ОБОЗРЕНИЕ
ПРИКЛАДНОЙ И ПРОМЫШЛЕННОЙ
Том 19 МАТЕМАТИКИ Выпуск 4
2012

А. В. Бернштейн, **А. П. Кулешов** (Москва, ИСА РАН, ИППИ РАН)**. Тангенциальная близость в процедурах снижения размерности.**

Задача снижения размерности формулируется как задача моделирования многообразий (Manifold Learning): по выборке $\mathbf{X}_n = \{X_1, X_2, \ldots, X_n\}$ из неизвестного q-мерного многообразия данных $\mathbf{X} = \{X = f(b) \in \mathbf{R}^p : \ b \in \mathbf{B} \subset \mathbf{R}^q\}$ в $\mathbf{R}^p$, покрытого одной координатной системой (картой) f, необходимо построить отображение вложения h многообразия $\mathbf{X}$ в q-мерное множество $\mathbf{Y} = h(\mathbf{X}) \subset \mathbf{R}^q$ меньшей размерности $(q < p)$ и отображение восстановления g множества $\mathbf{Y}$ в $\mathbf{R}^p$, обеспечивающие приближенные равенства $r(X) \equiv g(h(X)) \approx X$ для всех $X \in \mathbf{X}$. Отображение h есть отображение снижения размерности, позволяющее вместо p-мерного вектора X рассмотреть q-мерный вектор $y = h(X)$ без значимой потери информации о векторе X — по вектору y можно построить вектор $X^* = g(y)$, близкий к исходному вектору X: $X^* \approx X$.

Построенное по данным отображение r переводит многообразие $\mathbf{X}$ в q-мерное эмпирическое многообразие $\mathbf{X}_{\text{эмп}} = r(\mathbf{X}) = \{X = g(y) \in \mathbf{R}^p : \ y \in \mathbf{Y} = h(\mathbf{X}) \subset \mathbf{R}^q\}$ в $\mathbf{R}^p$, обеспечивая близость этих многообразий: $\mathbf{X} \approx \mathbf{X}_{\text{эмп}}$. Пусть q-мерные аффинные пространства $T(X) \equiv X \oplus L(X)$ и $T_{\text{эмп}}(r(X)) \equiv r(X) \oplus L_{\text{эмп}}(r(X))$ являются касательными пространствами к многообразиям $\mathbf{X}$ и $\mathbf{X}_{\text{эмп}}$ в точках $X \in \mathbf{X}$ и $r(X) \in \mathbf{X}_{\text{эмп}}$ соответственно, а q-мерные линейные подпространства L и $L_{\text{эмп}}$ рассматриваются как элементы многообразия Грассмана [1], состоящего из q-мерных линейных подпространств в $\mathbf{R}^p$. В [2] показано, что для обеспечения хорошей обобщающей способности процедуры снижения размерности необходимо, чтобы пара отображений $\theta = (h, g)$ обеспечивала не только близость многообразий $\mathbf{X}$ и $\mathbf{X}_{\text{эмп}}$, но и тангенциальную близость $L(X) \approx L_{\text{эмп}}(X)$ в некоторой метрике на многообразии Грассмана для всех $X \in \mathbf{X}$.

Множество $T\mathbf{X} = \{(X, L(X)) : \ X \in \mathbf{X}\}$ называется *касательным расслоением* (tangent bundle) многообразия $\mathbf{X}$. Рассмотрим новую задачу моделирования многообразий с касательным расслоением (Tangent Bundle Manifold Learning), в которой по выборке $\mathbf{X}_n$ необходимо построить пару отображений $\theta = (h, g)$, определяющих эмпирическое касательное расслоение $T\mathbf{X}_{\text{эмп}} = \{(X', L_{\text{эмп}}(X')) : \ X' \in \mathbf{X}_{\text{эмп}}\} \equiv \{(r(X), L_{\text{эмп}}(r(X)) : \ X \in \mathbf{X}\}$, близкое к $T\mathbf{X}$, обеспечивая близость многообразий $\mathbf{X} \approx \mathbf{X}_{\text{эмп}}$ и близость тангенциальных q-мерных подмногообразий $\mathbf{L}$ и $\mathbf{L}_{\text{эмп}}$ в многообразии Грассмана: $\mathbf{L} = \{L(X) : \ X \in \mathbf{X}\} \approx \mathbf{L}_{\text{эмп}} = \{L_{\text{эмп}}(r(X)) : \ X \in \mathbf{X}\}$.

Предложен GE-алгоритм (Grassmann Eigenmaps) решения Задачи моделирования многообразий с касательным расслоением, основанный на спектральном вложении касательных пространств, состоящий из нескольких этапов. Сначала по выборке $\mathbf{X}_n$ строится эмпирическое подмногообразие $\mathbf{L}_{\text{эмп}} \approx \mathbf{L}$, элементы которого будут касательными пространствами к построенному в дальнейшем эмпирическому многообразию $\mathbf{X}_{\text{эмп}}$. Элементы $L_{\text{эмп}}(X) \in \mathbf{L}_{\text{эмп}}$ имеют вид $L_{\text{эмп}}(X) = \text{Span}\,(H(X))$, где $(p \times q)$-матрицы $H(X)$ строятся как решения усредненной Прокрустовой задачи. Отображения h и g будут построены в дальнейшем таким образом, чтобы выполнялись

соотношения $H(X) \approx J_g(h(X))$ для всех X, где $J_g(h(X))$ – якобиан отображения $g(y)$ в точке $y = h(X)$, тогда для близких точек $X, X' \in \mathbf{X}$ будут выполняться соотношения

$$g(h(X')) - g(h(X)) \approx H(X)\,(h(X') - h(X)), \tag{1}$$

которые при условии $g(h(X)) \approx X$ могут быть заменены соотношениями

$$X' - X \approx H(X)(h(X') - h(X)). \tag{2}$$

Рассматривая соотношения (2) для близких точек выборки как регрессионные уравнения относительно $h(X)$, строится отображение h, а затем при помощи уравнения (1) строится отображение g.

В докладе будут также приведены результаты сравнительного анализа предложенного алгоритма и известных алгоритмов, используемых в задаче моделирования многообразий.

СПИСОК ЛИТЕРАТУРЫ

1. *Lee J. M.* Introduction to smooth manifolds. New-York: Springer–Verlag, 2003.
2. *Бернштейн А. В.* Локальная обобщающая способность в задаче восстановления по данным нелинейного многообразия. — Обозрение прикл. и промышл. матем., 2012, т. 19, в. 3.