Алгоритмы обработки многомерных данных

Н.С. Гарколь, В.В. Стругайло

Рассматриваются алгоритмы визуализации многомерных данных

При обработке многомерных данных возникает задача такой их графической визуализации и последующей кластеризации, которая была бы четкой и понятной для пользователя. Это предполагает проекцию распределения данных их многомерного пространства в двух- (реже трех-) мерном пространстве при сохранении основных характеристик распределения в многомерном пространстве.

Предлагаемый авторами метод может применяться и в задачах кластеризации или категоризации многомерных образов «без учителя», а также исследования свойств экспериментальных данных, с последующим поиском паттернов. Паттерны представляют собой закономерности, свойственные подвыборкам данных. Они отражают "скрытые" знания, которые должны быть компактно выражены в наглядной форме представления экспериментальных данных или результатов теоретического исследования. Их поиск производится методами, не ограниченными рамками априорных предположений о структуре выборки и виде распределений значений анализируемых показателей.

Идея метода состоит в следующем: пусть имеется множество N-мерных векторов *Xi*(*i=1,…,m*), которым ставятся в соответствие такое же количество векторов в двухмерном пространстве (обозначим их *Yi*, *i=1,…,m*). Далее в *N*-мерном и двухмерном пространствах необходимо рассчитать расстояния-метрики между векторами (D_{ij}, d_{ij} соответственно). Но следует учитывать, что любая нормировка данных приводит к тому, что изменяются взаимные расстояния между точками данных. Выбор метрики является важным моментом в любой методике анализа структуры данных.

Введем понятие матрицы связи признаков и матрицы удаленностей между объектами.

Матрица связи – квадратная симметричная матрица размерами *mxm* типа «признак-признак», где на пересечении *i*-строки и *j*-столбца стоит мера взаимосвязанности *i*-го и *j*-го признаков. Наиболее часто используемая мера связи – коэффициент корреляции Пирсона, который вычисляется по формуле:

$$r_{ij} = \frac{s_{kj}}{s_{kk} s_{jj}}, \text{ где}$$

$$s_{kj} = \frac{1}{m-1} \sum_{i=1}^{m} (x_{ik} - x_k)(x_{ij} - x_j) .$$

Таким образом, матрицей связи становится корреляционная матрица вида:

$$R = \begin{bmatrix} r_{11} \dots r_{1m} \\ \dots \quad \dots \quad \dots \\ r_{m1} \dots r_{mm} \end{bmatrix}.$$

Матрица удаленностей (или матрица расстояний) - квадратная симметричная матрица размерами *MxM* типа «объект-объект», где на пересечении *i*-строки и *j*-столбца стоит мера «расстояния » между объектами *i*- и *j*.

$$D = \begin{bmatrix} d_{11} \dots d_{1M} \\ \dots \quad \dots \quad \dots \\ d_{M1} \dots d_{MM} \end{bmatrix}.$$

Данные исходно могут быть заданы в виде матриц связи или удаленностей. Тогда возникает задача по заданной матрице восстановить в каком-либо смысле исходное множество точек данных таким образом, чтобы для матриц связности или расстояний имела заданный вид точно или приближенно.

Правило вычисления расстояний между объектами может сильно изменяться в зависимости от специфики задачи. Если такое правило уже задано, то говорят, что в пространстве признаков ведена метрика.

Рассмотрим квадратичные метрики расстояний, для которых квадрат расстояния между объектами является квадратичной формой от разностей значений их координат:

$$d_{ij} = \sqrt{(X_i - X_j)^T G (X_i - X_j)} ,$$

где G- симметричная положительно определенная матрица. В качестве матрицы G размерами *MxM* можно выбрать:

а) единичную матрицу *G=E*, в результате получим обычное Евклидово расстояние:

$$d_{ij} = \sqrt{\sum_{k=1}^{M} (x_{ik} - x_{jk})^2} \ .$$

б) диагональную матрицу $G=diag(g_1, g_2,\ldots, g_M)$, в результате получим взвешенную Евклидову метрику:

$$d_{ij} = \sqrt{\sum_{k=1}^{M} g_k (x_{ik} - x_{jk})^2}$$

.

в) матрицу, обратную ковариационной матрице $G=S^{-1}$:

$$S=\begin{bmatrix} s_{11} \ldots s_{1M} \\ \ldots \ \ldots \ \ldots \\ s_{M1} \ldots s_{MM} \end{bmatrix}, \quad \text{где}$$

$$s_{kj} = \frac{1}{m-1} \sum_{i=1}^{m} (x_{ik} - x_k)(x_{ij} - x_j) \ ,$$

тогда получим меру *Махаланобиса*.

С ковариационной матрицей связано понятие эллипсоида рассеяния облака точек. Осями эллипсоида рассеяния являются направления собственных векторов S (т.к. S-симметричная матрица, то собственные вектора образуют полную ортогональную систему векторов), длины осей выбираются равными значениями соответствующих собственных чисел. Основным преимуществом использования квадратичных метрик является тот факт, что производная от квадрата расстояния, измеренного в такой метрике, является линейной функцией от «координат» объектов, что может быть использовано при решении различных прикладных задач (например, задач оптимизации, классификации и т.п.).

Следующим шагом в изучении свойств исследуемых объектов является настройка метрики. Так, например, при использовании метрик остаются неопределенными веса признаков, хотя иногда условия задачи позволяют выделить те признаки, которые являются «более значимыми» при измерении расстояний, и назначить для них значения весов. Если никаких условий нет, то можно использовать итерационный метод настройки весов метрик:

рассмотрим метрику $d_{ij} = \sqrt{(X_i - X_j)^T G (X_i - X_j)}$, где d_{ij} - расстояние между *i*- и *j*- объектами. Предположим, что *det G*=1. При таком выборе G результаты остаются не менее общими, но при этом возможно рассматривать решения, отличающиеся друг от друга только преобразованием гомотетии (т.е. равномерным растяжением по всем осям). Предположим, что на наборе объектов уже существует определенная система отношений– объекты разбиты на k-непересекающихся классов. Введем матрицу внутриклассового разброса W:

$$W=\frac{1}{N} \sum_{i=1}^{k} \sum_{X_l, X_q \in K_i} (X_l - X_q)(X_l - X_q)^T \ ,$$

где *T*- операция транспонирования; K_i- множество объектов, принадлежащих *i*-классу.

Если выбрать $G=\alpha W^{-1}$, где α - числовой множитель (заданная константа), то минимальной из всех квадратичных метрик оказывается величина внутриклассового разброса

$$W=\frac{1}{N} \sum_{i=1}^{k} \sum_{X_l, X_q \in K_i} d^2 (X_l - X_q) \ ,$$

где $d^2(X_l - X_q)$- квадрат расстояния между *l*-м и *q*-м объектами. В итоге получаем, что классы оказываются максимально компактными[Айвазян Приклад статис].

Если разбиение на классы на классы не задано, то возможна такая настройка метрики, при которой данные будут разбиты на *k*-кластеров «наиболее контрастно».

Предлагаем следующий итерационный алгоритм:

Шаг 1: При некоторой фиксированной метрике $G^{(t)}$ производится разбиение множества данных на *k*-кластеров любым из известных алгоритмов кластерного анализа [Ту Гонсалес]. Число кластеров задается или определяется в начале алгоритма и далее не меняется.

Шаг 2. По полученной классификации строится матрица внутриклассового разброса *W* и вводится метрика $G^{(t+1)} = (W^{(t+1)})^{-1}$.

Данный алгоритм повторяется до тех пор, пока относительные изменения элементов не меньше заданного числа ε :

$$\left|\mathrm{G}^{(t+1)} - \mathrm{G}^{t}\right| \le \varepsilon ,$$

где ε -заданное малое число.

Но необходимо также рассмотреть и такой вариант распределения данных, как например, равномерный, или наоборот, структура сгущения оказалась бы более контрастной. В этом случае необходима процедура контрастирования структуры данных. Тогда рассмотрим такой вариант римановой метрики, где элемент объема риманова пространства вычисляется по формуле:

$$dV = |g| dx^1 dx^2 .. dx^n \text{, где}$$

|*g*| – определитель матрицы метрического тензора.

Оценим нормированную на одну точку плотность распределения данных в исходном пространстве с помощью следующей непараметрической оценки:

$$p(x) = \frac{1}{|x| \cdot r^n} \sum_{i=1}^{N} \prod_{j=1}^{n} K(\frac{x_j - x_j^i}{r})$$

где *K(x)* – некоторая функция, удовлетворяет условию:

$$\int_{\infty}^{\infty} K(x)\,dx = 1$$

В качестве такой функции может быть выбрана:

$$K(x) = \frac{1}{\pi} e^{-x^2}$$

r – радиус «облака» тензора, который является свободным параметром или его оценка возможна методами непараметрической статистики.

Предположим, что в новом пространстве плотность распределения будет постоянна: $p=p_0$, где p_0, можно выбрать, как $p_0 = \frac{1}{V}$, где *V* – исходный объем пространства экспериментальных данных (очевидно, что он конечен).

Предположим, что метрика пространства имеет конформно-плоский вид:

$$dS^2 = \gamma(x) \sum_{ij} \delta_{ij} dx^i dx^j ,$$

где $\gamma(x)$ – конформный множитель, зависящий от точки *x*.

Для того, чтобы плотность данных в новом пространстве стала постоянной, необходимо выбрать:

$$\gamma(x) = (\frac{p_0}{p(x)})^{2/n}$$

или

$$\gamma(x) = (\frac{p(x)}{p_0})^{\alpha} \text{, } \alpha>0$$

получаем метрику, в которой сгущение данных выглядит тем контрастней, чем больше параметров α. Малые расстояния между точками данных в такой метрике становятся еще меньше, большие – еще больше.

Однако, при работе с реальными экспериментальными данными или с результатами теоретических исследований, зачастую отдельные значения координат точек данных могут оказаться либо недостоверными, либо вообще неизвестными и тогда возникает вопрос, как представить такие объекты в многомерном пространстве и каким образом измерить расстояние между объектами в этом случае?

В этих случаях такие объекты необходимо представлять не точкой, а прямой или гиперплоскостью, параллельной координатным осям. Следовательно, и расстояния необходимо вычислять, как расстояние между соответствующими объектами (между точкой и прямой, точкой и плоскостью, и т.д.).

В этих случаях получаем, что вычисления расстояний между объектами с пропущенными значениями признаков: расстояния вычисляются в том подпространстве, в котором значения координат у объектов известны полностью, т.е. будем опускать те слагаемые, в которых отдельные значения признаков не известны.

Например, пусть объект имеет следующие значения признаков:

$$X(x_1, x_2, .. x_{k-1}, E \ldots x_{k+1}, \ldots x_m),$$

где *E* – неизвестное значение признака x_k, обозначим через $x_{(k)}$ – значение которого у объекта *X* неизвестно, и введем:

$X^0(k)=(x_1, x_2 \ldots x_{k-1}, 0, x_{k+1}, \ldots x_m)$, где x_k - заменено нулем.

Тогда геометрический образ, который можно сопоставить объекту x_k – прямая: $X=x^0(k)+e_k t$, где e_k – единичный вектор, k – координата оси.

Пусть значения признаков объекта Y известны полностью, тогда найдем кратчайшее расстояние между $X(x_1,x_2,..,x_m)$ и $Y(y_1, y_2,.., y_m)$,

$$\frac{d}{dt}((X-Y)^2)=0,$$

$(X - Y)e_k = 0$,

$t = Ye_k = y_k$,

Следовательно, при вычислении расстояния неизвестное значение K-ого признака у X необходимо заменить значением некоторого признака Y, но тогда:

$(X - Y)^2 = (X^0(k) - Y)^2 - y_k = (X^0(k) - Y^0(k))^2$.

То есть в этом случае достаточно приравнять к нулю значения K-ого признака объектов X и Y, тогда получаем, что расстояние между объектами:

$X(x_1,x_2,..,x_m)$ и $Y(y_1,y_2,..,y_m)$ можно вычислить

$$d(X-Y)^2=\sum_{i=1}^{m}(x_i-y_i)^2,\ x_i \neq E.$$

Рассматривая ситуацию, когда у объекта X неизвестно значение признака x_k, а у Y значение y_k, тогда:

$Y = Y^0(l) + e_l s$,

$$\left.\begin{aligned}\frac{d}{dt}((X-Y)^2)=0\\ \frac{d}{ds}((X-Y)^2)=0\end{aligned}\right\},$$

$$\left\{\begin{aligned}(X-Y)^2 e_k=0\\ (X-Y)^2 e_l=0\end{aligned}\right\},$$

$$\left.\begin{aligned}-Y^0(l)e_k+t-\delta_{kl}s=0\\ -X^0(k)e_l+\delta_{kl}t-s=0\end{aligned}\right\},$$

если $k \neq l$, то $t = y_k$, $s = x_k$ и результат тот же, что и в случае точки и прямой.

Если $k = l$, то $t = s$, $(X - Y)^2 = (X^0(k) - Y^0(k))^2$, и тогда просто пропускаем неизвестное значение признака.

Тогда при реализации алгоритма проектирования N-мерных данных в в двух- (реже трех-) мерное пространство при сохранении основных характеристик распределения в многомерном пространстве решение задачи сводится к минимизации погрешности E, определяемую по формуле: $E=\frac{1}{2}\sum_{i=1,\ldots i\neq j}^{m}(D_{ij}-d_{ij})^2$,

$$d_{ij}=\sqrt{(y_1^i-y_1^j)^2+(y_2^i-y_2^j)^2}.$$

Расстояния D_{ij} –меры близости N-мерных векторов - являются неизменными, а d_{ij} –расстояния в двухмерном пространстве итерационно будут пересчитываться по мере корректировки компонент y_1 и y_2 для каждого i-го вектора по следующему правилу: $y_1= y_1-\Delta y_1$, $y_2= y_2-\Delta y_2$, где

$\Delta y_1=\partial E/\partial y_1$, $\Delta y_2=\partial E/\partial y_2$,

$$\frac{\partial E}{\partial y_1^i}=-\sum_{i=1,i\neq j}^{m}\frac{D_{ij}-d_{ij}}{d_{ij}}(y_1^i-y_1^j),$$

аналогично и для компоненты y_2

$$\frac{\partial E}{\partial y_2^i}=-\sum_{i=1,i\neq j}^{m}\frac{D_{ij}-d_{ij}}{d_{ij}}(y_2^i-y_2^j).$$

Геометрическую метафору облака точек в многомерном пространстве можно сопоставить данным, изначально представленным в виде таблицы «объект-признак». Главное достоинство предлагаемых алгоритмов становится заметным только при обработке многомерных данных, пространственное размещение которых человек уже не в состоянии себе представить. Механизмы предложенных алгоритмов функционируют независимо от размерности задачи. Такой подход позволяет определить зоны концентрации данных в многомерном пространстве и основные характеристики их распределения, существенные с точки зрения пользователя.

Литература:

1. Форсайт, Д. Компьютерное зрение. Современный подход : [пер. с англ.] / Д. Форсайт, А. Дэвид, Жан Понс. – М.: Издательский дом "Вильямс", 2004. –928 с.: ил.

2. Круглов, В.В. Искусственные нейронные сети. Теория и практика / В.В. Круглов, В.В. Борисов. – М.: Горячая линия – Телеком, 2001. – 382 с. : ил.

3. Осовский С. Нейронные сети для обработки информации / Пер. с польского И.Д. Рудинского. – М.: Финансы и статистика, 2002.