УДК 004.032.26:519.174.2

А.В. Шкловец, Н.Г. Аксак

Харьковский национальный университет радиоэлектроники, Харьков

ПРОЕЦИРОВАНИЕ БОЛЬШОГО КОЛИЧЕСТВА МНОГОМЕРНЫХ ДАННЫХ НА ДВУМЕРНУЮ КУСОЧНО-ГЛАДКУЮ САМООРГАНИЗУЮЩУЮСЯ КАРТУ КОХОНЕНА

В работе предложен метод отображения многомерных данных на двумерную кусочно-гладкую самоорганизующуюся карту Кохонена полученную путем аппроксимации двумерной кусочно-линейной самоорганизующейся карты Кохонена кубической параметрической сплайн поверхностью для решения задачи визуализации большого количества многомерных данных. Использование таких карт позволяет увеличить точность визуализации и различать данные на карте при незначительном увеличении количества скалярных операций, необходимых на их построение.

Ключевые слова: *визуализация многомерных данных, двумерная кусочно-гладкая карта Кохонена, кубическая параметрическая сплайн поверхность.*

Введение

В связи со стремительным накоплением информации о системах или процессах и методов её анализа, всё более актуальным становится вопрос о поиске закономерностей, заключенных в большом количестве параметров данных. Для анализа многомерных данных путем наглядного представления их структуры или результатов исследования, необходимо решить задачу визуализации многомерных данных.

Классическим методом решения задачи визуализации многомерных данных является метод главных компонент [1], основная идея которого заключается в построении линейного многообразия (карты), на которую отображаются данные. Этот метод эффективно решает поставленную задачу в случае простой, унимодальной структуры данных. Для более сложных структур данных был разработан ряд нелинейных методов визуализации многомерных данных, таких как карты Саммона [2], метод главных многообразий [3], самоорганизующиеся карты Кохонена [4], упругие карты [5] и другие.

Карты Саммона, метод главных многообразий и упругие карты предполагают нелинейную структуру карты. Алгоритм построения этих карт заключается в последовательном решении большого числа сложных задач оптимизации, что делает эти методы неэффективными при визуализации большого числа многомерных данных. Простой и быстрый алгоритм обучения самоорганизующихся карт Кохонена позволяет строить кусочно-линейные карты, быстро аппроксимирующие структуру многомерных данных.

Данный метод позволил решить глобальную проблему визуализации многомерных данных со сложной топологией, однако приводит к ряду других проблем. Разработанные модификации самоорганизующихся карт Кохонена, такие как метод регуляризации [6], алгоритмы Bacth SOM [7], Adaptive SOM [8], иерархические алгоритмы [9], нейронный газ [10] и другие, ликвидируют многие недостатки самоорганизующихся карт Кохонена. Однако проблема отображения данных на ребра или вершины двумерных кусочно-гладких карт Кохонена вызвана кусочно-линейной структурой карты, и не может быть решена путем модификации алгоритма обучения нейронной сети Кохонена. Решение приведенной проблемы основано на построении кусочно-гладких самоорганизующихся карт Кохонена путем аппроксимации двумерных кусочно-линейных самоорганизующихся карт Кохонена кубическими параметрическими сплайн поверхностями [11 – 14].

Постановка задачи. Пусть после обучения карты Кохонена на множестве данных X в n-мерном евклидовом пространстве множество выходных нейронов $W = \{W^1,...,W^l\}$ имеет координаты $W^i = \left(w_1^i,...,w_n^i\right)^T$, $i = \overline{1,l}$, l – количество выходных нейронов.

На множестве W задана триангуляция $T = \left\{T^1,...,T^M\right\}$, где $T^k = \left\{T_1^k, T_2^k, T_3^k\right\}$, $k = \overline{1,M}$, M – количество треугольников, T_p^k – номер нейрона в множестве W, $T_p^k \in \{1,N\}$, $p = \overline{1,3}$. Двумерная кусочно-линейная самоорганизующаяся карта Кохонена $\langle W,T\rangle$ была аппроксимирована кубической параметрической сплайн поверхностью

$$S(t_1,t_2) = \bigcup_{k=1}^{M} S^k(t_1,t_2)\Big|_{T^k}, \qquad (1)$$

где

$$S^k(t_1,t_2)\Big|_{T^k} = \left(s_1^k(t_1,t_2),..,s_n^k(t_1,t_2)\right),$$

$$s_p^k(t_1,t_2)=\sum_{i=0}^{3}\sum_{j=0}^{3}a_{ijp}^k t_1^i t_2^j,\ k=\overline{1,M},\ p=\overline{1,n},\quad (2)$$

$A=\left(a_{ijp}^k\right)_{i=\overline{0,3};j=\overline{0,3};p=\overline{1,n}}^{k=\overline{1,M}}$ – коэффициенты, определённые из условий гладкости и минимизации отклонения кусочно-линейной самоорганизующейся карты Кохонена кубической параметрической сплайн поверхностью. Построенная кубическая параметрическая сплайн поверхность является двумерной кусочно-гладкой самоорганизующейся картой Кохонена.

Требуется отобразить множество данных X на двумерную кусочно-гладкую самоорганизующуюся карту Кохонена $S(t_1,t_2)$.

Отображение многомерных данных на двумерную кусочно-гладкую самоорганизующуюся карту Кохонена

Из рис. 1 видно, что проекции точки данных элемента данных $X^c\in X$ на двумерную кусочно-гладкую самоорганизующуюся карту Кохонена $S(t_1,t_2)$, определяются решением системы уравнений

$$\begin{cases}\dfrac{\partial S(t_1,t_2)}{\partial t_1}.\left(S(t_1,t_2)-X^c\right)=0,\\ \dfrac{\partial S(t_1,t_2)}{\partial t_2}.\left(S(t_1,t_2)-X^c\right)=0,\end{cases}\quad (3)$$

$$c=\overline{1,N},$$

где $(.)$ – операция скалярного умножения. Учитывая (1) и (2) система уравнений (3) распишется в виде M систем уравнений

$$\begin{cases}F_1^k(t_1,t_2)=\dfrac{\partial S^k(t_1,t_2)}{\partial t_1}.\left(S^k(t_1,t_2)-X^c\right)=0,\\ F_2^k(t_1,t_2)=\dfrac{\partial S^k(t_1,t_2)}{\partial t_2}.\left(S^k(t_1,t_2)-X^c\right)=0,\end{cases}\quad (4)$$

$$k=\overline{1,M}\ ,\ c=\overline{1,N},$$

причем $S^k(t_1,t_2)$ ограничено областью T^k.

Раскрыв скобки в системах (4), видно, что её первые функции $F_1^k(t_1,t_2)$ являются полиномами соответственно 5-го и 6-го порядка, а вторые функции $F_2^k(t_1,t_2)$ являются полиномами соответственно 6-го и 5-го порядка относительно переменных t_1 и t_2. Таким образом, система (4) не может быть решена аналитически. Учитывая, что сплайн поверхность $S(t_1,t_2)$ минимально отклоняется от кусочно-плоской карты, а значит, она не претерпевает большого числа изгибов, и поэтому в области T^k поверхности $S^k(t_1,t_2)$ не может существовать большого числа действительных корней системы (4).

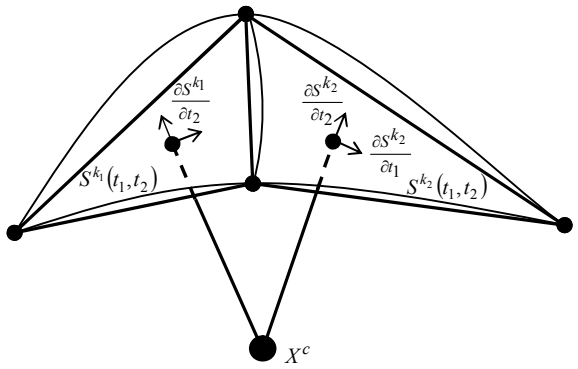

Рис. 1. Отображение элемента X^c на двумерную кусочно-гладкую самоорганизующуюся карту Кохонена $S(t_1,t_2)$

В области T^k распределим равномерно p (положим $p=10$) точек $\left({}^jt_1^0,\ {}^jt_2^0\right)$ $j=\overline{1,p}$, являющихся начальными приближениями для поиска корней.

Для нахождения i-го приближения воспользуемся итерационными формулами метода Ньютона

$$\begin{cases}{}^jt_1^{i+1}={}^jt_1^i+\dfrac{{}^jT_1^i}{{}^j\Im^i},\\ {}^jt_2^{i+1}={}^jt_2^i+\dfrac{{}^jT_1^i}{{}^j\Im^i}\end{cases},\ j=\overline{1,p},\quad (5)$$

где

$${}^j\Im^i=\det\begin{Vmatrix}\dfrac{\partial F_1^k\left({}^jt_1^i,\ {}^jt_2^i\right)}{\partial t_1} & \dfrac{\partial F_1^k\left({}^jt_1^i,\ {}^jt_2^i\right)}{\partial t_2}\\ \dfrac{\partial F_2^k\left({}^jt_1^i,\ {}^jt_2^i\right)}{\partial t_1} & \dfrac{\partial F_2^k\left({}^jt_1^i,\ {}^jt_2^i\right)}{\partial t_2}\end{Vmatrix},$$

$${}^jT_1^i=\det\begin{Vmatrix}F_1^k\left({}^jt_1^i,\ {}^jt_2^i\right) & \dfrac{\partial F_1^k\left({}^jt_1^i,\ {}^jt_2^i\right)}{\partial t_2}\\ F_2^k\left({}^jt_1^i,\ {}^jt_2^i\right) & \dfrac{\partial F_2^k\left({}^jt_1^i,\ {}^jt_2^i\right)}{\partial t_2}\end{Vmatrix},\quad (6)$$

$${}^jT_2^i=\det\begin{Vmatrix}F_1^k\left({}^jt_1^i,\ {}^jt_2^i\right) & \dfrac{\partial F_1^k\left({}^jt_1^i,\ {}^jt_2^i\right)}{\partial t_1}\\ F_2^k\left({}^jt_1^i,\ {}^jt_2^i\right) & \dfrac{\partial F_2^k\left({}^jt_1^i,\ {}^jt_2^i\right)}{\partial t_1}\end{Vmatrix}.$$

Если точка $\left({}^{j'}t_1^{i+1},\ {}^{j'}t_2^{i+1}\right)\notin T^k$, то она более не рассматривается. Предполагается, что точка, вышедшая за границы области T^k, более в неё не вернется. Таким образом, значительно снижается коли-

чество скалярных операций, необходимых для нахождения всех корней систем уравнений (4).

Критерием останова может быть как некоторое заданное количество итераций i_{max} (положим $i_{max} = 10$), так и условие

$$\begin{cases} \left| {}^{j}t_1^{i+1} - {}^{j}t_1^{i} \right| < \varepsilon \\ \left| {}^{j}t_2^{i+1} - {}^{j}t_2^{i} \right| < \varepsilon \end{cases}, j = \overline{1,p}. \quad (7)$$

Если для двух точек $\left({}^{j}t_1^{i}, {}^{j}t_2^{i} \right)$ и $\left({}^{j'}t_1^{i}, {}^{j'}t_2^{i} \right)$ выполняется условие

$$\begin{cases} \left| {}^{j}t_1^{i} - {}^{j'}t_1^{i} \right| < \varepsilon \\ \left| {}^{j}t_2^{i} - {}^{j'}t_2^{i} \right| < \varepsilon \end{cases}, j = \overline{1,p}, \quad (8)$$

что означает, что начальные приближения $\left({}^{j}t_1^{0}, {}^{j}t_2^{0} \right)$ и $\left({}^{j'}t_1^{0}, {}^{j'}t_2^{0} \right)$ сошлись к одному корню, поэтому точку $\left({}^{j'}t_1^{i}, {}^{j'}t_2^{i} \right)$ можно не рассматривать.

Проекцией элемента данных X^c на карту назовем точку

$$\tilde{X}^c = \min_{\substack{j=\overline{0,p^k} \\ k=\overline{1,M}}} \left\| X^c - S^k \left({}^{j}t_1^{i}, {}^{j}t_2^{i} \right) \right\|, \quad (9)$$

$$c = \overline{1,N}.$$

Множество $\tilde{X} = \left(\tilde{X}^1, ..., \tilde{X}^N \right)$ является плоской визуализацией данных X.

Численные эксперименты

Для определения качества визуализации данных картами Кохонена предлагается использовать следующие критерии.

Критерий различимости данных (количество данных отображаемых в одну и ту же точку на карте):

$$K_r = \sum_{c=1}^{M} \left(m\left(\tilde{X}^c \right) - 1 \right), \quad (10)$$

где $m\left(\tilde{X}^c \right)$ – количество элементов $\tilde{X}^c \in \tilde{X}$, $\sum_{c=1}^{M} m\left(\tilde{X}^c \right) = N$, M – количество точек на карте.

1. Количество точек данных, отображенных в вершину или на границу карты Кохонена

$$K_g = \left| \bar{\bar{X}} \right|,$$

$$\bar{\bar{X}} = \left\{ \tilde{X}^c \left| \tilde{X}_1^c \in Z \vee \tilde{X}_1^c = 1 - b \vee \tilde{X}_1^c = 1 + b \vee \right. \right. \quad (11)$$

$$\left. \left. \vee \tilde{X}_2^c \in Z \vee \tilde{X}_2^c = 1 - b \vee \tilde{X}_2^c = 1 + b \right\}, \; c = \overline{1,N}.$$

2. Среднее квадратичное расстояние точек в n-мерном евклидовом пространстве до карты Кохонена

$$K_o = \sqrt{\frac{1}{N} \sum_{c=1}^{N} \rho^2 \left(X^c, \tilde{X}^c \right)}. \quad (12)$$

3. Критерий ошибки визуализации данных в евклидовой метрике (оценка изменения в структуре данных после отображения на карту)

$$K_v = \sqrt{\frac{2}{N(N-1)} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \left(\frac{\rho\left(X^i, X^j \right)}{V} - \frac{\rho\left(\tilde{X}^i, \tilde{X}^j \right)}{\tilde{V}} \right)^2}, \quad (13)$$

где

$$V = \max_{\substack{i=\overline{1,N-1} \\ j=\overline{i+1,N}}} \rho\left(X^i, X^j \right), \; \tilde{V} = \max_{\substack{i=\overline{1,N-1} \\ j=\overline{i+1,N}}} \rho\left(\tilde{X}^i, \tilde{X}^j \right). \quad (14)$$

Численные эксперименты проводились для базы данных компаний из стран СНГ рынка нефтехимии. Визуализация компаний проводилась на основе прямоугольной карты Кохонена размером 9×9. Компании характеризовались следующими показателями: страна, вид полимера, прирост с начала года в тоннах и процентах, прирост за последний месяц в тоннах и процентах, доля в импорте с начала года и с начала месяца в процентах, импорт с начала года и за последний месяц в тоннах.

Результаты сравнительного анализа методов визуализации данных по предложенным критериям приведены в табл. 1.

Из таблицы видно, что использование кусочно-гладких карт Кохонена в виде параметрической сплайн поверхности позволяет практически полностью различить данные на карте. Кроме того, значительно уменьшилось среднее квадратичное отклонение данных от карты и уменьшилась ошибка визуализации многомерных данных.

Таблица 1

Сравнение методов визуализации многомерных данных

Критерий	Отображение в нейроны	Кусочно-плоская карта	Кусочно-гладкая карта
K_r	186	25	3
K_g	250	123	13
K_o	0,0486	0,0460	0,0270
K_v	0,3001	0,2940	0,2580

Заключение

Для улучшения отображения структуры большого количества многомерных данных при их визуализации с помощью двумерных кусочно-плоских карт Кохонена в работе рассмотрен метод построения двумерных кусочно-гладких карт Кохонена и впервые предложен метод отображения многомерных данных на них. Проведенные численные эксперименты показали, что использование двумерных кусочно-гладких самоорганизующихся карт Кохонена позволяет различать многомерные данные на карте и благодаря малому количеству скалярных операций, необходимых на реализацию приведенных методов, эффективно решает задачу визуализации большого числа многомерных данных.

Список литературы

1. Pearson K. On lines and planes of closest fit to systems of points in space / K. Pearson // Philosophical Magazine, (1901) 2. – P. 559-572.

2. Sammon JW. A nonlinear mapping for data structure analysis / JW. Sammon // IEEE Transactions on Computers (1969)18. – P. 401-409.

3. Hastie T. Principal Curves and Surfaces. Ph. D. Dissertation / Hastie T. – Stanford Linear Accelerator Center, Stanford University, Stanford, California, US, November 1984.

4. Kohonen T. Self-Organizing Maps / T. Kohonen. – Springer: Berlin-Heidelberg, 1997.

5. Зиновьев А.Ю. Визуализация многомерных данных [Текст] / А.Ю. Зиновьев. – Красноярск: Изд-во КГТУ, 2000. – 168 с.

6. Goppert J. Regularized SOM-Training: A Solution to the Topology-Approximation Dilemma? / J. Goppert // Proc. Of International Conference on NetWorks. Washington, DC, 1996. Vol.1. – P. 38-44.

7. LeBlank M. Adaptive principal surfaces / M. LeBlank, N. Tibshorany // Journal of the American Statistical Association. – 1994. – Mar. V. 89, No. 425. – P. 53-66.

8. Kiviluoto K. Two-level self-organizing-map's for analysis of financiel statement / K. Kiviluoto, P. Bergius // Proc. Of international joint conference on Neural NetWorks. – Washington, DC, 1998.

9. Back B. Analyzing Financial Performance with Self-Organized Maps / B. Back, K. Sere, H. Vanharanta // Proc. of International joint conference on Neural NetWorks. – Washington, DC, 1998.

10. Fritzke Bernd. Growing cell structures – a self-organizing network for unsupervised and supervised learning / Bernd Fritzke // In Neural Networks, 7(9):-1460, 1994.

11. Шкловец А.В. Аппроксимация двумерных карт Кохонена кубическими сплайн поверхностями [Текст] / А.В. Шкловец, Н.Г. Аксак // Сб. науч. тр. – Евпатория: МОИНУ, 2010. – 225 с.

12. Шкловец А.В. Метод аппроксимации сплайн поверхностями кусочно-плоских карт Кохонена для визуализации многомерных данных [Текст] / А.В. Шкловец, Н.Г. Аксак // XIII всероссийская научно-техническая конференция «Нейроинформатика-2012»: Сб. науч. трудов. В 3 частях. Ч.1. М.: НИЯУ МИФИ, 2011. – 208 с.

13. Шкловец А.В. Построение четырёхугольных карт Кохонена на основе триангуляции Делоне для визуализации многомерных данных [Текст] / А.В. Шкловец, Н.Г. Аксак // Бионика интеллекта: науч. техн. журнал. – 2011. – 3(77). – С. 94-97.

14. Шкловец А.В. Метод построения кусочно-гладких карт на основе четырёхугольных кусочно-плоских карт Кохонена [Текст] /А.В. Шкловец // Системи обробки інформації. – Х.: ХУ ПС, 2012. – Вип. 2 (100). – С. 168-175.

Поступила в редколлегию 2.03.2012

Рецензент: д-р техн. наук, проф. О.Г. Руденко. Харьковский национальный университет радиоэлектроники, Харьков.

ПРОЕКТУВАННЯ ВЕЛИКОЇ КІЛЬКОСТІ БАГАТОВИМІРНИХ ДАНИХ НА ДВОВИМІРНУ ЧАСТИНО-ГЛАДКУ КАРТУ КОХОНЕНА, ЩО САМООРГАНІЗУЄТЬСЯ

А.В. Шкловець, Н.Г. Аксак

В статі запропоновано метод відображення багатовимірних даних на двовимірну частино-гладку мапу Кохонена, отриману шляхом апроксимації двовимірної частино-лінійної мапи Кохонена кубічною параметричною сплайн поверхнею для розв'язку задачі візуалізації великої кількості багатовимірних даних. Застосування цих мап дозволяє збільшити точність візуалізації та відокремити дані на мапі при незначному збільшені кількості скалярних операцій, необхідних на їх побудову.

Ключові слова: *візуалізація багатовимірних даних, двовимірна кусочно-гладка карта Кохонена, кубічна параметрична сплайн поверхня.*

PROJTCTION A LARGE NUMBER OF MULTIDIMENSIONAL DATA TO TWO-DIMENSIONAL PIECEWISE SMOOTH SELF ORGANIZATION KOHONEN MAP

N.G. Axak, A.V. Shklovets

A method of displaying to a two-dimensional piecewise smooth self organization Kohonen map, obtained by approximating two-dimensional piecewise-linear self organization Kohonen map by cubic parametric spline surface for solving a problem of visualization a large number of multidimensional data were proposed. Using this map can increase the accuracy of visualization and distinguish between data on a map with slight increase in the number of scalar operations required for their construction.

Keywords: *visualization of multidimensional information, двумерная piece-smooth card of Kokhonena, cube self-reactance spline surface.*