d для разной мерности пространства параметров m. Верхняя кривая 1 соответствует мерности пространства $m = 2$, средняя $2 – m = 4$ и нижняя $3 – m = 6$. Как видно из рисунка, с увеличением мерности пространства значение R падает с ростом отношения D/d более резко.



Рисунок

Эту вероятность ошибки R согласно формуле (3) предложено принять за критерий качества кластеризации: чем меньше вероятность ошибки R , тем выше качество.

Алгоритм исключения неинформативных признаков

При вычислении евклидова 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 приведет к меньшему, чем в предыдущем случае, изменению отношения D/d, т.е. вероятность ошибки будет больше.

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

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

• Вычисляем вероятность ошибки R по формуле (3) для каждого случая.

• Параметр, исключение которого дает наименьшее значение вероятности ошибки R, исключаем как неинформативный.

• 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.

Описание тестового примера

Для подтверждения работоспособности описанного алгоритма проведено его испытание на тестовом примере, а также смоделирована смешанная выборка из 20 точек по четыре параметра каждая. В плоскости первых двух параметров все данные группировались в два разнесенных 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 четырехмерном пространстве параметров значение вероятности ошибки R было равным 0,1. При переходе в пространство трех 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. После исключения параметра, давшего наименьшее значение R, и перехода в пространство двух измерений получены 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 $R = 0{,}06$.

Заключение. Таким образом, если предположительно при изменении входных переменных объекта могут существовать локальные области повышенной плотности, то с использованием критерия качества кластеризации в форме вероятности ошибки R согласно формуле (3) можно исключить из дальнейшего анализа те параметры, которые ухудшают качество кластеризации. Так как вероятность ошибки R зависит от отношения D/d экспотенциально, то ее использование в качестве критерия предпочтительнее простой оценки качества по отношению D/d. Значение R является достаточно надежным индикатором для процедуры исключения неинформативных параметров.

1. *Васильев В.И., Шевченко А.И.* Искусственный интеллект: формирование и опознавание образов. – Донецк: ДонГИИИ, 2000. – 360 с.
2. *Коваль П.Н.* Кластеризация на основе скалярной оценки ближайшего расстояния // Зб. пр. Міжнар. сем. з індуктивного моделювання. – Київ: Міжнар. наук.-навч. центр інформ. технол. та систем НАН та МОН України, 2005. – 370 с.

Поступила 12.04.2010

Тел. для справок: (044) 5-2-6337 (Киев)
E-mail: *dep*175@*irtc.org.ua*