Український журнал телемедицини
та медичної телематики

Том 7, №2, 2009

УДК 61:31+004.93'14+004.852

Использование contrast-статистики при кластеризации результатов методом построения самоорганизующихся карт

Ю.Е.Лях, В.Г.Гурьянов, О.Г.Горшков, Ю.Г.Выхованец

Донецкий национальный медицинский университет им. М.Горького, Донецк, Украина

РЕЗЮМЕ, ABSTRACT
Предложен метод оценки оптимального числа кластеров при проведении кластеризации путем построения SOM (Contrast-статистика). Проведен сравнительный анализ эффективности применения метода с уже существующими процедурами: Gap-статистика, Silhouette-статистика. Обоснована возможность применения метода для анализа больших баз медико-биологических данных (Укр.ж.телемед.мед.телемат.-2009.-Т.7,№2.-С.149-153).
Ключевые слова: нейронные сети, кластерный анализ, оптимальное число кластеров

Ю.Є. Лях, В.Г.Гур'янов, О.Г.Горшков, Ю.Г.Вихованець
ВИКОРИСТАННЯ CONTRAST-СТАТИСТИКИ ПРИ КЛАСТЕРИЗАЦІЇ РЕЗУЛЬТАТІВ МЕТОДОМ ПОБУДОВИ КАРТ, ЩО САМООРГАНІЗУЮЮТЬСЯ
Донецький національний медичний університет ім. М. Горького, Донецьк, Україна
Запропоновано метод оцінки оптимального числа кластерів при проведенні попереднього аналізу даних (Contrast-статистика). Метод може бути ефективно використаний при проведенні кластеризації шляхом побудови SOM. Проведено порівняння ефективності методу з раніше розробленими процедурами. Обґрунтовано можливість ефективного використання методу для аналізу великих баз даних у практичних задачах медицини (Укр.ж.телемед.мед.телемат.-2009.-Т.7,№2.-С.149-153).
Ключові слова: нейронні мережі, кластерний аналіз, оптимальне число кластерів

Yu.E.Liakh, V.G.Gurianov, O.G.Gorshkov, Yu.G.Vykhovanets
CONTRAST STATISTIC METHOD FOR SELF-ORGANISING MAP CLUSTERING ALGORITHM
M. Gorky Donetsk National Medical University, Donetsk, Ukraine
A new method (the Contrast statistic) for estimating the number of clusters in a set of data is proposed. The technique uses the output of self-organising map clustering algorithm, comparing the change in dependency of Contrast value upon clusters number to that expected under a uniform distribution. A simulation study shows that the Contrast statistic can be used successfully both, when variables describing the object in a multi-dimensional space are independent (ideal objects) and dependent (real biological objects) (Ukr. z. telemed. med. telemat.-2009.-Vol.7,№2.-P.149-153).
Keywords: neural networks, cluster analysis, optimal number of clusters

Для представления многомерных результатов медико-биологических исследований привлекаются методы кластерного анализа. К задачам кластерного анализа относятся такие, в которых необходимо распределить совокупность некоторых объектов на однородные группы (кластеры) в многомерном пространстве признаков, описывающих эти объекты.

Для решения задач кластеризации медико-биологических данных, которые отличаются, с одной стороны, высоким значением размерности пространства признаков, в котором находятся объекты, с другой стороны, высокой степенью коррелированности этих признаков, с ус-

пехом применяется метод самоорганизующихся карт (SOM – self-organizing map) – нейронные сети Кохонена [1]. Сеть Кохонена обучается без «учителя», воспринимая саму структуру входных данных. Она может быть использована в задачах распознавания образов, разведочном анализе данных. Сеть Кохонена может распознавать кластеры в данных, а также устанавливать близость классов и, таким образом, улучшить понимание структуры данных. Сеть состоит из входного и выходного слоя. Входной слой состоит из элементов (нейронов), каждый из которых соответствует одному признаку, характеризирующему объект. Нейроны этого слоя служат для преобразования входных данных в стандартный вид (обычно эти элементы преобразуют значения от 0 до 1). Выходной слой составлен из радиальных (RBF – radial basis function) элементов (выходной слой называют слоем топологической карты), количество которых равно количеству кластеров, которые будет распознавать сеть. Элементы топологической карты располагаются в некотором (как правило, двумерном) пространстве, что позволяет после проведения кластерного анализа получить наглядное представление о структуре анализируемых данных.

В то же время одной из основных, и нетривиальных, проблем, возникающих при проведении кластерного анализа, является выбор числа кластеров, в которые производится распределение объектов. При выборе числа кластеров меньше оптимального в одну группу могут быть отнесены существенно различающиеся объекты, при выборе числа кластеров больше оптимального однотипные объекты могут быть разделены в разные кластеры. К настоящему времени разработано множество методов оценки оптимального числа кластеров. Достаточно подробно сравнительный анализ этих методов дан в работах [2, 3]. Однако применение этих алгоритмов к анализу больших, многомерных баз реальных медико-биологических данных указало на некоторые проблемы и сложности их реализации.

Цель исследования

Разработка алгоритма оценки оптимального числа кластеров для анализа многомерных массивов медико-биологических данных больших размеров.

Материал и методы

В работе [4] представлено описание Contrast-статистики, которая предназначена для вычисления оптимального числа кластеров при проведении кластеризации методом построения нейронных сетей Кохонена.

Пусть $X = \{x_1, x_2, \ldots, x_N\}$ – множество N точек в m–мерном пространстве признаков. Используя процедуру кластеризации метода SOM, разобъем это множество в k кластеров. Для оценки качества разбиения предлагается рассчитать показатель контрастности кластеризации (Contrast-статистика):

$$\text{Contrast} = \frac{1}{N}\sum_{i=1}^{N}\frac{R_i}{r_i}, \quad (1)$$

где суммирование производится по всем N точкам, R_i – эвклидово расстояние от i-й точки до центра ближайшего к ней кластера, которому она не принадлежит, r_i – эвклидово расстояние от этой точки до центра кластера, которой она принадлежит. Из определения показателя можно предполагать, что при успешном проведении кластеризации показатель Contrast будет иметь большое значение, в случае же деления однородно распределенных точек показатель Contrast будет иметь малое значение.

Для анализа зависимости показатель Contrast-статистики от числа кластеров разбиения в случае однородного распределения точек были проведены численные эксперименты. В экспериментах было смоделировано равномерное распределение точек ($N=10^4$) в пространст-

вах различных размерностей (m=3, 4, …, 12). Для каждого набора данных была проведена кластеризация методом построения SOM (последовательное деление в k=2, 3, …, 49 кластеров). Из анализа полученных результатов применения Contrast-статистики к случаю однородного распределения можно сделать следующие выводы:

1) при небольшом (для данной размерности пространства m) числе кластеров разбиения показатель контрастности разбиения связан с количеством кластеров (k) соотношением

$$\text{Contrast} = a \cdot \text{k}^{\text{л}} \quad (2),$$

где a и λ – некоторые константы, значение которых зависит от размерности пространства, в котором расположены точки;

2) значение константы λ соотношения (2) уменьшается с увеличением размерности пространства m, но всегда остается большим 0;

3) при большом (для данной размерности пространства m) числе кластеров разбиения показатель контрастности не зависит от количества кластеров (Contrast=Const, значение зависит от m).

На рисунке приведен типичный пример примения Contrast-статистики к случаю анализа модельного неоднородного распределения точек (три кластера в 4-мерном пространстве признаков, признаки независимы).

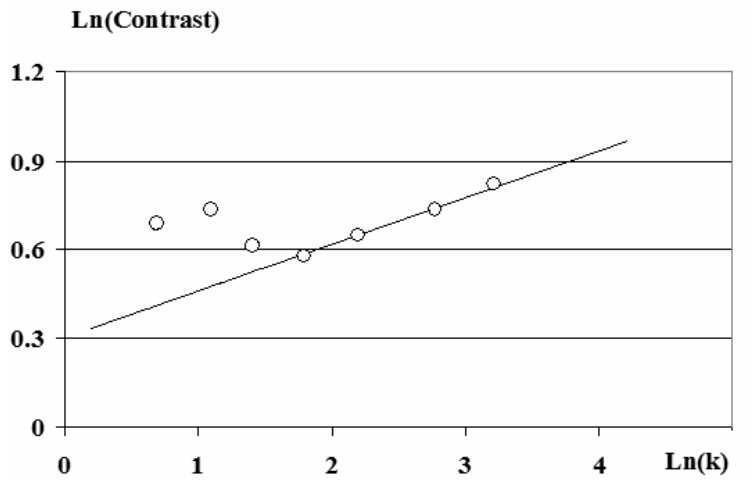

Рисунок. Зависимость показателя Contrast от числа кластеров разбиения для модельных данных (три кластера в 4-мерном пространстве признаков, признаки независимы)

Из рисунка видно, что зависимость значения показателя Contrast-статистики в случае, когда число кластеров значительно превышает количество реально присутствующих групп однородных объектов (k>3), от числа кластеров разбиения описывается формулой (2) – однородное распределение в 4-мерном пространстве признаков. При этом наибольшее отклонение от этой зависимости наблюдается для k=3 (что соответствует количеству реально присутствующих в этом модельном распределении групп однородных объектов).

Отсюда может быть предложен алгоритм вычисления оптимального числа кластеров для некоторого набора многомерных данных:

1) производится кластеризация предложенного множества объектов в 2, 3,..., k кластеров;

2) для каждого разбиения производится расчет Contrast-статистики;

3) анализируется кривая зависимости Contrast(k); в случае, когда значения превышают величину, рассчитанную по (2), можно говорить об эффективном разбиении;

4) за оптимальное число кластеров берется то, для которого зависимость в наибольшей степени отклоняется от теоретической для равномерного распределения.

Результаты и обсуждение

Для оценки эффективности Contrast-статистики был проведен анализ расчета оптимального числа кластеров к различным модельным задачам (число кластеров заранее известно) и реальным медико-биологическим базам данных (анализ записей ЭЭГ [5], анализ реестра больных диабетом [6], анализ стабилограмм [7] и т.д.). Следует отметить, что в модельных задачах исследовались как случаи пространства независимых признаков, так и коррелированных признаков, в реальных медико-биологических данных признаки были коррелированны.

Для проведения сравнения эти же данные были проанализированы с помощью методов Gap-статистики [2] и Silhouette-статистики [3].

Для применения Gap-статистики [2] после проведения разбиения данных в k кластеров рассчитывается показатель g:

$$g(k)=\ln\left(\frac{MSE_{X^*}(k)}{MSE_{X^*}(1)}\right)-\ln\left(\frac{MSE_X(k)}{MSE_X(1)}\right), \quad (3)$$

где $MSE_{X^*}(i)$ – среднее евклидово расстояние от объекта до центра своего кластера при делении референтного (равномерного) распределения объектов в *i* кластеров, $MSE_X(i)$ – среднее евклидово расстояние от объекта до центра своего кластера при делении анализируемого распределения объектов в *i* кластеров (при этом количество объектов в референтном распределении равняется количеству объектов в анализируемом распределении). В [2] предлагается рассчитывать значение $MSE_{X^*}(i)$ для нескольких (B) референтных выборок, по которым производится усреднение и рассчитывается стандартное отклонение этой величины sd(*i*). За оптимальное число кластеров выбирается такое минимальное k, для которого g(k)≥g(k+1) – sd(k+1) [2].

При расчете Silhouette-статистики в случае кластеризации в k кластеров для каждого *j*-го объекта анализируемого распределения рассчитывается величина *s* [3]:

$$s(j)=\frac{b(j)-a(j)}{b(j)}, \quad (4)$$

где *a(j)* – среднее евклидово расстояние от объекта до других объектов, принадлежащих тому же кластеру, *b(j)* – среднее евклидово расстояние от объекта до объектов, принадлежащих ближайшему к объекту кластеру, которому он не принадлежит. Оптимальным считается такое количество кластеров k, для которого среднее (по всем объектам) значение *s(j)* максимально [3].

Обобщая полученные результаты применения трех методов расчета, могут быть сделаны следующие выводы.

1. Применение Gap-статистики является эффективным методом оценки оптимального числа кластеров в случае независимых (или слабо связанных) признаков. Для этих случаев Silhouette-статистика и Contrast-статистика дают сходный по эффективности результат.

2. В случае сильной связи между признаками для эффективного применения Gap-статистики необходима предварительная оценка размерности пространства, в котором генерируется референтное распределение, в противном случае оптимальное число кластеров не может быть определено. В то же время Silhouette-статистика и Contrast-статистика для этих случаев эффективно решают задачу и дают сходный результат.

3. В случае небольших по объему выборок объектов время, затраченное на расчеты Silhouette-статистики и Contrast-статистики, невелико, причем результаты Silhouette-статистики обладают меньшей дисперсией.

4. Для больших по объему массивов данных сложность (а следовательно и время) расчета Silhouette-статистики существенно больше, чем для Contrast-статистики, при этом сложность расчетов Gap-статистики не отличается от сложности для Contrast-статистики (если исключить необходимость оценки референтного распределения).

5. Применение Contrast-статистики позволяет не только дать оценку опти-

мального числа кластеров разбиения, но также оценить число независимых переменных, которые описывают объект анализа.

Выводы

Таким образом, предложена новая процедура оценки оптимального числа кластеров при проведении кластеризации методом построения SOM (Contrast-статистика). Процедура сопоставима по эффективности с уже существующими методами оценки – Gap-статистикой, Silhouette-статистикой – при кластеризации небольших выборок в пространстве независимых признаков. При анализе больших выборок предлагаемая процедура превосходит Silhouette-статистику по скорости расчета. При анализе объектов, распределенных в пространстве коррелированных признаков, предлагаемая процедура превосходит Gap-статистику по эффективности расчета. Это позволяет рекомендовать предложенную процедуру для анализа больших баз реальных медико-биологических данных.

Литература и веблиография

1. *Kohonen T.* Self-Organizing Maps of Massive Databases. Engineering Intelligent Systems, 2001. Vol. 4, pp. 179-185.

2. *Tibshirani R., Walther G., Hastie . T.* (2000) Estimating the Number of Cluster in a Dataset via the Gap Statistic. Technical report, Department of Biostatistics, Stanford University .

3. *Dudoit S., Fridlyand J.* A prediction – based resampling method for estimating the number of clusters in a dataset// Genome Biology.– 2002.– Vol. 3, № 7.

4. *Лях Ю. Е., Гурьянов В. Г.* Обоснование выбора оптимального числа кластеров для метода самоорганизующихся карт Кохонена // Клиническая информатика и телемедицина. – 2005. – Т2. – №1. – С.124.

5. *Островая Т.В., Черний В.И., Гурьянов В.Г, Андронова И.А.* Изучение реактивных перестроек ЭЭГ в ответ на фотостимуляцию с помощью спектрального анализа у пациентов укладывающихся в понятие „норма” // IV з’їзд УБФТ. – Донецк, 2006. – С. 208-209.

6. *Халангот М.Д., Гур’янов В.Г., Кравченко В.І., Тронько М.Д.* Визначення у популяціях хворих на цукровий діабет груп ризику розвитку тяжків наслідків хвороби за допомогою побудови „нейронних мереж кохонена” // Журнал АМН України. – 2007. – Т.13, №1. – С. 133-140.

7. *Лях Ю.Е., Выхованец Ю.Г., Гурьянов В.Г., Прокопец В.И., Черняк А.Н., Грецкая И.Р.* Нейросетевая классификация параметров стабилограммы// Матеріали Всеукраїнської науково-практичної відеоконференції: Актуальні питання дистанційної освіти та медицини 2008. - Запоріжжя-Київ, 2008. - С. 8-10.

Надійшла до редакції: 12.02.2009.

Кореспонденція: Лях Ю.Є.,
Пр-т Ілліча, 16, 83003, Донецьк, Україна
E-mail: info@dsmu.edu.ua