Мусаев А.А., Иниваткин А.А.

Прогнозирование состояния нестационарных процессов на основе выявления и обработки ситуаций-аналогов

Санкт-Петербургский Государственный Технологический Институт
(Россия, Санкт-Петербург)

doi:10.18411/spc-12-05-2019-13

idsp: sciencepublic-12-05-2019-13

Аннотация

Рассматривается задача прогнозирования на основе выявления ситуаций-аналогов в массивах ретроспективных данных. В случае, когда системная компонента наблюдаемого процесса представляет собой непериодический колебательный временной ряд (реализацию хаотической динамики), традиционные схемы распознавания оказываются малоэффективными. Задача состоит в построении системы выявления и обработки аналогов, позволяющей получить прогноз для условий нестационарной динамики.

Ключевые слова: нестационарный случайный процесс, прогнозирования, аналоги, прецедент, метод ближайшего соседа, метрическая классификация.

Введение. Рассматривается задача прогнозирования нестационарного случайного процесса, характерного, например, для управления газодинамическими технологическими процессами (ТП). Особенностью решаемой задачи является системная компонента рассматриваемого процесса, представляющая собой реализацию детерминированного хаоса, т.е. колебательный непериодический процесс [1, 2]. В этих условиях традиционные методы экстраполяционного прогноза оказываются не эффективными. Вариант альтернативного прогноза, рассмотренный в настоящей работе, основан на методе ближайших соседей [3, 4]. При таком подходе, методами распознавания в базе данных (БД) ретроспективных наблюдений отыскиваются один (прецедент) или несколько аналогов, подобных текущему состоянию процесса. Последующие за ними интервалы наблюдений принимаются в качестве прототипов прогноза. Достаточно ожидаемо, что для нестационарного процесса с хаотической системной составляющей последствия, используемые в качестве прототипов прогноза, могут существенно отличаться от реальных изменений процесса. Анализ эффективности такого подхода и возможности его улучшения на основе совместной обработки найденных ситуаций-аналогов посвящена настоящая работа.

Постановка задачи. Имеем одномерный временной ряд $Y(k)$, $k = 1, \dots, N$, формируемый системой мониторинга реального нестационарного процесса. Простейшая модель такого процесса задается известной аддитивной схемой $Y(k) = X(k) + v(k)$, $k = 1, \dots, N$, где $v(k)$ – случайный процесс, имитирующий шумы системы. Примерами таких процессов являются вариации вектора состояния газодинамической среды ТП. Простейший анализ независимости и стационарности исследуемого процесса, например, на основе медианного критерия серий или критерия восходящих и нисходящих серий [5], показывает, что он не удовлетворяет указанным условиям. Более того, системная компонента $X(t)$ данного процесса, порождается хаотической динамикой, и представляет собой непериодический колебательный процесс. Отсюда непосредственно следует, что традиционные методы экстраполяционного прогноза случайных процессов неприемлемы по своей сути: невозможно пролонгировать динамику с изменяющимися по неизвестному правилу параметрами процесса. Однако в силу ограниченной инерционности многих реальных

процессов, могут оказаться приемлемыми некоторые методы прецедентного анализа, широко используемые в задачах распознавания образов.

В качестве метрики близости $\rho(Y_w(k), Y_w(k-i))$ используется квадратичное расстояние между текущим окном наблюдения $Y_w(k) = (Y(k), \dots, Y(k-L))$ и окнами данных $Y_w(i) = Y(i, i+L)$, $i = 1, \dots, N_0 - L$, формируемые в процессе сканирования БД ретроспективных наблюдений объемом в N_0 отсчетов. Здесь L - размер окон состояния и сканирования. Для нестационарных процессов величины L и τ представляет собой опции, допускающие уточнение в процессе адаптации модели наблюдения.

Простейший вариант прогноза сводится к методу ближайшего соседа, при котором в качестве естественного прецедента выбирается окно сканирования с номером i^* , обладающее максимальным подобием окну текущего состояния $Y_w(k)$, т.е.

$$Y_w(i^*): \rho[Y_w(k), Y_w(i^*)] = \min \quad \forall i = 1, \dots, N_0 - L. \quad (1)$$

В качестве прогноза $\tilde{Y}(k+\tau)$ в этом случае используется соответствующее последствие за окном прецедента, т.е. значение наблюдения из БД, отстоящее от границы окна сканирования на интервал прогнозирования τ $\tilde{Y}(k+\tau) = Y(i^* + L + \tau)$.

В более общем случае [6-8] в качестве прогноза используется усреднение последствий по группе из m_a аналогов $(i_1^*, \dots, i_{m_a}^*)$, обладающих наименьшими значениями метрики $\rho[Y_w(k), Y_w(i)] \quad \forall i = 1, \dots, N_0 - L$, т.е.

$$\tilde{Y}(k+\tau) = \frac{1}{m_a} \sum_{j=1}^{m_a} Y(i_j^* + L + \tau). \quad (2)$$

Недостатком последнего соотношения является равнозначность всех аналогов, независимо от степени их подобия текущей ситуации $Y_w(k)$. В связи с этим целесообразно использовать в качестве прогноза взвешенную сумму последствий с весами, пропорциональными степени подобия. Если для прецедента (1) этот вес выбрать равным единице $w_1 = 1$, то для остальных аналогов данные веса равны $w_j = \rho_1 / \rho_j = 1 / \rho_j$, $j = 2, \dots, m_a$. Тогда взвешенное значение прогноза определяется из соотношения

$$\tilde{Y}(k+\tau) = \frac{1}{S(m_a)} \sum_{j=1}^{m_a} \omega_j Y(i_j^* + L), \quad \text{где} \quad S(m_a) = \sum_{j=1}^{m_a} w_j. \quad (3)$$

Следует отметить, что последовательный поиск аналогов, отвечающих критерию (1), как правило, приводит к формированию групп подобия, состоящих из рядом стоящих окон сканирования. Следовательно, предложенное расширение метода ближайшего соседа практически не привносит новой информации и не улучшает качество прогноза. Для устранения указанного недостатка необходимо усложнить выбор аналогов, полученных в процессе сканирования БД. В качестве варианта программного решения может быть предложена вычислительная схема, основанная на искусственно формируемой области запрета выбора аналогов, последовательно включающей в себя Δ_j - окрестности отсчётов i_j , $j = 1, \dots, m_a$, отвечающих выбранным

аналогам. В случае, если последующее по величине возрастания метрики близости в упорядоченном массиве данных окно сканирования окажется в области запрета, то оно не может использоваться в качестве аналога. Размер области запрета Δ_j также является опцией, в качестве первого приближения его выбирают равным половине окна сканирования $L/2$.

Значимой особенностью поиска аналогов является требование предварительного сглаживания прогнозируемого процесса. Динамическое сглаживание позволяет отыскивать подобные ситуации, ориентируясь на системную составляющую процесса. Такой подход позволяет повысить достоверность промежуточных решений по выбору ситуаций-аналогов. В данной работе использовалось простейшее экспоненциальное сглаживание $\hat{Y}(t) = \alpha Y(t) + (1 - \alpha)\hat{Y}(t-1)$ с коэффициентом сглаживания $\alpha = 0.02$. В общем случае, коэффициент α , так же как и другие опционные параметры, целесообразно включить в список показателей, подлежащих параметрической оптимизации.

Пример выбора ситуаций-аналогов представлен на рис. 1. Вертикальная линия разделяет зону ретроспективных данных, используемых для формирования прецедентного прогноза, от рабочей зоны, на которой осуществляется тестирование алгоритма. Начало и концы интервалов, отвечающих текущей ситуации и аналогам, отмечены, соответственно звёздочками и крестиками, значения прогноза – ромбиками.

На рис. 2 приведены развернутые графики окна текущего состояния (верхний график) и окон-аналогов. Второе сверху окно одновременно является прецедентом по методу ближайшего соседа. На сплошные линии аналогов пунктиром наложен график текущего состояния, позволяющий наглядно убедиться в качественном решении задачи селекции аналогов на основе квадратичной метрики.

Рис. 1. Пример текущей ситуации и окон-аналогов на общем временном графике сглаженного процесса.

Рис. 2. Окно текущей ситуации (верхний график) и отвечающие ему окна-аналоги

Для улучшения качества прогноза следует обратить внимание на степень подобия аналогов. В частности, второй аналог дал совершенно неправильный прогноз, однако степень подобия его по отношению прецедента составляет лишь 0.49. Это означает, что предварительная фиксация числа ситуаций аналогов для нестационарных процессов ухудшает качество прогноза и, следовательно, программа должна быть дополнена вторым уровнем селекции аналогов, определяемым пороговым уровнем допустимого подобия.

С другой стороны следует учитывать размер интервала прогнозирования для процессов данного класса. Очевидно, что уменьшение интервала прогнозирования

позволяет получить более точный результат. Так, например, снижение интервала прогноза с 200 до 100 отсчетов позволило снизить ошибку прогноза в рассмотренном примере на 17,5% и получить вполне удовлетворительное решение для решаемой прикладной задачи. Однако сокращение времени прогнозирования не всегда допустимо. Например, в задаче оценки параметров качества выходного потока ТП по значениям параметров его состояния интервал прогноза определяется структурой производства и не допускает изменения.

Рис. 3. Пример изменения относительной точности прогноза параметра технологического процесса с нестационарной динамикой

Улучшение качества прогнозирования может достигаться за счет использования системы распределенных прогнозаторов, основанных на слабых классификаторах с последующей совместной обработкой. В частности, могут независимо использоваться классификаторы, основанные на соотношениях (1-3) с выбором решения, отвечающего варианту с наиболее точным прогнозом на предшествующем скользящем участке наблюдения. Такая схема позволяет получить для определенного класса прикладных задач вполне удовлетворительные результаты. Так, например, прогнозирование текущей температуры в химическом реакторе, используемом в процессе производства ароматических углеводородов, позволяет получить, как это видно на рис. 3., относительную погрешность прогноза в пределах нескольких процентов. При этом чаще всего предпочтение отдавалось варианту прогноза с тремя ближайшими соседями с селекцией аналогов на уровне подобия менее 50% и экспоненциальным сглаживанием с астатизмом первого порядка и коэффициентом передачи $\alpha = 0.02$.

Для нестационарных процессов более сложной природы, не обладающих массовой или энергетической инерционностью, построить эффективной прогноз на основе рассмотренной простейшей схемы не удалось. Это означает необходимость в дальнейших исследованиях, основанных на учете внешних факторов и многомерных статистических метриках.

Заключение. Представленные исследования позволяют обосновать допустимость применения прецедентного анализа для некоторого класса задач прогнозирования нестационарных процессов с хаотической системной компонентой. В частности, приведенный подход обычно приемлем для процессов, обладающих инерционностью. Однако для безынерционных процессов, например, порожденных динамикой рыночных котировок, такой прогноз совершенно не соответствует реальным значениям. В связи с этим необходимо проведение дальнейших исследований в двух направлениях:

1. Построение самоорганизующихся алгоритмов прогнозирования, учитывающих изменения внешних факторов, влияющих на контролируемый процесс;

2. Применение многомерного анализа по группе коррелированных процессов с использованием метрик многомерного статистического анализа. В частности, представляет интерес исследовать эффективность прецедентного прогноза на основе расстояния Махаланобиса, или других, достаточно нестандартных метрик, таких, как след Хотеллинга или след Пиллаи [8].

Исследования, выполненные по данной тематике, проводились в рамках бюджетной темы №0073–2019–0004 и Международного проекта ERASMUS+, Capacity building in higher education, №73751-EPP-1-2016-1-DE-EPPKA2-CBHE-JP.

The research described in this paper is partially supported by the state research 0073–2019–0004 and International project ERASMUS+, Capacity building in higher education, # 73751-EPP-1-2016-1-DE-EPPKA2-CBHE-JP.

1. Peters, E. E. Chaos and order in the capital markets: a new view of cycles, prices, and market volatility (2nd ed.)/ E.E. Peters – NY: John Wiley & Sons, 1996. – 288p.
2. Омельченко, А. В. Нестационарные процессы в газовой динамике/ А.В. Омельченко, В. Усков, А. Кожемякин – СПб.: Издательство БГТУ, 2005.
3. Advances in Data Mining. Applications and Theoretical Aspects – 18th Industrial Conference, ICDM 2018, New York, NY, USA, July 11-12, 2018, Proceedings. Lecture Notes in Computer Science, 10933, Springer 2018. – 452p.
4. Larose, D.T. Discovering knowledge in data. Second ed./ D.T. Larose, C.D. Larose – Wiley: 2014. – 336p.
5. Айвазян, С.А. Прикладная статистика. Основы моделирования и первичная обработка данных/ С.А. Айвазян, И.С. Енюков, Л.Д. Мешалкин/ Под ред. С.А. Айвазяна – М.: Финансы и статистика, 1993. – 471с.
6. Гухман, А.А. Введение в теорию подобию/ А.А. Гузман – ЛКИ: 2010. – 296с.
7. Макшанов, А.В. Технологии интеллектуального анализа данных/ А.В. Макшанов, А.Е. Журавлев – Лань: 2018. – 212с.
8. Айвазян, С.А. Прикладная статистика. Классификация и снижение размерности/ С.А. Айвазян, В.М. Бухштабер, И.С. Енюков, Л.Д. Мешалкин/ Под ред. С.А. Айвазяна – М.: Финансы и статистика, 1989. – 607с.

Мусаев А.А., Фенин М.М.

**Робастификация алгоритмов регрессионного анализа в задачах оценивания
изменения состояния нестационарных систем**

*Санкт-Петербургский Государственный Технологический Институт
(Россия, Санкт-Петербург)*

doi:10.18411/spc-12-05-2019-14

idsp: sciencepublic-12-05-2019-14

Статистическая технология робастификации алгоритмов оценивания позволяет снизить чувствительность формируемых оценок к вариациям вероятностных характеристик наблюдаемого процесса. В частности, робастные оценки менее чувствительны к вариациям функции распределения, и в определенных случаях применимы к обработке нестационарных процессов. При этом предполагается, что вариации функции распределения не выводят их за пределы некоторого априори заданного класса, характеризуемого известным набором ограничений. Расширение класса допустимых вариаций функции распределения неизбежно приводит к ориентации на более «загрубленные» методы обработки, и, как следствие к той или иной форме снижения качества оценивания. В частности, для алгоритмов формирования системной компоненты динамики котировок, снижение чувствительности следящего фильтра ведет к росту задержки оценки (лага), который, в свою очередь, приводит к снижению качества управления торговыми операциями.

Таким образом, возникает необходимость в построении версий робастификации алгоритмов обработки данных, позволяющий, с одной стороны, сохранить их устойчивость к вариациям статистических характеристик наблюдаемых процессов, а с