ИССЛЕДОВАНИЯ СОВРЕМЕННЫХ РЫНКОВ

А. А. Жеребцов, Ю. А. Куперин

ПРИМЕНЕНИЕ САМООРГАНИЗУЮЩИХСЯ КАРТ КОХОНЕНА ДЛЯ КЛАСТЕРИЗАЦИИ ИНДЕКСОВ DJIA И NASDAQ 100

В настоящей работе решена задача сегментации (кластеризации) наборов корпоративных акций, формирующих финансовые индексы Dow-Jones Industrial Average (DJIA) и NASDAQ 100 методом самоорганизующихся карт Кохонена (СОК). В работе мы развиваем применение данного метода как альтернативу метода ультраметрических пространств. В статье показано, что нелинейный метод СОК является более точным, гибким и перспективным в задачах кластеризации большого объема плохо структурированных данных.

ВВЕДЕНИЕ

В работе представлены две взаимосвязанные проблемы: кластеризации и выявления новых структур и паттернов в массивах финансовых данных методом самоорганизующихся карт Кохонена (СОК) и сравнения метода искусственных нейронных сетей (искусственного интеллекта) с традиционными методами финансовой математики — подходом ультраметрических пространств. Прикладная часть исследования выполнена на анализе обоими методами реальных финансовых данных: портфелях акций, составляющих фондовые индексы DJIA и NASDAQ 100.

Указанные проблемы тесно связаны с задачами менеджмента организаций, поскольку принятие решений корпоративными финансовыми менеджерами зависит как от понимания ими сложных (нелинейных) закономерностей потоков данных, так и от современных методов извлечения знаний (data mining) в целях оперативного и стратегического управления. Кроме того, получение знаний из имеющейся информации является базой для управления капиталом и любой инвестиционной деятельности вообще. Новизна проблематики состоит в развитии интеллектуальных онлайновых

© А. А. Жеребцов, Ю. А. Куперин, 2005

технологий анализа данных и сравнении их с традиционными подходами финансовой математики. Проблемы кластеризации больших многомерных плохо структурированных данных, рассматриваемых в настоящей работе, встречаются в различных областях управления финансовыми потоками, например, в процессе маркетинговой работы организации, которая ищет доходный сегмент рынка, соответствующий ее ресурсам и возможностям. Управление инвестиционным портфелем также требует решения задачи кластеризации данных, поскольку для реализации портфельного инвестирования очень важно выявление корреляций как самих активов, так и их групп.

В последние годы в теории кластеризации данных наметился переход от простых концептуальных моделей к статистическим методам. В этом ключе особое место стал занимать кластерный анализ. Термин «кластерный анализ» (впервые ввел Трайон [Tryon, 1939]) в действительности включает в себя набор различных алгоритмов классификации. Техника кластеризации применяется в самых разнообразных областях. В общем, всякий раз, когда необходимо классифицировать «горы» информации на пригодные для дальнейшей обработки группы, кластерный анализ оказывается весьма полезным и эффективным.

Существует несколько методов обобщения наборов данных или статистических таблиц. Простейшие из них позволяют получить сводки информации. Например, наименьшие и наибольшие значения в наборе данных, медиану, первый и третий квартили. Такие простые методы очень полезны при обобщении наборов данных небольшой размерности. Обобщать и визуализировать данные большой размерности труднее. В данной работе мы развиваем метод, который может быть использован для обобщения и визуального представления больших наборов многомерных данных.

Обычно выборочные совокупности данных состоят из множества величин, которые могут соответствовать некоторому набору статистических показателей. Последний может быть выражен вектором, что означает просто упорядоченный набор числовых величин.

В том случае, когда существует только два или три измерения, достаточно легко построить простые двумерные и трехмерные графики. Однако если размерность данных больше, то изобразить вектор или взаимосвязь между различными векторами графически весьма не просто. Именно поэтому необходимы другие методы визуального представления.

Самоорганизующаяся карта (СОК) представляет собой специальный тип искусственной нейронной сети, которая позволяет осуществлять нелинейную регрессию и отображать многомерные данные на двумерную плоскость с сохранением расстояний в изначальном пространстве данных [Kohonen, 1982; 1990; 1995; Kohonen et al., 1996]. Методы СОК имеют мно-

жество применений в менеджменте, экономике, финансах и маркетинге [Back et al., 1996; Demartines, 1994; Carlson, 1991; Cheng et al., 1994; Garavaglia, 1993; Martin-del-Brio, Serrano-Cinca, 1993; Marttinen, 1993; Serrano-Cinca, 1996; Ultsch, 1993; Дебок, Кохонен, 2001]. Например, могут быть использованы в целях:

- анализа финансовой отчетности компаний;
- изучения перспективности инвестиций во взаимные фонды;
- оценки недвижимости;
- исследования товарных рынков на основе анализа потребительских предпочтений и аттитюдов;
- сегментации покупателей и клиентов;
- информационного обеспечения выработки маркетинговых стратегий;
- долгосрочного прогнозирования динамики процентных ставок и выявления предпосылок к банкротству предприятий.

Список, приведенный выше, может быть, разумеется, продолжен (см., напр.: [Дебок, Кохонен, 2001]).

Переходя к предмету настоящей работы, отметим, что в книге [Mantegna, Stanley, 2001] использовался метод ультраметрических пространств для кластеризации индексов DJIA и NASDAQ 500. Здесь мы применяем как ультраметрический, так и методы СОК для кластеризации DJIA и NASDAQ 100 портфелей. Мы показываем, что нелинейный метод СОК является более точным, гибким и перспективным для задач кластеризации большого объема плохо структурированных данных. Таким образом, общей целью является построение альтернативных нелинейных адаптивных методов для решения задач кластеризации и выявления новых структур и паттернов в массивах финансовых данных.

НЕЙРОННЫЕ СЕТИ И МЕТОДОЛОГИЯ САМООРГАНИЗУЮЩИХСЯ КАРТ КОХОНЕНА

Любая искусственная нейронная сеть состоит из относительно простых, в большинстве случаев однотипных, функциональных элементов, имитирующих работу нейронов мозга. Каждый нейрон (рис. 1) обладает группой синапсов — однонаправленных входных связей, соединенных с выходами других нейронов, а также имеет аксон — выходную связь данного нейрона, с которой сигнал (возбуждения или торможения) поступает на синапсы следующих нейронов. Каждый синапс характеризуется величиной синаптической связи или ее весом w_{ij} (вес связи нейрона i с нейроном j), который по физическому смыслу эквивалентен электрической проводимости. Текущее состояние нейрона определяется как взвешенная с его синаптическими весами сумма n входов:

$$S = \sum_{i=1}^{n} X_i \cdot w_i,$$

где стоящая справа сумма интерпретируется как состояние нейрона или мера активности нейрона. Выход нейрона есть функция его состояния: $Y = f(S)$. Нелинейная функция f называется функцией активации нейрона. Одной из наиболее употребимых функций активации является нелинейная функция с насыщением: так называемая логистическая функция, или сигмоид (т. е. функция S-образного вида) [Ежов, Шумский, 1998]:

$$f(X) = \frac{1}{1+e^{-\alpha X}}.$$

При уменьшении параметра α сигмоид становится более пологим, в пределе при $\alpha = 0$ вырождаясь в горизонтальную линию на уровне 0,5, при увеличении α сигмоид приближается по внешнему виду к функции единичного скачка с порогом, равным 1 в точке $X = 0$. Из выражения для сигмоида очевидно, что выходное значение нейрона лежит в диапазоне [0, 1]. Следует отметить, что сигмоидная функция дифференцируема на всей оси абсцисс, что используется в некоторых алгоритмах обучения. Кроме того, она обладает свойством усиливать слабые сигналы лучше, чем сильные, и предотвращает насыщение от больших сигналов.

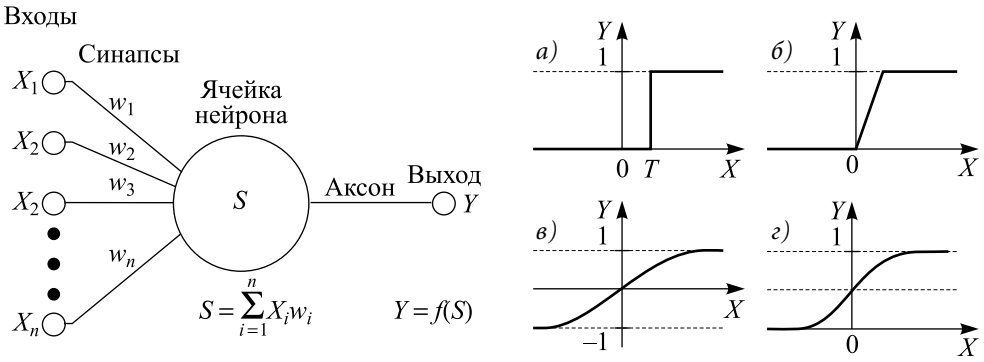

Рис. 1. Схематическое представление нейрона (*слева*) и типы функций активации (*справа*: *а* — функция единичного скачка; *б* — линейный порог (гистерезис); *в* — сигмоид — гиперболический тангенс; *г* — сигмоид

Нейроны, собранные в систему определенной архитектуры, называются нейронной сетью, причем вид архитектуры этой системы опреде-

ляется типами задач, которые планируется решать с помощью сети. Заметим, что любая архитектура имеет входной и выходной слои (данные подаются во входной слой и после обработки сетью получаются из выходного слоя). Обучением нейронной сети называется изменение синапсов (весов каждого нейрона) сети. При этом обучение может быть «с учителем» или «без учителя». В первом случае для каждого входного паттерна сети представляется «эталонный» выходной паттерн, и по разности отклика сети с «эталонным» паттерном происходит подстройка весов. Обучение без учителя предполагает, что сеть самостоятельно формирует свои выходы, адаптируясь к поступающим на ее вход сигналам. «Учителем» сети могут служить лишь сами данные, т. е. имеющаяся в них информация, закономерности, отличающие входные данные от случайного шума. При этом, в отличие от обучения с учителем, решается задача не минимизации целевого функционала (функции ошибки сети), а оптимального кодирования информации в прототипы. Разумеется, как и при любом обучении, необходимо сформулировать правила обучения. Ниже мы описываем три из них: правило обучения Ойя, соревновательное обучение и обучение по Кохонену.

Самоорганизующиеся сети представляют собой тип нейронных сетей, обучающихся без учителя (рис. 2).

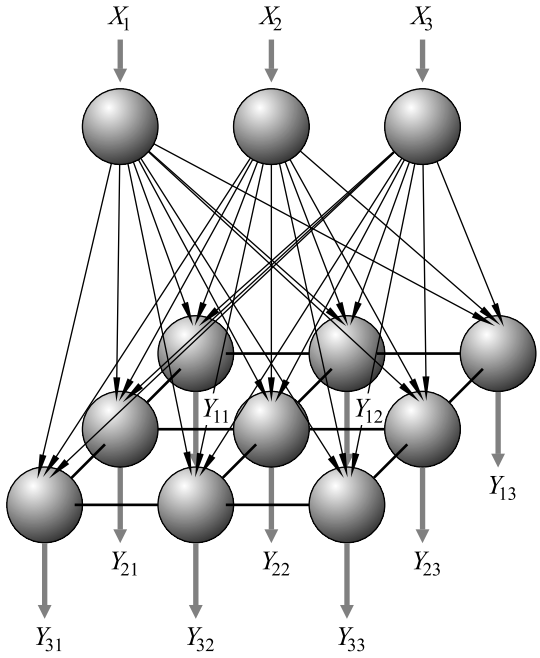

Рис. 2. Схематическое представление самоорганизующейся сети

Остановимся на принципах функционирования и обучения СОК [Kohonen et al., 1996]. Отметим сразу, что размерности входного вектора $\vec{X}$ и вектора, составленного из весов нейронов выходного слоя, совпадают. Это очевидно, если заметить, что каждый нейрон выходного слоя соединен со всеми нейронами входного слоя. Будем обозначать вектор весов i-го нейрона через $\vec{w}_i$.

Для обучения сети используется алгоритм так называемого соревновательного обучения. Опишем кратко этот алгоритм. При таком обучении выходы сети максимально коррелированны: при любом значении входа активность (или длина вектора $\vec{w}_i$) всех нейронов, кроме так называемого нейрона-победителя, одинакова и равна нулю. Такой режим функционирования сети называется «победитель забирает все». Нейрон-победитель (с индексом i^* и весом $\vec{w}_{i*}$) свой для каждого входного вектора $\vec{X}$. Поэтому победитель выбирается так, что его вектор весов $\vec{w}_{i*}$, определенный в d-мерном пространстве входных данных, находится ближе к данному входному вектору $\vec{X}$, чем у всех остальных нейронов: $\left|\vec{w}_{i*} - \vec{X}\right| \le \left|\vec{w}_i - \vec{X}\right|$, $\forall i \ne i^*$, где $|a| = \sqrt{a_1^2 + a_2^2 + ... + a_d^2}$ — обычная евклидова норма d-мерного пространства). Если, как это обычно делается, применять правила обучения нейронов, обеспечивающие одинаковую нормировку всех весов, например, $|w_i| = 1$, то победителем окажется нейрон, дающий наибольший отклик на данный входной «стимул». Выход такого нейрона усиливается до единичного, а остальных — подавляется до нуля.

Количество нейронов в соревновательном слое (в данном случае соревновательный слой совпадает с выходным) определяет максимальное разнообразие выходов и выбирается в соответствии с требуемой степенью детализации входной информации. Обученная сеть может затем классифицировать входы: нейрон-победитель определяет, к какому классу относится данный входной вектор.

В отличие от обучения с учителем, самообучение не предполагает априорного задания структуры классов. Входные векторы должны быть разбиты по категориям (кластерам), согласуясь с внутренними закономерностями самих данных. В этом и состоит задача самоорганизованного обучения соревновательного слоя нейронов.

Согласно правилу Ойа [Oja, Ogawa, Wangviwattana, 1991], $\Delta\vec{w}_i^r = \eta Y_i^r \times$ $\times\left(\vec{X}^r - Y_i^r \vec{w}_i^r\right)$, где $\Delta\vec{w}_i^r$ — изменение веса i-го нейрона при предъявлении ему r-го примера, $\vec{X}^r$ — входной вектор, Y_i^r — отклик i-го нейрона на r-й пример, η — скорость обучения. Это основной алгоритм обучения соревновательного слоя. В согласии с выписанным выше правилом, корректируются лишь веса нейрона-победителя, так как он один имеет ненулевой (единичный) выход. Тогда для победителя правило обучения принимает следующий вид: $\Delta\vec{w}_{i*}^r = \eta(\vec{X}^r - \vec{w}_{i*})$.

В 1982 г. Т. Кохонен [Kohonen, 1982; 1990; 1995] предложил ввести в базовое правило соревновательного обучения информацию о расположении нейронов в выходном слое. Для этого нейроны выходного слоя упорядочиваются, образуя одно- или двумерные решетки, т. е. положение нейронов в такой решетке маркируется двухкомпонентным векторным индексом $\mathbf{i}$. Такое упорядочение естественным образом вводит расстояние между нейронами $|\mathbf{i} - \mathbf{j}|$ в слое. Модифицированное Кохоненым правило соревновательного обучения учитывает расстояние нейронов от нейрона-победителя:

$$\Delta \vec{w}_{\mathbf{i}}^{r} = \eta \Lambda\left(|\mathbf{i} - \mathbf{i}^{*}|\right)\left(\vec{X}^{r} - \vec{w}_{\mathbf{i}}^{r}\right).$$

Функция соседства $\Lambda\left(|\mathbf{i} - \mathbf{i}^{*}|\right)$ равна единице для нейрона-победителя с индексом $\mathbf{i}^{*}$ и постепенно спадает с расстоянием, например, по закону $\Lambda(a) = \exp\left(-a^2/\sigma^2\right)$, где σ называется радиусом обучения. Как темп обучения η, так и радиус обучения нейронов σ постепенно уменьшаются в процессе обучения, так что на конечной стадии обучения мы возвращаемся к базовому правилу Ойя адаптации весов только нейронов-победителей. При такой индивидуальной подстройке прототипов (весов нейронов) обучение по Кохонену напоминает натягивание «эластичной сетки» прототипов на массив данных из обучающей выборки. По мере обучения эластичность сетки постепенно увеличивается, чтобы не мешать окончательной (тонкой) подстройке весов нейронной сети.

Удобным инструментом визуализации данных является раскраска топографических карт аналогично тому, как это делают на обычных географических картах (рис. 3).

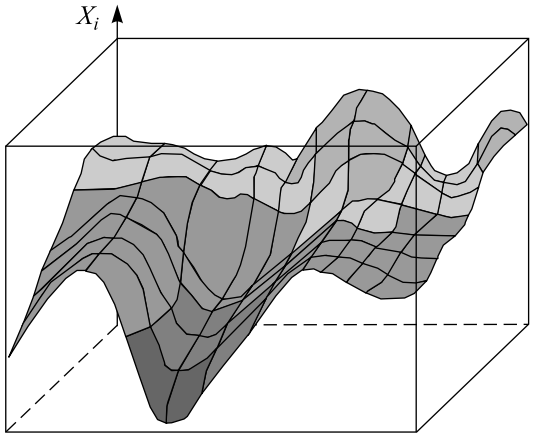

Рис. 3. Раскраска топографической карты, индуцированная *i*-ой компонентой входных данных

И с т о ч н и к: [Ежов, Шумский, 1998, с. 86].

Каждый признак данных порождает свою раскраску ячеек карты — по величине среднего значения этого признака у данных, попавших в данную ячейку. Собрав воедино карты всех интересующих нас признаков, получим топографический атлас, дающий интегральное представление о структуре многомерных данных.

МЕТОД УЛЬТРАМЕТРИЧЕСКИХ ПРОСТРАНСТВ

Одним из современных статистических методов кластеризации большого количества данных является построение иерархических деревьев в ультраметрическом пространстве [Mantegna, Stanley, 2001]. С общей точки зрения, ультраметрические пространства дают наиболее естественный способ описания иерархии любой сложной системы, предположительно имеющей какую-либо структуру. В каком-то смысле концепция ультраметричности есть одна из формализаций понятия иерархии. Более того, таксономия, базирующаяся на обычной евклидовой метрике с неравенством треугольника, требует дополнительной гипотезы относительно топологии пространства n объектов, которые должны быть классифицированы. Такую топологическую гипотезу реализует, в частности, понятие ультраметричности. Необходимые для понимания текста понятия и определения приводятся ниже.

Рассмотрим m временных рядов $S_i(t)$, $0 < i \le m$, $0 < t \le n$, представляющих, например, логарифмические приращения цен акций. Определим коэффициент корреляции между рядами S_i и S_j стандартным образом:

$$\rho_{ij} = \frac{<S_i S_j> - <S_i><S_j>}{\sqrt{<S_i^2> - <S_i>^2 > - <S_j^2> - <S_j>^2 >}},$$

где треугольные скобки обозначают усреднение по t.

Нормируем ряды следующим образом:

$$\hat{S}_i = \frac{S_i - \langle S_i \rangle}{\sqrt{\langle S_i^2 \rangle - \langle S_i \rangle^2}}$$

и будем рассматривать их как векторы $\vec{\hat{S}}_i$ в n-мерном пространстве.

В таком пространстве можно, например, использовать следующие метрики [StatSoft, 2001]:

- евклидова метрика: $d_{ij} = \left\| \vec{\hat{S}}_i - \vec{\hat{S}}_j \right\| = \sqrt{\sum_{k=1}^{n} \left(\hat{S}_{ik} - \hat{S}_{jk} \right)^2}$;
- квадратичная евклидова метрика: $d_{ij} = \left\| \vec{\hat{S}}_i - \vec{\hat{S}}_j \right\|^2 = \sum_{k=1}^{n} \left(\hat{S}_{ik} - \hat{S}_{jk} \right)^2$;

- метрика городских кварталов (манхэттенское расстояние): $d_{ij} = \sum_{k=1}^{n} \left| \hat{S}_{ik} - \hat{S}_{jk} \right|$;
- метрика Чебышева: $d_{ij} = \max_k \left| \hat{S}_{ik} - \hat{S}_{jk} \right|$;
- метрика Пирсона: $d_{ij} = 1 - \rho_{ij}$, где коэффициенты корреляции ρ_{ij} определены приведенной выше формулой.

Эти метрики имеют обычные свойства, удовлетворяющие неравенству треугольника (см. свойство 3):

1) $d_{ij} > 0$; $d_{ij} = 0$, если $i = j$;
2) $d_{ij} = d_{ji}$, $\forall i, j$;
3) $d_{ij} \le d_{ik} + d_{kj}$, $\forall i, j, k$.

Ультраметрическое пространство — это пространство с ультраметрикой $\hat{d}_{ij}$, подчиняющееся следующим свойствам, среди которых неравенство треугольника заменено более сильным свойством (см. свойство 3 ниже) [Mantenga, Stanley, 2001]:

1) $\hat{d}_{ij} > 0$, $\hat{d}_{ij} = 0$, если $i = j$;
2) $\hat{d}_{ij} = \hat{d}_{ji}$;
3) $\hat{d}_{ij} \le \max_k \{\hat{d}_{ik}, \hat{d}_{kj}\}$.

При условии, что в пространстве задана «обычная» метрика, можно построить несколько ультраметрических пространств, кластеризующих набор объектов. Введение ультраметрики позволяет, как было отмечено выше, строить необходимые для кластеризации иерархические деревья. При построении последних важны правила объединения ближайших объектов, т. е. правила формирования кластеров и их соединений. Существует множество типов связи [StatSoft, 2001], мы приведем только три из них.

Одиночная связь, когда расстояние между двумя кластерами определяется расстоянием между двумя наиболее близкими объектами (ближайшими соседями) в различных кластерах.

Полная связь (*метод наиболее удаленных соседей*), если расстояния между кластерами определяются наибольшим расстоянием между любыми двумя объектами в различных кластерах (т. е. «наиболее удаленными соседями»). Этот метод обычно работает очень хорошо, когда объекты происходят на самом деле из различных «рощ». Если же кластеры имеют в некотором роде удлиненную форму или их естественный тип является «цепочечным», то этот метод непригоден.

Невзвешенное попарное среднее. В этом методе расстояние между двумя различными кластерами вычисляется как среднее расстояние между всеми

парами объектов в них. Метод эффективен, когда объекты в действительности формируют различные «рощи», однако он работает одинаково хорошо и в случаях протяженных («цепочного» типа) кластеров.

Отметим, что ультраметрический метод имеет ряд недостатков, связанных с тем, что при выборе различных метрик и типов связи получается различная структура кластеров, поэтому при использовании данного метода необходима дополнительная статистическая обработка данных.

ИССЛЕДОВАНИЕ ИНДЕКСА DOW-JONES INDUSTRIAL AVERAGE

Прежде чем приступить к кластеризации, отобразим ряды S_i на промежуток [0, 1] нелинейным преобразованием: $\tilde{S}_i = f\left(\frac{S_i - \overline{S}_i}{\sigma_i}\right)$, где $f(a) = \frac{1}{1 + e^{-a}}$, $\overline{S}_i = \frac{1}{P}\sum_{j=1}^{P} S_i^j$ — математическое ожидание и $\sigma_i = \frac{1}{P-1}\sum_{j=1}^{P}(S_i^j - \overline{S}_i)^2$ — среднеквадратичное отклонение (волатильность) i-го актива. О преимуществах такой нормировки по сравнению со стандартной линейной можно прочитать в книге [Ежов, Шумский, 1998].

Мы исследовали недельные котировки акций входящих в индекс DJIA, за периоды с 10.01.1994 по 27.10.1997 и с 10.11.1997 по 27.08.2001[1]. Данный индекс включает в себя акции 30 компаний, следовательно, количество различных коэффициентов корреляции ρ_{ij} равно $(30 \times 29)/2 = 435$. В таблице 1 представлены максимальные и минимальные значения ρ_{ij} за каждый исследуемый период. Наибольший коэффициент корреляции за период с 10.01.1994 по 27.10.1997 соответствует компаниям JP Morgan Chase и American Express, а с 10.11.1997 по 27.08.2001 — компаниям JP Morgan Chase и Citigroup. Все три компании занимаются предоставлением финансовых услуг. Соответствующие графики, дающие визуальное представление о скоррелированности логарифмов цен акций этих компаний, приведены на рис. 4 (а, б).

Таблица 1

Максимальные и минимальные значения коэффициентов корреляции компаний из индекса DJIA за исследуемые периоды

Временной период	$\mathbf{Min}\,\rho_{ij}$	$\mathbf{Max}\,\rho_{ij}$
с 10.01.1994 по 27.10.1997	−0,04	0,61
с 10.11.1997 по 27.08.2001	−0,06	0,72

Дополнительную информацию о поведении матрицы коэффициентов корреляции во времени можно получить, изучив эмпирическую функцию

[1] Данные взяты с сайта http://finance.yahoo.com

плотности распределения вероятности $P(\rho_{ij})$ полного набора 435 коэффициентов корреляции. Известно, что для большинства финансовых временных рядов $P(\rho_{ij})$ имеет форму колокола, причем среднее значение $P(\rho_{ij})$ слабо зависит от времени, тогда как среднеквадратичное отклонение почти не меняется [Mantegna, 1997].

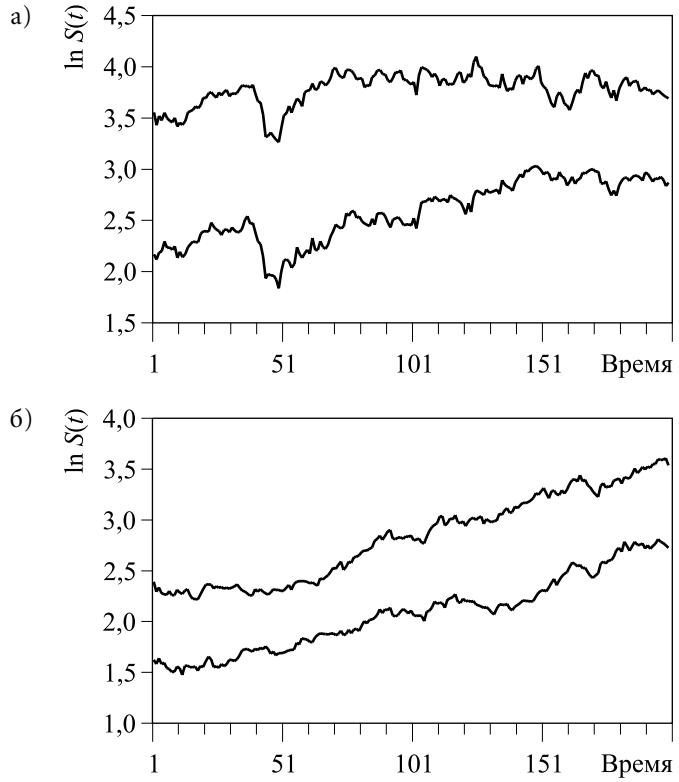

Рис. 4. а) Изменение во времени ln $S(t)$ компаний JP Morgan Chase (верхний график) и American Express (нижний график) с 10.01.1994 по 27.10.1997;
б) Изменение во времени ln $S(t)$ компаний JP Morgan Chase (верхний график) и Citigroup (нижний график) с 10.11.1997 по 27.08.2001

На рис. 5 изображены иерархические деревья для акций 30 компаний, входивших в индекс DJIA за периоды с 10.01.1994 по 27.10.1997 и с 10.11.1997 по 27.08.2001. Иерархические деревья строились на основе метрики Пирсона при использовании невзвешенной попарной средней связи между кластерами. Отчетливо видна кластерная структура данных. В частности, в один кластер попали компании, предоставляющие финансовые услуги своим клиентам, — AXP, C, JPM, производители компьютерной техники и про-

граммного обеспечения — IBM, INTC, HWP, MSFT, компании CAT (машиностроение), IP (производство бумаги и упаковочных материалов), AA (металлургия), и компании PG (косметическая промышленность), MRK (производство фармацевтических препаратов), KO (производство продуктов питания), JNJ (косметическая промышленность). Заметим, что структура кластеров, образованных этими компаниями, сохраняется во времени. Также видно, что телекоммуникационные компании T (левая часть рис. 5) и SBC (правая часть) находятся далеко от остальных компаний и, следовательно, слабо коррелируют с ними.

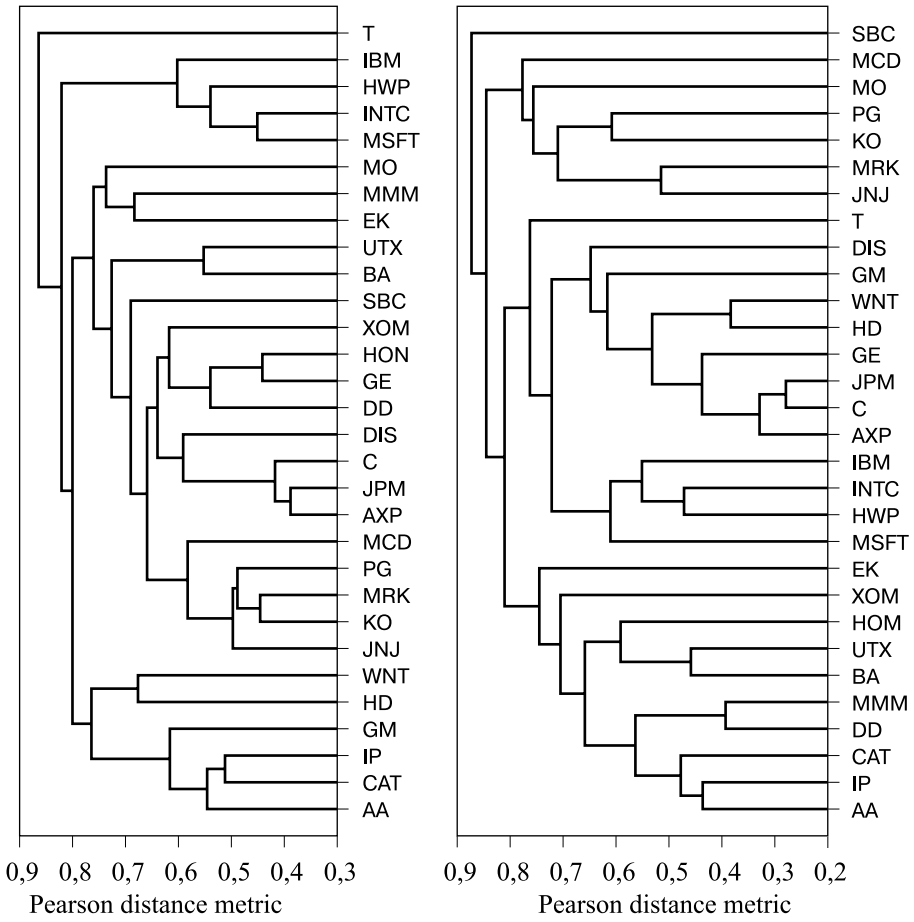

Рис. 5. Иерархические деревья для 30 компаний, формирующих индекс DJIA. С 10. 01. 1994 по 27. 10. 1997 (*слева*) и с 10.11.1997 по 27.08.2001 (*справа*)

Таблица 2 содержит параметры обучения СОК. За оба исследуемых периода параметры одинаковые. Процесс обучения делится на грубую и точ-

ную подстройку. На этапе грубой подстройки происходит грубая кластеризация. Для этого этапа характерна достаточно большая коррекция синаптических весов нейросети по прохождению каждой эпохи обучения. Под последней понимается однократное предъявление ей всех примеров из обучающей выборки. На этапе точной подгонки величина коррекции синаптических весов значительно уменьшается. Следовательно, радиус обучения σ в начале должен быть соизмерим с размерами карты, а в конце — достаточно малым. Сперва он может принимать значение от половины меньшего значения размера карты до этого значения. Чем больше радиус обучения, тем медленнее обучается СОК, так как приходится корректировать большое количество нейронов.

Таблица 2

Параметры обучения СОК для компаний, формирующих DJIA (с 10.01.1994 по 27.10.1997 и с 10.11.1997 по 27.08.2001)

Скорость обучения изменяется в зависимости от размеров карты и количества эпох обучения. Мы использовали в качестве изменения модификации скорости обучения обратно-пропорциональную зависимость от числа обучающих примеров, так как это дает наиболее плавное изменение синаптических весов нейронов. Инициализация синаптических весов реализовалась случайными значениями. На рис. 6 изображена кластерная структура индекса DJIA, построенная методом СОК. Все пространство компаний разделилось на группы векторов, расстояние между которыми внутри каждой группы наименьшее с учетом выбора масштаба кластеризации.

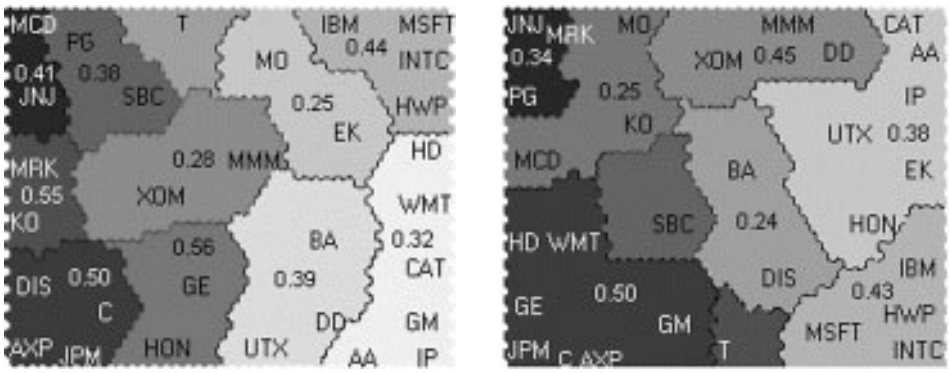

Рис. 6. Кластерная структура СОК для индекса DJIA за периоды с 10.01.1994 по 27.10.1997(*слева*) и с 10.11.1997 по 27.08.2001 (*справа*)

Заметим, что чем ближе по «цвету» кластеры, тем больше коэффициент корреляции между ценами акций компаний, принадлежащих этим кластерам, и наоборот. В таблицах 3 и 4 представлены корреляции между акциями компаний, попавших в самый светлый и самый темный кластеры.

Таблица 3

Матрица корреляций для компаний, попавших в два самых удаленных кластера за период с 10.01.1994 по 27.10.1997

	AA	CAT	GM	HD	IP	WMT
JNJ	– 0,03	0,16	0,14	0,19	0,07	0,16
MCD	0,04	0,20	0,04	0,17	0,05	0,07

Таблица 4

Матрица корреляций для компаний, попавших в два самых удаленных кластера за период с 10.11.1997 по 27.08.2001

	AA	CAT	EK	HON	IP	UTX
JNJ	0,07	0,17	0,04	– 0,04	0,07	0,21
MRK	– 0,01	0,14	0,02	0,13	0,10	0,25
PG	0,06	0,16	0,08	0,14	0,07	0,28

Приведенная выше статистика показывает, что метод ультраметрических пространств и метод СОК вполне согласуются на данных из выборки DJIA.

ИССЛЕДОВАНИЕ ИНДЕКСА NASDAQ 100

Индекс NASDAQ 100 включает в себя акции 100 компаний[2], следовательно, количество различных коэффициентов корреляции ρ_{ij} равно $(100 \times 99)/2 = 4950$. Исследования проводились на дневных ценовых данных за период с 14.03.2001 по 31.12.2001. Наибольший коэффициент корреляции $\rho^{MAX} = 0{,}93$ соответствует компаниям, разрабатывающим микропроцессоры — Novellus Systems, Inc. (NVLS) и KLA-Tencor Corporation (KLAC) (графики логарифмов цен акций этих компаний изображены на рис. 7).

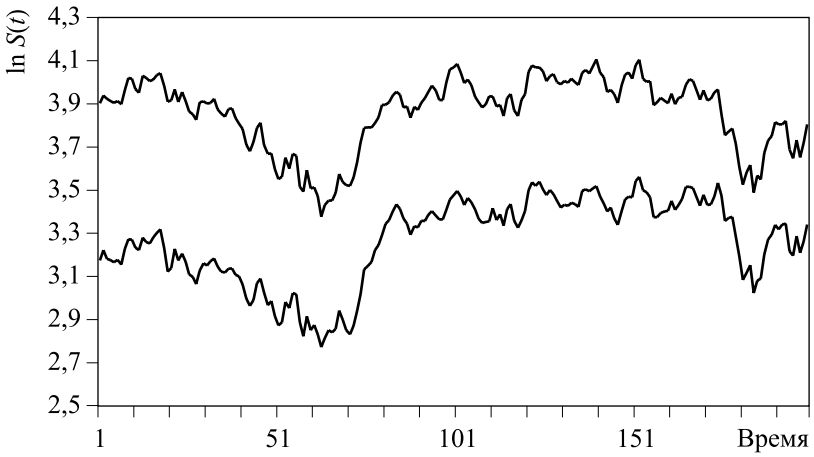

Рис. 7. Изменение во времени ln $S(t)$ компаний KLAC (верхний график) и NVLS (нижний график) с 14.03.2001 по 31.12.2001

На рис. 8 изображено иерархическое дерево, полученное методом ультраметрических пространств, для 100 компаний, входящих в индекс NASDAQ 100.

Параметры обучения СОК для компаний, формирующих NASDAQ 100 за период с 14.03.2001 по 31.12.2001, представлены в табл. 5.

На рис. 9 изображена кластерная структура СОК для компаний, формирующих NASDAQ 100 с 14.03.2001 по 31.12.2001. Цифры на каждом кластере соответствуют среднему коэффициенту корреляции между курсами акций компаний, попавших в один кластер. Видно, что СОК поделила все пространство на кластеры с довольно высокими корреляциями, чего нельзя сказать об ультраметрическом дереве, где кластеры имеют размазанную структуру. Также видно, что максимально коррелированные компании KLAC и NVLS из всего исследуемого набора заняли одну ячейку на карте (верхний правый угол).

[2] Подробную информацию о компаниях можно получить на сайте: http://finance.yahoo.com

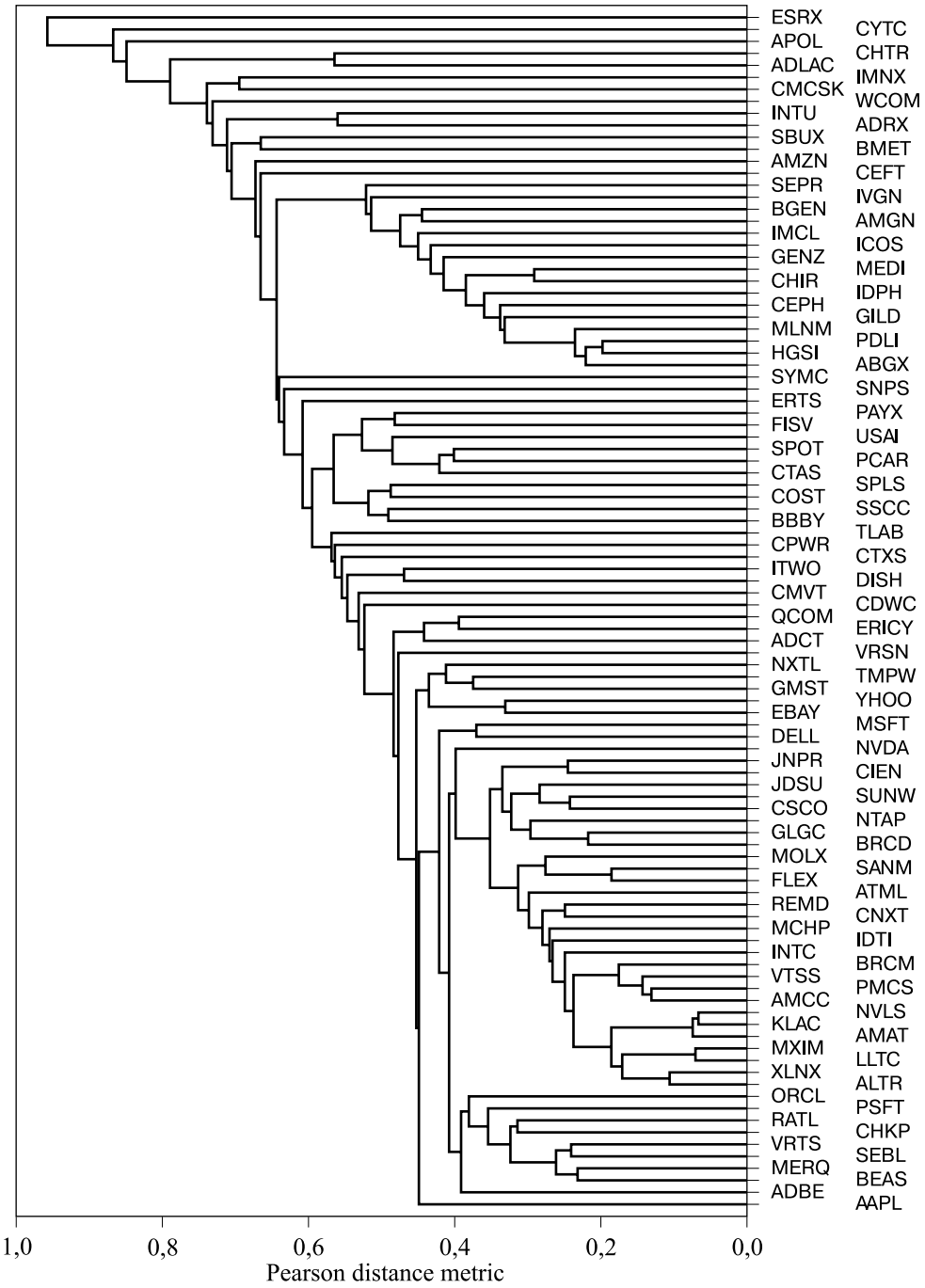

Рис. 8. Иерархическое дерево для индекса NASDAQ 100 за период *с* 14.03.2001 по 31.12.2001

Таблица 5

Параметры обучения СОК для компаний, формирующих NASDAQ 100, с 14.03.2001 по 31.12.2001

Параметр	Значение
Размер карты	40x40 нейронов
Форма ячеек	Шестиугольники
Количество эпох при грубой подстройке	100 000
Количество эпох при точной подстройке	100 000
Скорость обучения при грубой подстройке	0,3
Скорость обучения при точной подстройке	0,05
Начальный радиус обучения	20
Конечный радиус обучения	1
Модификация скорости обучения	Обратно-пропорциональная
Начальная инициализация весов	Случайными значениями
Время обучения	4 ч. 10 мин.

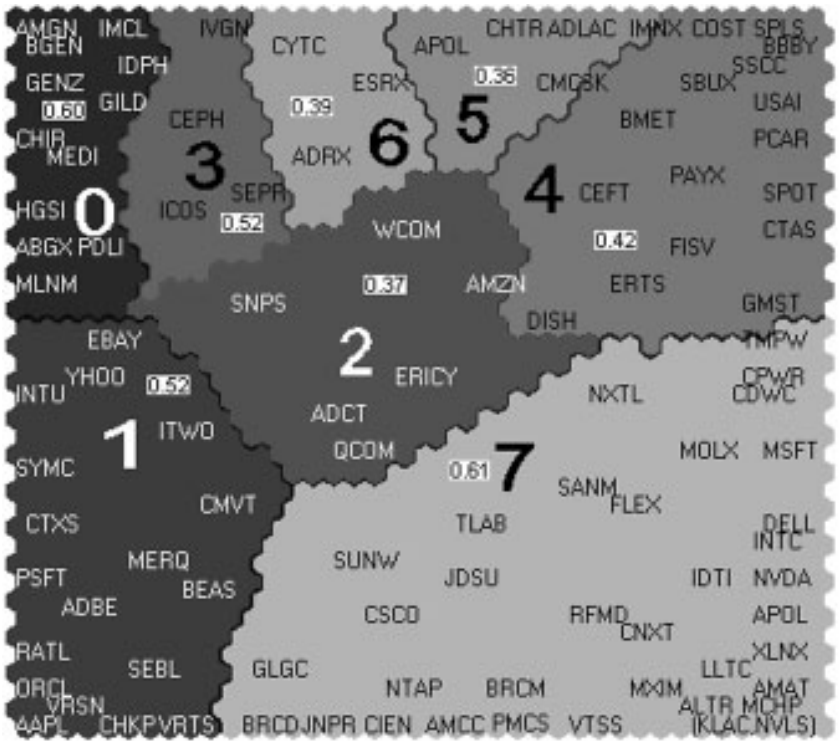

Рис. 9. Кластерная структура СОК для индекса NASDAQ100 за период с 14.03.2001 по 31.12.2001

Нами было подсчитано, что среднее значение и среднеквадратичное отклонение матрицы корреляций компаний формирующих NASDAQ 100 равно 0,47 и 0,18 соответственно. Среднее же значение матрицы корреляций между курсами акций компаний, попавших в самый светлый и самый темный кластеры, равно 0,31. Это свидетельствует о том, что СОК успешно справилась с задачей кластеризации.

ЗАКЛЮЧЕНИЕ

В данной работе изучалась возможность использования самоорганизующихся карт в качестве альтернативы традиционным моделям таксономии массивов данных. Был представлен обзор модели самоорганизующейся нейронной сети, обучаемой без учителя. В результате получена кластеризация индексов DJIA и NASDAQ 100 с помощью ультраметрических деревьев и самоорганизующихся карт Кохонена. В исследовании было показано, что:

1) в случае DJIA результаты обоих методов оказались вполне согласованными. Значит, на небольших выборках данных ультраметрические методы практически не дают ошибки кластеризации;
2) при кластеризации NASDAQ 100 выяснилось: результаты метода СОК оказались более понятными и естественными, нежели результаты ультраметрических методов, что свидетельствует об ограниченной применимости обычных статистических методов в приложении к кластеризации больших, слабо структурированных наборов данных.

Специальное внимание уделено недостаткам метода иерархических деревьев. К недостаткам СОК можно отнести, разве что, значительное время, необходимое для обучения (4 ч. 10 мин.). Вместе с тем при выбранной в работе временной нарезке данных (дневные цены акций), метод СОК можно считать on-line методом обработки данных. Наконец, время обучения СОК можно существенно сократить благодаря использованию более высокопроизводительных компьютеров, нежели ПК, на котором проводилось обучение СОК в настоящей работе.

Мы показали, что метод СОК более перспективен в применении к задачам по нахождению неизвестных заранее структур в больших массивах данных, или, в более общем виде, извлечению новых знаний из имеющейся информации. Не требует объяснений применимость технологии СОК к решению самых разнообразных задач оперативного и стратегического менеджмента, чем должен быть занят финансовый менеджер, формирующий и управляющий диверсифицированным портфелем ценных бумаг, для которого наибольший интерес представляют наименее коррелированные наборы активов.

В продолжение работы планируется создание интеллектуальной автоматической системы принятия торговых или инвестиционных решений на основе управления динамическим инвестиционным портфелем, состав которого меняется во времени. Формирование такого портфеля подразумевает кластеризацию активов по большему числу параметров. В качестве основных из них можно выбрать доходность актива, объемы торговли, открытый интерес и т. д. По этим параметрам будет проведена оптимизация стратегии управления портфелем ценных бумаг. Подобные интеллектуальные технологии работы с данными могут представлять интерес для финансовых менеджеров инвестиционных компаний, банков, паевых инвестиционных фондов и дилинговых компаний.

Литература

Дебок Г., Кохонен Т. Анализ финансовых данных с помощью самоорганизующихся карт. М.: Альпина Паблишер, 2001.

Ежов А. А., Шумский С. А. Нейрокомпьютинг и его применения в экономике и бизнесе. (сер. «Учебники экономико-аналитического института МИФИ») / Под ред. проф. В. В. Харитонова. М.: МИФИ, 1998.

Back B., Sere K., Vanharanta H. Data Mining Accounting Numbers Using Self Organizing Maps // Proceedings of STeP'96, Finnish Artificial Intelligence Conference / Eds. J. Alander, T. Honkela, M. Jakobsson. Finnish Artificial Intelligence Society: Vaasa, Finland, 1996. P. 35–47.

Carlson E. Self-Organizing Feature Maps for Appraisal of Land Value of Shore Parcels // Artificial Neural Networks. Proceedings of ICANN'91, International Conference on Artificial Neural Networks. Vol. II. North-Holland, Amsterdam, 1991. P. 1309–1312.

Cheng G., Liu X., Wu J. X. Interactive Knowledge Discovery through Self-Organizing Feature Maps // Proceedings of WCNN'94. World Congress on Neural Networks. Vol. IV. Lawrence Erlbaum: Hillsdale, NJ, 1994. P. 430–434.

Demartines P. Analyse de Données par Réseaux de Neurons Auto-Organisés (Data Analysis through Self-Organized Neural Networks). PhD Thesis. Institut National Polytechnique de Grenoble: Grenoble, France, 1994.

Garavaglia S. A Self-Organizing Map Applied to Macro and Microanalysis of Data with Dummy Variables // Proceedings of WCNN'93, World Congress on Neural Networks. Lawrence Erlbaum and INNS Press: Hillsdale, NJ, 1993. P. 362–368.

Kohonen T. Self-Organized Formation of Topologically Correct Feature Maps // Biological Cybernetics. 1982. Vol. 43. P. 59–69.

Kohonen T. The Self-Organizing Map // Proceeding of the IEEE. 1990. Vol. 78. P. 1464–1480.

Kohonen T. The Self-Organizing Maps. Springer: Berlin, 1995.

Kohonen T., Oja E., Simula O., Visa A., Kangas J. Engineering Applications of the Self-Organizing Map // Proceedings of IEEE. 1996. Vol. 84. P. 1358–1389.

Mantegna R. N., Degree of Correlation Inside a Financial Market., in Applied Nonlinear Dynamics and Stochastic Systems Near the Millennium / Eds. P. 362–368. J. B. Kadtke, A. Bulsara. AIP Press: N. Y., 1997. P. 197–202.

Mantegna R. N., Stanley H. E. An Introduction to Econophysics. Correlation and Complexity in Finance. Cambridge University Press: Cambridge, 2001.

Martнn-del-Brнo B., Serrano-Cinca C. Self-Organizing Neural Networks for the Analysis and Representation of Data: Some Financial Cases // Neural Computing & Applications. 1993. Vol. 1. P. 193–206.

Marttinen K. SOM in Statistical Analysis: Supermarket Customer Profiling // Proceedings of the Symposium on Neural Network Research in Finland / Eds. A. Bulsari, B. Saxén. Finnish Artificial Intelligence Society: Turku, Finland, 1993. P. 75–80.

Oja E., Ogawa H., Wangviwattana J., Learning in Nonlinear Constrained Hebbian Network // Artificial Neural Networks / Eds. Kohonen T. et al. North-Holland, Amsterdam. 1991. P. 385-390.

Serrano-Cinca C. Self-Organizing Neural Networks for Financial Diagnosis // Decision Support Systems. 1996. Vol. 17. July. P. 227-238.

StatSoft. Электронный учебник по промышленной статистике. М., 2001 [Электронный ресурс]. — Режим доступа: http://www.statsoft.ru/home/portal/textbook_ind/default.htm

Tryon R. C. Cluster Analysis. Ann Arbor: Edwards Brothers, 1939.

Ultsch A. Self-Organizing Neural Networks for Visualization and Classification // Information and Classification / Eds. O. Opitz, B. Lausen, R. Klar. Springer-Verlag: Berlin, 1993. P. 307–313.

Статья поступила в редакцию 24 ноября 2004 г.