АНАЛИЗ ДИНАМИЧЕСКИХ ХАРАКТЕРИСТИК СОСТОЯНИЯ РЫНКА ЦЕННЫХ БУМАГ

А. А. Мусаев[1], И. А. Барласов[2]

[1]Санкт-Петербургский институт информатики и автоматизации РАН, [2]НОУ «Международный Банковский Институт»

[1]СПИИРАН, 14-я линия ВО, д. 39, Санкт-Петербург, 199178; [2]НОУ «МБИ», Невский пр., д., 58, Санкт-Петербург, 191023

[1]<amusaev@szma.com>, [2]<barlasov@yandex.ru>

УДК 681.3.01

Мусаев А. А., Барласов И. А. **Анализ динамических характеристик состояния рынка ценных бумаг** // Труды СПИИРАН. Вып. 6. — СПб.: Наука, 2008.

Аннотация. *Рассматривается проблема построения автоматизированной системы анализа свойств информационных потоков как первоначального этапа применения алгоритмов интеллектуального анализа данных (Data Mining, DM), лежащих в основе нового аналитического типа автоматизированного управления многомерными динамическими процессами. Применение указанных алгоритмов требует проведения статистического и динамического анализа данных, формирующего представления о структуре исходных информационных потоков.* — Библ. 19 назв.

UDC 681.3.01

Musayev A. A., Barlasov I. A. **Dynamic parameters analyses of the stock market state** // SPIIRAS Proceedings. Issue 6. — SPb.: Nauka, 2008.

Abstract. *The article is devoted to automatic analyzing system of information as the first stage of DM (Data Mining) algorithms which are the basis of new type of automatic control of multivariate dynamic processes. It is unavailable to use these algorithms without statistic and dynamic analyses forming the information flows structure.* — Bibl. 19 items.

1. Введение

Современное состояние систем анализа рядов наблюдений за изменением инвестиционной ситуации (товарного, валютного и фондового рынков, отраслей промышленности и ведущих корпораций, параметров экономической и политической среды погружения и т.д.) характеризуется сложившимися противоречиями:

- между огромным объемом массивов данных, формируемых в процессе мониторинга инвестиционной ситуации, и возможностью человеческого мозга по их восприятию и аналоговой (качественной) переработке в интересах выработки управляющих решений;
- между возможностями современной прикладной (компьютерной) математики и крайне низким уровнем ее применения в интересах количественного анализа инвестиционной ситуации с целью совершенствования технологических управлений.

Разрешение указанных противоречий ведет к актуализации проблемы создания математического и программного инструментария, обеспечивающего возможность извлечения знаний из массивов оперативной и ретроспективной информации, накопленной в процессе автоматического мониторинга состояния инвестиционной ситуации. Решение данной проблемы привело к возникновению аналитических информационных технологий (Data Mining).

Однако непосредственное применение продуктов Data Mining к массивам накопленной ретроспективной информации, как правило, не обеспечивает по-

лучение достоверных результатов. В большинстве случаев это связано с тем, что те или иные алгоритмы прогнозирования и принятия решений базируются на математических методах, применение которых ограничено набором условий. Данные условия обычно относятся к статистическим и динамическим свойствам исходных данных. При невыполнении этих условий соответствующие прикладные алгоритмы в лучшем случае не обеспечивают формирование оптимальных решений, а в некоторых обстоятельствах, вообще, приводят к ложным выводам.

Таким образом, применению формализованных методов прогнозирования и принятия решений должен предшествовать комплексный анализ структуры и свойств информационных потоков, отображающих изменений состояния инвестиционной ситуации. Отсюда непосредственно вытекает задача разработки программно-алгоритмических средств, обеспечивающих проведение указанного автоматизированного анализа информационных потоков. Результаты анализа должны обеспечить возможность содержательной интерпретации состояния инвестиционного климата, а также графическую и текстовую визуализацию их свойств.

С точки зрения инвестиционного анализа, речь идет о выявлении структуры данных, позволяющем обнаруживать искажения информационных потоков и устранить (или снизить) влияние этих деформаций на формирование математических моделей, лежащих в основе алгоритмов прогнозирования и формирования инвестиционных решений.

Базовый вариант построения системы анализа свойств информационных потоков включает в себя решение следующих задач:

1. Дескриптивный статистический анализ наблюдаемого параметра;
2. Анализ динамических характеристик наблюдаемого параметра;
3. Выявление значимых взаимосвязей для наблюдаемого параметра;
4. Идентификация формы (характера) взаимосвязей между любыми парами наблюдаемых параметров;
5. Выявление аномальных наблюдений;
6. Корреляционный анализ групп наблюдаемых параметров;
7. Обобщенное (агрегированное) представление групп наблюдений и их визуализация;
8. Выявление и анализ несоответствий в группах наблюдаемых параметров.

Перечисленные задачи образуют базовые функциональности анализатора, функциональная структура которого приведена на рис. 1. В настоящей статье рассмотрены лишь первые две функциональности из перечисленного списка. Решение остальных задач предполагается привести в последующих выпусках трудов СПИИРАН.

2. Дескриптивный статистический анализ исходных данных. Проблема устойчивости статистических моделей

Пусть x_n - выборка наблюдений случайной величины X с некоторым, в общем случае неизвестным распределением $f(x)$. В качестве выборочных характеристик этого распределения в процессе дескриптивного анализа обычно используют:

Труды СПИИРАН. 2008. Вып. 6. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2008. Issue 6. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

– выборочное среднее $x = \frac{1}{n}\sum_{i=1}^{n} x_i$,

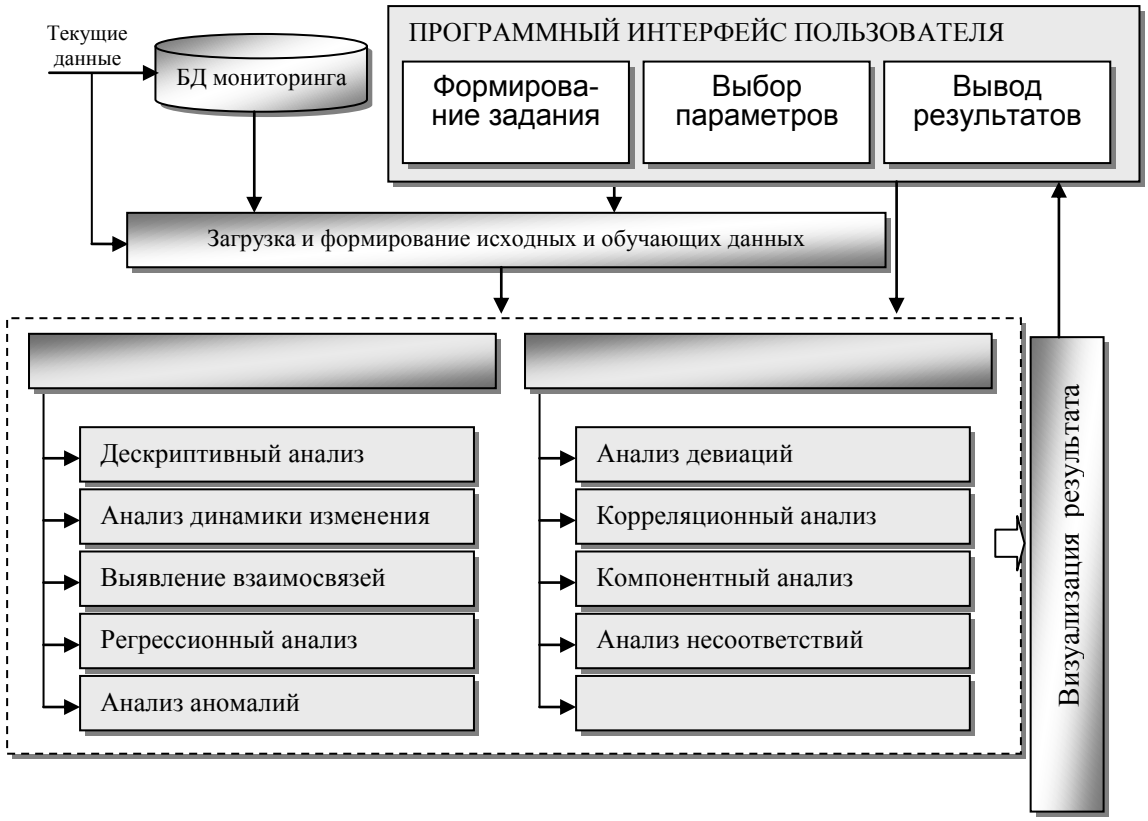

Рис. 1. Функциональная структура анализатора информационных потоков.

– выборочную дисперсию $D(x) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$;

– среднеквадратическое отклонение $\sigma = \sqrt{D(x)}$;

– коэффициент вариации $V = \frac{\sigma}{\bar{x}}$;

– коэффициент асимметрии $\mathrm{As} = \frac{1}{n}(\sum_{i=1}^{n}(x_i - \bar{x})^3)/\sigma^3$;

– коэффициент эксцесса $\mathrm{Ex} = \frac{1}{n}(\sum_{i=1}^{n}(x_i - \bar{x})^4)/\sigma^4 - 3$.

Если число членов невелико ($n < 30$), то при вычислении σ^2 в знаменателе используется $(N-1)$ и уточняются значения коэффициентов асимметрии эксцесса:

$$\mathrm{As}_1 = \frac{n-1}{n-2}\,\mathrm{As};$$

Труды СПИИРАН. 2008. Вып. 6. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2008. Issue 6. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

$$\mathrm{Ex}_1 = \frac{\sqrt{n(n-1)}}{(n-2)(n-3)}[(n+1)\cdot \mathrm{Ex} + 6$$

Медиана выборочного ряда определяется для четного n в виде среднего арифметического двух средних значений вариационного (упорядоченного по возрастанию или убыванию) ряда (например, $x_{(1)} \le x_{(2)} \le \ldots \le x_{(n)}$), т.е. $\mathrm{M_e} = \frac{1}{2}(x_{(n/2)} + x_{(n/2+1)})$, а для нечетного n — в виде срединного элемента $\mathrm{M_e} = \frac{1}{2}x_{(n/2)}$.

Максимальный и минимальные элементы определяются соответственно как крайние порядковые статистики вариационного ряда $x_{(1)} \le x_{(2)} \le \ldots \le x_{(n)}$. При упорядочении (сортировке) по возрастанию $x_{\max} = x_{(n)};\ x_{\min} = x_{(1)}$. Сортировка осуществляется любым известным способом, например методом Шелла.

Размах выборки определяется разностью $R = x_{\max} - x_{\min}$.

Для определения моды (значение выборки или ряда, отвечающее наибольшей частоте) осуществляют частотную группировку значений ряда. Ширина интервала либо задается пользователем, либо автоматически по формулам $L = 1 + 3.22\ \lg\ (n)$, $h = R/L$, где L — количество интервалов (при вычислении округляется в большую сторону), n — объем выборки, h — размах, R — ширина интервала.

Каждый интервал характеризуется определенной частотой f_i и частотностью попадания в него наблюдений из имеющейся выборки:

$$n = \sum_{j=1}^{L} f_j\ ,\ w_j = \frac{f_j}{n}.$$

При определении моды первоначально определяется модальный интервал. Пусть x_{Mo} — начало модального интервала, w_{Mo} — его частотность, w_{M_o-1}, w_{M_o+1} — частотности предшествующего модальному и последующего интервалов. Внутри этого интервала мода определяется по формуле

$$\mathrm{Mo} = \mathrm{x_{Mo}} + \mathrm{h}\cdot\frac{\mathrm{w_{Mo}} - \mathrm{w_{Mo-1}}}{(\mathrm{w_{Mo}} - \mathrm{w_{Mo-1}}) + (\mathrm{w_{Mo}} - \mathrm{w_{Mo+1}})}.$$

В завершении оценочных процедур осуществляется расчет доверительных интервалов для среднего и СКО. Доверительные интервалы для среднего вычисляются для 95, 99 и 99,9% уровней в виде

$$\Delta_\gamma = [\bar{x} - \frac{t\sigma}{\sqrt{n}};\ \bar{x} + \frac{t\sigma}{\sqrt{n}}].$$

Величина t определяется по таблице распределения t - статистики [8].

Доверительные интервалы для СКО для тех же уровней доверия определяется из соотношения

$\Delta_\gamma = [s(1-q);\ \ s(1+q)]$, *при* $q < 1$, или $\Delta_\gamma = [0; s(1+q)]$, *при* $q \ge 1$, где величина q определяется по таблице распределения q - статистики [8].

Труды СПИИРАН. 2008. Вып. 6. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2008. Issue 6. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

Все найденные числовые характеристики вероятностного распределения рядов наблюдений выводятся в командное окно и дублируются в виде текстового сообщения в графическом окне (рис. 2) В качестве исходных данных в настоящем подразделе используются ряды наблюдений за изменениями котировок акций компании Microsoft (MS).

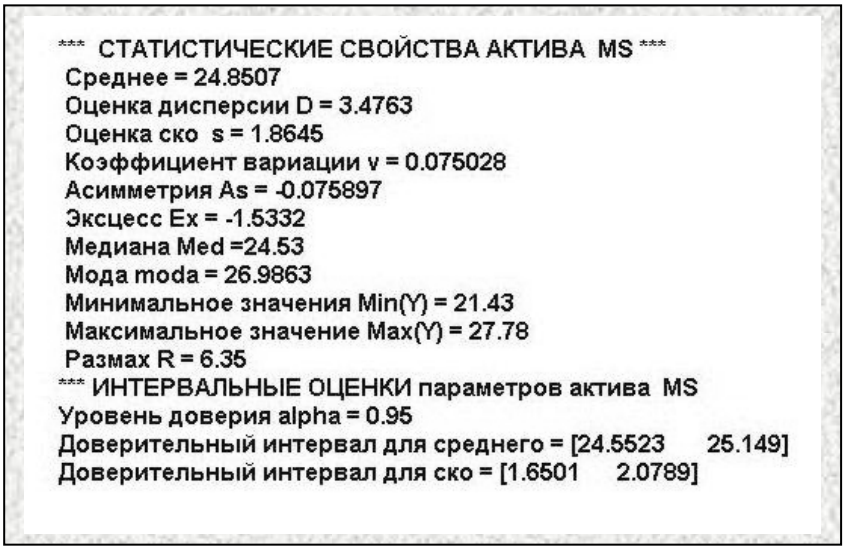

Рис. 2. Вид окна отображения результатов дескриптивного анализа данных.

Еще раз заметим, что все перечисленные характеристики относятся к параметрам распределения, описывающего вероятностную структуру исходных данных. Однако само понятие функции распределения предполагает наличие стационарных участков динамики состояния инвестиционной ситуации. В случае если динамика нестационарна, то само понятие распределения данных теряет определенность.

Разумеется, природа любых рынков является по сути своей нестационарной и, как правило, анализ вероятностной структуры данных относится не к базовым понятиям рынка (фондовые активы, котировки валют и т.п.), а к аддитивной стохастической компоненте, называемой *шумами системы*. Под этим понятием обычно подразумевается некоторый случайный процесс, возникающий под суммарным влиянием большого числа неконтролируемых возмущающих факторов. В случае, когда этих факторов достаточно много и ни один из них не является явно доминирующим, результирующий процесс оказывается близким к стационарному и, в соответствии с центральной предельной теоремой, его распределение стремится к нормальному [4,7].

Отклонение аддитивной модели шумов системы (или шумов наблюдений), приводит к существенному снижению точности прогноза и качества формируемых решений. В связи с этим в процессе дескриптивного анализа данных возникает необходимость в контроле таких свойств информационных потоков, образованных в результате мониторинга инвестиционной ситуации, как стационарность и нормальность. С этой целью обычно используется общая теория проверки статистических гипотез.

Процедура сопоставления истинности высказанной гипотезы с имеющимися выборочными данными $Z_1, Z_2, \ldots, Z_n$ осуществляется с помощью *статистического критерия* и называется *статистической проверкой гипотез*. В

результате проверки гипотезы устанавливается либо отрицательный результат (данные противоречат высказанной гипотезе), либо неотрицательный (но не положительный!) — данные не противоречат гипотезе.

Неотрицательный результат не означает оптимальности решения. Более того, он не означает, что утверждение верно.

Логическая схема статистической проверки гипотезы описана в [14].

В качестве первого примера применения данной методологии рассмотрим задачу проверку гипотезы о нормальности выборочных данных $H_0: \; x \in N(E, D)$ на основе анализа значений коэффициентов асимметрии и эксцесса [1]. Гипотеза H_0 не отвергается, если $\mathrm{As} \le 3S_{As}$, $\mathrm{Ex} \le 5S_{Ex}$, где S_{As}, S_{Ex} — их ско:

$$s_{As} = \sqrt{\frac{6n(n-1)}{(n-2)(n+1)(n+3)}};$$

$$s_{Ex} = \sqrt{\frac{24n(n-1)^2}{(n-3)(n-2)(n+3)(n+5)}}.$$

В качестве второго примера рассмотрим задачу проверки независимости и стационарности данных на основе медианного критерия серий [1].

Для уровня доверия $P = 1 - \mathrm{alpha} = 0.9 - 0.95$ гипотеза о стохастической независимости наблюдений H_0 отвергается, если окажется нарушенным хотя бы одно из условий

$$v > N + 1 - 1.96 * (sqrt(N-1))/2,$$

$$\mathrm{tau\_max} < (3.3 * \log_{10}(N+1)).$$

В противном случае считается, что гипотеза не противоречит исходным данным. Здесь v — количество серий (последовательностей) наблюдений, расположенных над уровнем и ниже уровня медианы, $\mathrm{tau\_max}$ — длина наибольшей серии.

В случае если гипотезы о стационарности и нормальности не выполняются, то возникает проблема статистической устойчивости формируемых решений.

3. Анализ динамических характеристик инвестиционных ситуаций. Проблема нестационарности информационных потоков и некоторые подходы к ее разрешению

Большинство ОИА и инвестиционных сред относятся к классу открытых нестационарных систем с нелинейной динамикой эволюции состояния. В качестве примера, на рис. 3 приведен график изменения котировок акций компании MS на интервале времени в 150 суток. Достаточно очевидно, что данный процесс является нестационарным и содержит явно выраженные нелинейные тренды.

Для исследования динамических свойств таких процессов целесообразно использовать традиционную полиномиальную аппроксимацию с подгонкой по МНК.

На рис. 3 приведены графики самого процесса, его среднего значения, линейной тенденции и полиномиальной модели второго порядка, а на рис. 4 представлено графическое окно вывода тестовой информации, отображающей

значения коэффициентов полиномиальной модели. Очевидно, что с ростом порядка аппроксимации точность аппроксимации будет возрастать. Однако до бесконечности порядок модели увеличивать нельзя; начиная с некоторого его значения, система нормальных уравнений, используемая при реализации МНК, становится плохо обусловленной.

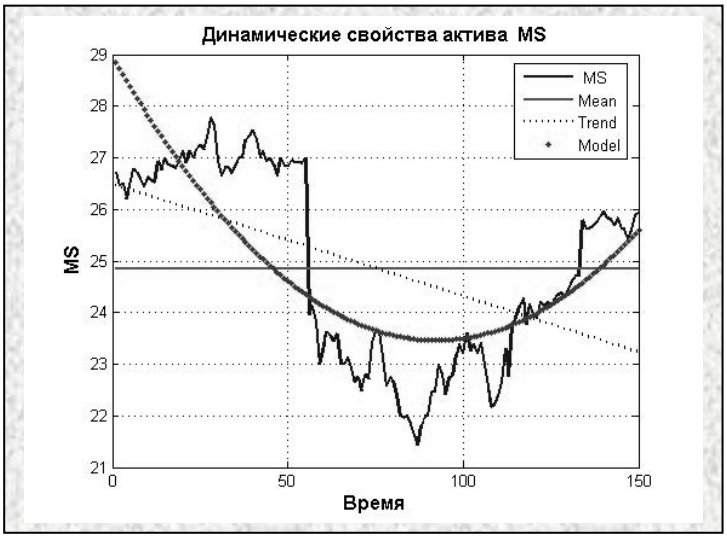

Рис. 3. Изменение котировок акций MS и МНК аппроксимация.

Следует заметить, что важнейшим объектом анализа динамики котировок является первая конечная разность исследуемого процесса, определяющая знак изменения котировок. В качестве примера на рис. 5. представлен график

Рис. 4. Окно вывода параметров аппроксимирующей модели.

изменения первой разности котировок акций MS. В отличие от исходного процесса, динамика изменения первой конечной разности оказывается близкой к стационарной. Более того, на достаточно больших участках наблюдения методами проверки статистических гипотез подтверждается гипотеза о независимости первых разностей, что существенно усложняет возможность краткосрочного прогноза изменения знака котировок.

Однако и прогноз самих котировок также связан с наличием серьезных проблем. Как правило, большинство ОИА относятся к классу открытых нестационарных систем с нелинейной динамикой эволюции состояния. В то же время основные методы статистической обработки данных, лежащие в основе анализа, прогнозирования и управления инвестиционным процессом, ориентированы на стационарные процессы. Для разрешения данного противоречия использу-

Труды СПИИРАН. 2008. Вып. 6. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2008. Issue 6. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

ется технология скользящего окна наблюдения, ограничивающая размер обучающей выборки векторных наблюдений.

Достаточно очевидно, что для высоко динамичных участков процесса эволюции состояния инвестиционной ситуации, окно наблюдения должно быть уменьшено, что позволяет снизить динамическую ошибку процесса слежения. В

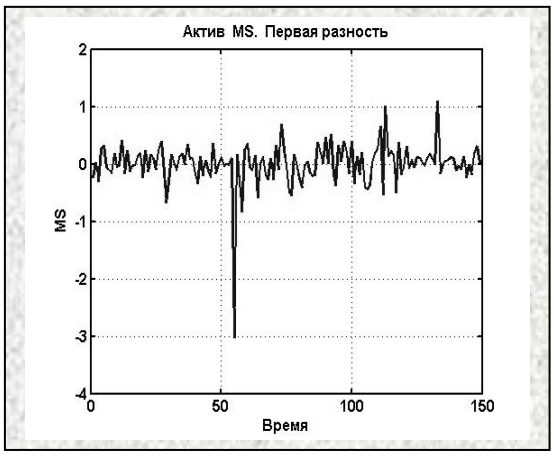

Рис. 5. Окно вывода параметров аппроксимирующей модели.

то же время, снижение размера окна наблюдения неизбежно влечет уменьшение размера выборки, по которой строятся оценки наблюдаемых параметров, что приводит к росту дисперсионной компоненты погрешности управления.

Отсюда вытекает необходимость выбора оптимального размера окна наблюдения, причем величина этого окна будет зависеть от текущей динамики контролируемого процесса. Наиболее адекватное решение в таких ситуациях дают адаптивные методы, устанавливающие размер используемой ретроспективной памяти L в зависимости от величины *полного квадрата ошибки* (ПКО) прогнозирования на окне наблюдения. При этом под ПКО понимается сумма квадрата смещения оценки (динамическая ошибка скользящего прогноза) b^2 и дисперсии оценки (статистическая ошибка) оцениваемых параметров σ^2:

$$d^2 = b^2 + \sigma^2 .$$

Организация скользящего окна наблюдения и отвечающая ему схема структуризации данных представлены на рис. 6.

При этом информационная платформа анализа представлена в виде двумерной таблицы, в левой части которой представлены данные наблюдений за инвестиционной ситуацией, используемые в качестве предикторов в модели прогноза состояния ОИА (фондовые индексы, параметры экономической ситуации, котировки акций компаний, связанных с ОИА и т.п.). Заметим, что для задач прогнозирования в левую часть таблицы могут входить и параметры самого ОИА.

В правой части таблицы представлены параметры, описывающие состояние ОИА (например, котировки акций активов, используемых при формировании инвестиционного портфеля). При этом для задач прогнозирования осуществляется сдвиг ретроспективных данных, соответствующих состоянию ОИА, относительно данных, описывающих состояние инвестиционной среды (среда

Труды СПИИРАН. 2008. Вып. 6. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2008. Issue 6. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

взаимодействия) на число шагов $k(\tau)$, отвечающих интервалу планируемого прогноза τ. В задачах анализа состояния ОИА такой сдвиг делать не нужно.

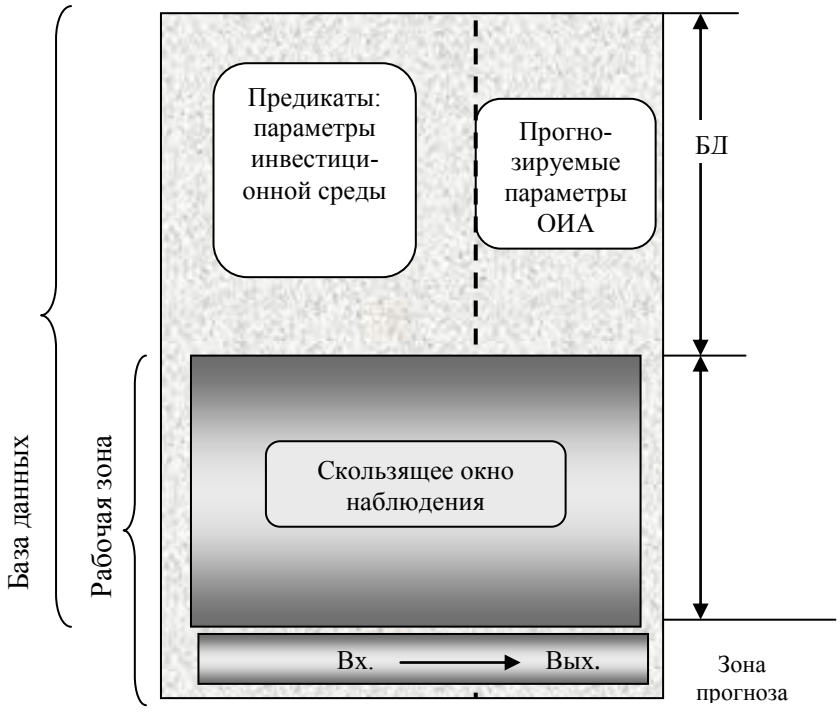

Рис. 6. Структуризация данных и организация скользящего окна наблюдений.

Совокупность ретроспективных данных за промежуток времени $(k-L,k-1)$, где k — отсчет времени, отвечающий текущему состоянию инвестиционной среды и текущему (для задач анализа) или сдвинутому на $k(\tau)$ шагов (для задач прогноза) состоянию ОИА, образует обучающую выборку наблюдений, используемую для построения математической модели инвестиционной ситуации.

Наличие скользящего окна наблюдения позволяет оперативно перестраивать модель как в части ее структуры, так и с точки зрения ее параметрической идентификации. Данный подход предоставляет возможность последовательно «отслеживать» динамику процесса, причем оставшиеся невязки аппроксимации образуют процесс, весьма близкий к стационарному гауссовскому шуму.

4. Заключение

Существующие способы оценки стоимости рыночных активов и модели прогнозирования рынка достаточно эффективны лишь в условиях значительных теоретических и практических ограничений. К сожалению, они мало учитывают изменения информационных потоков на финансовых рынках. Например, очень часто все параметры для анализа берутся из одного источника информации (например, баланса), при этом понятно, что при условии фальсификации или устаревании информации число учитываемых параметров не оказывает принципиального влияния на точность инвестиционных оценок. Такое положение дел облегчило недобросовестным участникам рынка процесс манипулирования ценами, фальсификацию рекомендаций. Не случайно в последнее время

Труды СПИИРАН. 2008. Вып. 6. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2008. Issue 6. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

все чаще поднимается вопрос о повышении ответственности со стороны инвестиционных аналитиков за их прогнозы.

Непосредственное применение программных технологий аналитических исследований к массивам накопленной ретроспективной информации, как правило, не обеспечивает получение достоверных результатов, поэтому выработка качественных инвестиционных решений невозможна без качественного предварительного анализа данных. Такими видами анализа могут быть статистический и динамический анализы рядов наблюдений в совокупности с адаптивными методами, позволяющими решить проблему нестационарности информационных потоков.

Однако, несмотря на очевидные преимущества использования метода «скользящего окна», в данной статье рассмотрены лишь первые функциональности автоматической системы анализа свойств информационных потоков. Чтобы получить действительно приемлемые результаты для прогнозирования и принятия качественных инвестиционных решений, необходимо провести все виды анализа, перечисленные во введении. Как уже отмечалось, данные исследования будут представлены в последующих сборниках трудов СПИИРАН.

Литература

1. Автоматизированное рабочее место для статистической обработки данных / *В. В Шураков., Д. М Дайитбеков., С. В Мизрохин., С. В Ясеновский*. М.: Финансы и статистика, 1990. 190 с.
2. *Аджиев В*. Mineset — визуальный инструмент аналитика. // Открытые системы. 1997. № 3. С. 72–77.
3. *Гершберг А. Ф., Мусаев А. А., Нозик А. А., Шерстюк Ю. М.* Концептуальные основы информационной интеграции АСУ ТП нефтеперерабатывающего предприятия. СПб: Альянс-строй, 2003. 128 с.
4. *Вентцель Е С.* Теория вероятностей. М.: Наука, 1969. 576с.
5. *Дюк В., Самойленко А.* Data Mining: Учебный курс. СПб.: Питер, 2001. 366 с.
6. *Заде Л.* Основы нового подхода к анализу сложных систем и процессов принятия решений // Математика сегодня. М.: Знание, 1974. С. 5–48.
7. *Кендалл М., Стьюарт А*. Статистические выводы и связи // Пер. с англ. под ред. А. Н Колмогорова. М.: Наука, 1973. 900 с.
8. *Киселев М., Соломатин Е.* Средства добычи знаний в бизнесе и финансах // Открытые системы, 1997. № 4. С. 41–44.
9. *Кречетов Н., Иванов П*. Продукты для интеллектуального анализа данных // Computer Week. 1997. № 14–15. С. 32–39.
10. *Кривда Ш.* Раскопки сокрытых знаний. // ЛАН. 1996. № 4. С 17–23.
11. *Львов В.* Создание систем поддержки принятия решений на основе хранилищ данных // СУБД. 1997. № 3. С. 30–40.
12. *Львович О.* Data Warehousing – выход из кризиса оперативного анализа // Read Me. 1998. № 6. С. 44–45, 66.
13. *Мусаев А. А*. Виртуальные анализаторы: концепция построения и применения в задачах управления непрерывными технологическими процессами // Автоматизация в промышленности. 2003. № 8. С. 28–33.
14. Прикладная статистика: Основы моделирования и первичная обработка данных. Справочное изд. / *С. А. Айвазян, И. С. Енюков, Л. Д. Мешалкин.* М.: Финансы и статистика, 1983. 471 с.
15. *Фогель Дж., Оуэнс Дж., Уолш Л.* Эволюционное моделирование и искусственный интеллект. М.: Мир, 1969. 219 с.
16. *Шапот М.* Интеллектуальный анализ данных в системах поддержи принятия решений // Открытые системы. 1998. № 1. С. 30–35.
17. *Codd E. F., Codd S. B., Salley C. T.* Providing OLAP (On-Line Analytical Processing) to User-Analysts: An IT Mandate. E. F. Codd Associates, 1993. 18 p.

18. *W. H. Inmon.* Building the Data Warehouse. Wellesley, MA: QED Publishing Group, 1992.
19. *Musaev A.* Intelligent Control Systems for Refinery Technological Processes // Proceedings of conf. ICPI'02 (Intelligent computing for the petroleum industry. Mexico, 2002. vol. 2. P. 6–17.