УДК 004.8

И. Д. ПОЛОСУХИН, студент НТУ «ХПИ»

ДИНАМИЧЕСКАЯ КЛАСТЕРИЗАЦИЯ ВРЕМЕННЫХ РЯДОВ С ИСПОЛЬЗОВАНИЕМ АГРЕГИРОВАННЫХ ПОКАЗАТЕЛЕЙ

В статті розглядається задача кластеризації часових рядів стосовно котирування акцій. У роботі були використані: метод отримування головних компонент «Гусениця» і коефіцієнт Херста для отримання параметрів ряду; метод *k*-середнього та евклідова відстань для кластерізація.

В статье рассматривается задача кластеризации временных рядов применительно к котировкам акций. В работе были использованы: метод получения главных компонент «Гусеница» и коэффициент Хёрста для подсчета параметров ряда; метод *k*-среднего и евклидово расстояние для кластеризации.

In this paper was examined a problem of time series based on stocks market history prices. For this purpose was used: method of Singular Spectrum Analysis and Hurst exponent for parameters calculation; *k*-mean clustering and Euclidean distance for clustering.

Введение. Успешное решение задачи прогнозирования рядов в значительной мере определяется соответствием выбранной модели истинной структуре ряда. Задача выбора вида модели не имеет формального решения и в значительной мере опирается на эвристические соображения в сочетании со статистическими методами оценивания параметров модели и последующей проверкой ее адекватности. Очевидно, что решение указанной задачи существенно облегчается, если предварительно сгруппировать исследуемые временные ряды в группы, содержащие ряды, в определенном смысле близкие по структуре. Для решения этой вспомогательной задачи можно использовать известные методы кластеризации, однако при этом возникает проблема выбора обоснованных показателей «близости» временных рядов на основе вычисляемых статистических характеристик.

Проблема кластеризации временных рядов. Различные походы к решению задачи кластеризации временных рядов рассматривались в ряде работ [1, 2]. При этом для решении задачи использовались такие методы кластеризации, как перегруппированная кластеризация (*relocation clustering*), агломеративная иерархическая кластеризация, метод *k*-среднего, метод нечеткого *c*-среднее и другие.

В [1] рассмотрены вопросы классификация временных рядов цен на акции по индустриальным категориям, таким как Media, IT, и др. и проведен анализ движения цен акций между различными категориями. При этом использовались следующие показатели:

- результаты усреднения ряда на недельной основе;
- процентные приросты цен на акции в определенные моменты времени;
- нормализованные значения процентных приростов цен акций;

Для решения задачи использовалась иерархическая конгломеративная кластеризация с функциями стоимости *Single link* (*min*), *Complete link* (*max*), *average link*, *ward's method*. В качестве метрики была выбрано Евклидовое расстояние.

Основные подходы к задаче кластеризации временных рядов рассмотрены в [2]. Таким образом, основная проблема состоит в выборе системы показателей временного ряда, обеспечивающих формирование критериев их подобия (близости) и соответствующего метода кластеризации.

Выбор перечня показателей. Для решения задачи кластеризации были в работе выбраны следующие показатели временного ряда $X = \{x_i\}_{i=1}^{N}$:

- Показатель «математическое ожидание» – $M_x = \frac{1}{N}\sum_{i=1}^{N} x_i$.
- Показатель «среднее квадратичное отклонение» – $Std_x = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - M_x)^2}$.
- Показатель «тренд» – направленность роста временного ряда:

$$P^T = \begin{pmatrix} 1 & 1 & \dots & 1 \\ 1 & 2 & \dots & N \end{pmatrix}, \Theta = \left(P^T P\right)^{-1} P^T X, T_x = \Theta_2.$$

- Показатель, основанный на использовании коэффициентов разложения ряда по методу главных компонент. Для расчетов целесообразно воспользоваться методов «Гусеницы» (Singular Spectrum Analysis)» [3]. Расчеты производятся на основе использования так называемой «траекторной» матрицы, вычисленной на интервале времени $K = N - M + 1$, где $M \le \frac{N}{2}$:

$$P = \begin{pmatrix} x_1 & x_2 & \dots & x_K \\ x_2 & x_3 & \dots & x_{K+1} \\ \dots & \dots & \dots & \dots \\ x_M & x_{M+1} & \dots & x_N \end{pmatrix}, V = \frac{1}{K} P^T P. \quad (1)$$

Далее выполняется сингулярное разложение матрицы V: $V = U\Sigma V$, где Σ_x – диагональная матрица размера $M \times K$ с неотрицательными вещественными числами по диагонали. Эти числа и используются в качестве коэффициентов разложения по методу «Гусеницы».

- Показатель Хёрста [4] – показатель стохастичности ряда, позволяющий оценить, является ли ряд стохастичным, белым шумом или же имеет место наличие тренда. Показатель определяется из соотношения:

$$M\left[\frac{R(n)}{S(n)}\right] = Cn^H, n \longrightarrow \infty, \quad (2)$$

где $\frac{R(n)}{S(n)}$ – нормированный диапазон (rescaled range), C – константа, H – показатель Хёрста. В свою очередь нормированный диапазон можно вычислить следующим способом:

$$R(n) = \max(Z_1, Z_2, ..., Z_n) - \min(Z_1, Z_2, ..., Z_n), \; S(n) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - M_x)^2}, \quad (3)$$

где $y_t = x_t - M_x$, $z_t = \sum_{i=1}^{t} y_i$.

Для получения оценки $M\left[\frac{R(n)}{S(n)}\right]$, нужно усреднить $\frac{R(i)}{S(i)}$ для всех $i = 1, 2.., n$. Параметр Херста в свою очередь оценивается с использованием уравнения линейной регрессии, полученного путем логарифмирования (5):

$$\log(M\left[\frac{R(n)}{M(n)}\right]) = H\log(n) + \log(C). \quad (4)$$

Окончательно выражение для искомой оценки приобретает вид:

$$P^T = \begin{bmatrix} 1 & 1 & ... & 1 \\ \log(1) & \log(2) & ... & \log(n) \end{bmatrix}, Y^T = \left[\log(M\left[\frac{R(1)}{M(1)}\right]) \;\; \log(M\left[\frac{R(2)}{M(2)}\right]) \;\; ... \;\; \log(M\left[\frac{R(N)}{M(N)}\right])\right]; \quad (5)$$

$$\Theta = \left(P^T P\right)^{-1} P^T Y \;\; H_x = \Theta_2. \quad (6)$$

Выбор алгоритма кластеризации и метрики. В работе был использован известный метод кластеризации *k*-среднее. В качестве метрики было выбрано Евклидово расстояние в пространстве оцениваемых показателей.

Соответственно можно ввести показатель расстояния между двумя рядами:

$$P(X^i, X^j) = \sqrt{\sum_k (p_k^i - p_k^i)^2}, \quad (7)$$

где p_k^i – k-й показатель i-го временного ряда.

Вычислительный эксперимент. Для вычислительного эксперимента были использованы временные ряды цен на акции на бирже NASDAQ за последние 5 лет. Был проведен эксперимент, на 500 временных рядах, которые разбивались на 10 групп. В приведенной ниже таблице, указаны среднее расстояние, заданное формулой (7), между временными рядами внутри одной группы.

Как видно из таблицы, группы 1–6 имеют сравнительно малое среднее расстояния между временными рядами. Группы 7–10 в таблице не приведены, так как среднее расстояние в них на порядок больше, что означает, что эти группы содержат выбросы. Подсчёт среднего расстояния между центрами кластеров показал, что они удалены друг от друга, так как расстояние превышает 10^5.

Таблица полученных групп

	Номер группы					
	1	2	3	4	5	6
Количество рядов	88	46	81	33	161	65
Среднее Р в группе	2232	7150	4392	9293	1426	3101

Заключение. В перспективе, для лучшей кластеризации, можно использовать другие методы. Например, *c*-среднее – нечеткая кластеризация, которая позволит определить с какой степенью тот или иной ряд относится к какому-то кластеру. Целесообразно также использовать методы иерархической кластеризации, которые не требуют задания исходного числа кластеров, а позволяют найти их в процессе выполнения процедуры кластеризации. Так же для процедуры *c*-среднего можно использовать метрику относительного расстояния, вычисленную через корреляционный коэффициент Пирсона.

Представляет интерес так же использование других показателей, таких как коэффициенты разложения на прототипы функций (вейвлеты), коэффициенты разложения Фурье и другие возможные разложения, что, возможно, позволит повысить точность кластеризации.

Список литературы: 1. *Todd Wittman.* Time-Series Clustering and Association Analysis of Financial Data [Электронный ресурс] : сайт математического факультета Университета Калифорнии – Режим доступа: http://www.math.ucla.edu/~wittman/thesis/project.pdf. **2.** *T.Warren Liao*. Clustering of time series data — a survey. [Электронный ресурс]: архив статей Университета Пенсильвании – Режим доступа: http:// citeseerx.ist.psu.edu /viewdoc /download ?doi=10.1.1.115.6594 &rep=rep1 &type=pdf. **3.** Метод «Гусеница» [Электронный ресурс]: сайт об методе «Гусеница» – Режим доступа: http://www.gistatgroup.com/gus/ **4.** Показатель Хёрста [Электронный ресурс] : международная интернет энциклопедия – Режим доступа: http://en.wikipedia.org/wiki/Hurst_exponent.

Надійшла до редколегії 16.06.2011