УДК 681.518:004.93.1

ВИЗУАЛИЗАЦИЯ СТРУКТУР БАЗ ДАННЫХ НА ОСНОВЕ НЕПАРАМЕТРИЧЕСКОГО ОЦЕНИВАНИЯ

А.П. Чекалов, *канд.техн.наук;* ***М.С. Бабий,*** *канд.техн.наук;*
С.П. Шаповалов, *канд. физ.-мат. наук*
Сумский государственный университет, г. Сумы

Предложена схема представления многомерных статистических данных и отношений между данными в реляционных базах с помощью вложенных поверхностей, формирующих кластеры. Оптимизация сглаживающего параметра для поверхностей выполняется на основе непараметрического оценивания методом перекрестной проверки с удалением элементов по одному. Описан алгоритм построения вложенных поверхностей.

Запропоновано схему представлення багатовимірних статистичних даних і відношень між даними у реляційних базах за допомогою вкладених поверхонь, що формують кластери. Оптимізація згладжуючого параметра для поверхонь виконується на основі непараметричного оцінювання методом перехресної перевірки з видаленням елементів по одному. Описано алгоритм побудови вкладених поверхонь.

ВВЕДЕНИЕ И ПОСТАНОВКА ЗАДАЧИ

В современных условиях объемы информации, на основе которых необходимо принимать решения, становятся очень большими. Причиной этого являются повсеместная компьютеризация и автоматизация технологических процессов. В качестве примера можно привести серии автоматических измерений в физике, химии, медицине, данные мониторинговых и спутниковых систем.

В общем случае процесс анализа многомерного массива данных $D=\{d_1,...,d_n\}$ состоит в нахождении подмножества $D' \subseteq D$ и проверке гипотез относительно этого подмножества в контексте решаемой задачи. В числе этих гипотез могут быть выполнение свойств, которые поддерживаются для всех или для большинства элементов $e_i \in D'$, классификация D' на классы C_i в соответствии со свойствами P_i , функциональные зависимости $d_{i1}=F(d_{i2},...,d_{ik})$ или отношения $R\ (d_{i1},...,d_{ik})$ между двумя и более размерностями [1].

Визуальное представление данных достаточно удобно для поиска интересующих нас отношений между данными. Обычно для этой цели используют графики (диаграммы) рассеяния. Графики рассеяния удобны в случае двух измерений. В трехмерном случае при большом количестве данных плотные области скрывают друг друга, при этом сложно обнаружить какую-то закономерность. С другой стороны, трудно исследовать и разреженные области, так как точки в этих областях легко принять за шум.

В связи с изложенным поставлена задача построить в пространстве данных поверхности, объединяющие точки с одинаковым уровнем плотности вероятности. Каждому уровню будет соответствовать своя поверхность, в итоге будет получен набор вложенных поверхностей, формирующих кластеры с плотностью выше определенного уровня. Визуализация этих поверхностей позволит легче обнаруживать структуры данных, в том числе и множественные.

ОЦЕНКА ПЛОТНОСТИ ВЕРОЯТНОСТИ

Обозначим функцию плотности вероятности для трехмерной, случайной величины через $f\ (x_1.\ x_2,\ x_3)$. В общем случае x_k могут

представлять разнородные величины. Так как мы не имеем никакой априорной информации относительно вида распределения, наиболее целесообразно использовать непараметрические ядерные оценки типа Парзена-Розенблатта. Для одномерного случая оценка представляется в виде

$$\hat{f}_h(x) = \frac{1}{nh}\sum_{i=1}^{n} K(\frac{x - x_i}{h}) , \qquad (1)$$

где $K(u)$ – функция ядра;

h – ширина окна;

n – длина выборки.

Наиболее часто используются треугольное ядро $K(u) = (1 - |u|)\,[|u| \le 1]$, гауссовское – $K(u) = (2\pi)^{-1/2}\exp(-u^2/2)$, Епанечникова – $K(u) = \frac{3}{4}(1-u^2)\,[|u| \le 1]$ [2].

Для трехмерного случая используем представление ядра в виде произведения по отдельным координатам x_k, тогда оценка плотности

$$\hat{f}_h(x) = \frac{1}{n\,h_1 h_2 h_3}\sum_{i=1}^{n}\prod_{k=1}^{3} K(\frac{x_k - x_{ki}}{h_k}) , \qquad (2)$$

где h_1, h_2, h_3 – длины сторон элементарного параллелепипеда.

Более эффективной является оценка плотности, построенная на основе перекрестной проверки (скользящего контроля) с исключением элементов выборки по одному:

$$\hat{f}_{h\,i}(x) = \frac{1}{(n-1)\,h_1 h_2 h_3}\sum_{j \ne i}^{n}\prod_{k=1}^{3} K(\frac{x_k - x_{kj}}{h_k}) . \qquad (3)$$

Выбор размера h_k значительно больше влияет на качество восстановления плотности, чем выбор ядра [3]. При слишком маленьком h_k плотность концентрируется вблизи элементов выборки, а при слишком большом – плотность чрезмерно сглаживается и в пределе вырождается в константу.

Качество восстановления плотности на основе оценки $\hat{f}$ будем определять с помощью функционала среднеквадратической ошибки

$$I(h) = \int \{\ \hat{f}(x) - f(x)\ \}^2 dx ,$$

интегрирование здесь выполняется по всему пространству. В расширенном виде

$$I(h) = \int \hat{f}^2(x)dx - 2\int \hat{f}(x)f(x)\,dx + \int f^2(x)dx . \qquad (4)$$

Интеграл во втором слагаемом представляет собой математическое ожидание функции $\hat{f}(x)$. В случае использования оценки $\hat{f}_{h\,i}$

$$\mathrm{M}\,\hat{f}_{hi}(x_i) = \frac{1}{n}\sum_{i=1}^{n}\hat{f}_{h\,i}(x_i) ,$$

а функционал (4) примет вид

$$I(h) = \frac{1}{n}\sum_{i=1}^{n}\int \hat{f}_{hi}^2(x)dx - \frac{2}{n}\sum_{i=1}^{n}\hat{f}_{hi}(x_i) + \int f^2(x)dx\,. \qquad (5)$$

Может быть показано [4], что

$$\frac{1}{n}\sum_{i=1}^{n}\int \hat{f}_{hi}^2(x)dx = \int \hat{f}_h^2 dx + o(1/n^2 y)\ ,$$

тогда (5) преобразуется к виду

$$I(h) = \int \hat{f}_h^2(x)dx - \frac{2}{n}\sum_{i=1}^{n}\hat{f}_{hi}(x_i) + \int f^2(x)dx\,. \qquad (6)$$

В качестве $K(u)$ возьмем ядро Гаусса, которое является дифференцируемым на всем пространстве, тогда (2) запишется в виде

$$\hat{f}_h(x) = \frac{1}{n\,h_1h_2h_3}\sum_{i=1}^{n}\left\{\frac{1}{(\sqrt{2\pi})^3}\prod_{k=1}^{3}\exp\left[-\frac{(x_k - x_{ki})^2}{2h_k^2}\right]\right\}. \qquad (7)$$

Первое слагаемое выражения (6)

$$I_1(h) = \int \hat{f}_h^2(x)dx = \frac{1}{(n\,h_1h_2h_3)^2}\frac{1}{(2\pi)^3}\int\left[\sum_{i=1}^{n}\prod_{k=1}^{3}\exp\left(-\frac{(x_k - x_{ki})^2}{2h_k^2}\right)\right]^2 dx =$$

$$= \frac{1}{(2\pi)^3(nh_1h_2h_3)^2}\left(\begin{array}{l}\displaystyle\sum_{i=1}^{n}\int\prod_{k=1}^{3}\left\{\exp\left[-\frac{(x_k - x_{ki})^2}{2h_k^2}\right]\right\}^2 dx + \\ \displaystyle\sum_{i=1}^{n}\sum_{j\neq i}^{n}\int\prod_{k=1}^{3}\left\{\exp\left[-\frac{(x_k - x_{ki})^2}{2h_k^2}\right]\exp\left[-\frac{(x_k - x_{kj})^2}{2h_k^2}\right]\right\}dx\end{array}\right)$$

Нетрудно показать, что

$$\begin{aligned}&\exp\left[-\frac{(x_k - x_{ki})^2}{2h_k^2}\right]\exp\left[-\frac{(x_k - x_{kj})^2}{2h_k^2}\right] = \\ &= \exp\left[-\frac{(x_k - (x_{ki} + x_{kj})/2)^2}{h_k^2}\right]\exp\left[-\frac{(x_{ki} - x_{kj})^2}{4h_k^2}\right]\end{aligned} \qquad (8)$$

С учетом значения табличного определенного интеграла $\int_{-\infty}^{\infty}\exp(-a^2x^2)dx = \frac{\sqrt{\pi}}{a}$ и равенства (8) получим

$$I_1(h) = \frac{1}{(2\sqrt{\pi})^3 nh_1h_2h_3} + \frac{1}{(2\sqrt{\pi})^3 n^2h_1h_2h_3}\sum_{i=1}^{n}\sum_{j\neq i}^{n}\exp\left[-\sum_{k=1}^{3}\frac{(x_{ki} - x_{kj})^2}{4h_k^2}\right]. \qquad (9)$$

Второе слагаемое выражения (6)

$$I_2(h) = -\frac{2}{(\sqrt{2\pi})^3 n^2 h_1 h_2 h_3} \sum_{i=1}^{n} \sum_{j \neq i}^{n} \exp\left[-\sum_{k=1}^{3} \frac{(x_{ki} - x_{kj})^2}{2h_k^2} \right],$$

здесь для простоты $(n\text{-}1)$ мы заменили на n.

Третье слагаемое не зависит от h и при поиске минимума может быть отброшено.

Объединяя выражения $I_1(h)$ и $I_2(h)$ окончательно получим

$$I(h_1, h_2, h_3) = \frac{1}{(2\sqrt{\pi})^3 n h_1 h_2 h_3} + \frac{1}{(2\sqrt{\pi})^3 n^2 h_1 h_2 h_3} \times$$
$$\times \sum_{i=1}^{n} \sum_{j \neq i}^{n} \left\{ exp \left[-\sum_{k=1}^{3} \frac{(x_{ki} - x_{kj})^2}{4h_k^2} \right] - 2^{5/2} exp \left[-\sum_{k=1}^{3} \frac{(x_{ki} - x_{kj})^2}{2h_k^2} \right] \right\}. \qquad (10)$$

ВИЗУАЛИЗАЦИЯ ДАННЫХ

Так как вдали от точки минимума градиент функции (10) может быть близок к нулю, вместо обычных градиентных методов поиска более целесообразно использовать метод Монте-Карло. Если координаты x_k представляют однотипные величины, то в большинстве случаев размер h_k может быть взят одинаковым для всех x_k . В этом случае трехмерная задача нахождения минимума сводится к одномерной.

Полученную оценку плотности для простоты далее по тексту будем обозначать через $f(x, y, z)$.

Для формирования и визуализации кластеров будем использовать переменный порог α, $0 < \alpha < 1$. Точки пространства, в которых $f(x, y, z) \geq 0$, будем считать принадлежащими кластерам данных, соответственно точки, в которых $f(x, y, z) \geq 0$, будут представлять границы кластеров.

Алгоритм визуализации данных может быть представлен в виде последовательности следующих шагов.

Шаг 1 Ввод наблюдений $x[j], y[j], z[j], 1 \leq j \leq n$ из базы.

Шаг 2 Ввод k-, l-, m-количества узлов сетки по каждому измерению.

Шаг 3 Нахождение значений h_x , h_y , h_z , при которых достигается минимум функции I (h_x , h_y , h_z) .

Шаг 4 Вычисление оценки плотности $f(x, y, z)$ в узлах сетки.

Шаг 5 Выбор формы графического представления кластеров. Если выбран вывод сечений по плоскостям XY, YZ или XZ, выполняется шаг 6. Если выбран вывод пространственного изображения кластеров, выполняется шаг 7.

Шаг 6 Задается порог α. Если выбрано, например, сечение XY, то для каждого z_i , где $1 \leq I \leq m$, сканируются точки плоскости XY. Точки на плоскости, где $f(x, y, z) \geq 0$, выводятся на экран. Для дальнейшего просмотра можно выбрать другое α и другое направление сечений.

Шаг 7 Задается порог α. Пространственную фигуру наиболее удобно представлять ограничивающей ее поверхностью. Сама же поверхность будет представлена тремя наборами контуров во взаимно-перпендикулярных плоскостях XY, YZ, XZ.

Контуры в плоскости сечения строятся следующим образом. Вначале определяются граничные точки области, в которой $f(x,y,z) \geq 0$. В массив граничных точек включаются точки, которые хотя бы по одной из двух координат имеют хотя бы одну соседнюю точку, в которой $f(x,y,z) < 0$. В

общем случае массив граничных точек может представлять несколько контуров. Построение отдельного контура выполняется обходом отдельной компактной группы по часовой стрелке, начиная с крайней левой точки (рис. 1).

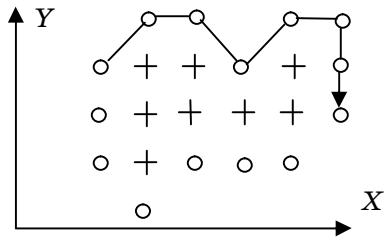

Рисунок 1 – Построение граничного контура:
+ – внутренние точки; o *– граничные точки*

После замыкания полигональной линии точки контура исключаются из рассмотрения, и выполняется переход к следующей группе точек.

Полученные наборы полигональных контуров в совокупности формируют изображения пространственных кластеров.

ВЫВОДЫ

1 Предложена схема представления многомерных статистических данных и отношений между ними в реляционных базах с помощью вложенных поверхностей, формирующих кластеры.

2 Получена непараметрическая оценка трехмерной плотности распределения на основе использования метода перекрестной проверки и ядер Гаусса.

3 Описан алгоритм построения поверхностей кластеров для визуализации на экране компьютера.

Внедрение данной методики позволит значительно эффективнее исследовать многомерные статистические данные в базах и отношения между ними.

SUMMARY

VISUALIZATION OF STRUCTURES IN DATABASES BASED ON NONPARAMETRIC ESTIMATION

O.P. Chekalov, M.S. Babiy, S.P. Shapovalov
Sumy State University

A scheme of presentation of multivariate statistical data and relations between data given in relational base by means of embedded surfaces forming clusters is offered there. The optimization of smoothing parameter for surfaces on the basis of nonparametric estimation by method of the cross validation with removing elements on by one is executed. The algorithm of the embedded surfaces construction is described.

СПИСОК ЛИТЕРАТУРЫ

1. Keim D.A., Kriegel H. Vizualization Techniques for Mining Large Databases: A Comparison // IEEE Transaction on Knowledge and Data Engineering. - 1996. - Vol. 8, No. 6.
2. Воронцов К. В. Лекции по статистическим (байесовским) алгоритмам классификации. //www.ccas.ru/voron/download/Bayes.pdf.
3. Turlach B. A. Bandwidth selection in kernel density estimation: a review. Statistic und Oekonometrie 9307, Humboldt Universitaet Berlin.
4. P. Hall. Large Sample Optimality of Least Squares Cross-Validation in Density Estimation Source: Ann. Statist. -1983. -Vol. 11, No. 4.

Поступила в редакцию 15 января 2009 г.