Тема работы:

Использование подходов искусственного интеллекта для разработки и исследования методов прогнозирования временных рядов.

Состав коллектива:

Чирихин Константин Сергеевич, аспирант кафедры Компьютерных систем ФИТ НГУ; младший научный сотрудник, Федеральный исследовательский центр информационных и вычислительных технологий

Рябко Борис Яковлевич, д.т.н., профессор кафедры Компьютерных систем ФИТ НГУ; главный научный сотрудник, Федеральный исследовательский центр информационных и вычислительных технологий

Информация о грантах

1. Грант РФФИ 19-37-90009 Аспиранты "Методы прогнозирования временных рядов, базирующиеся на алгоритмах сжатия данных и искусственного интеллекта", организация - ФИЦ ИВТ, руководитель - Рябко Б.Я.
2. Грант РФФИ 19-47-540001 р_а "Разработка когнитивных методов прогнозирования и их применение для предсказания социально-экономических процессов в Новосибирской области", организация - НГУ, руководитель - Рябко Б.Я.

Научное содержание работы

1. Постановка задачи.

Разрабатывается и исследуется метод прогнозирования временных рядов, основанный на алгоритмах сжатия данных и искусственного интеллекта и способный обнаруживать новые классы нестационарных закономерностей в данных. Для исследования точности предлагаемого метода при прогнозировании реальных процессов проводятся вычисления на физических и социально-экономических временных рядах.

2. Современное состояние проблемы.

Задача прогнозирования временных рядов имеет множество практических приложений и к настоящему времени было предложено достаточно большое количество разнообразных подходов к её решению. Среди наиболее популярных из них мы отметим модели экспоненциального сглаживания [1], авторегрессии-скользящего среднего [2], авторегрессионной условной гетероскедастичности [3], модели на основе нейронных сетей и машинного обучения [4]. Однако несмотря на разнообразие методов, во многих из них предполагается наличие линейной или простой нелинейной взаимосвязи между прошлыми и будущими значениями, что далеко не всегда выполняется на практике [5]. Существуют закономерности, очевидные для человека, которые известные методы прогнозирования временных рядов корректно экстраполировать не могут. В качестве простейшего примера

данных с закономерностью подобного класса можно привести последовательность 01001000100001… .

В данной работе мы развиваем теоретико-информационный подход к прогнозированию. Впервые использовать методы сжатия для прогнозирования временных рядов было предложено в [6]. Показано, что любой метод сжатия данных без потерь может быть использован для прогнозирования временных рядов, и если временной ряд порождён стационарным и эргодическим источником, а метод сжатия данных является универсальным кодом, то прогноз будет в определённом смысле оптимальным [7]. В предыдущих работах [8, 9] по прогнозированию данных реальных процессов с помощью методов сжатия для построения прогноза использовался только один универсальных код, хотя современные методы сжатия разнообразны и заранее не известно, какой из методов лучше подойдёт для рассматриваемого временного ряда. В данной работе мы используем комбинации из различных методов сжатия при прогнозировании, а также рассматриваем прогнозирование нестационарных данных с закономерностями некоторых классов.

1. Forecasting with exponential smoothing: the state space approach / R. Hyndman [et al.]. — Berlin/Heidelberg, Germany : Springer-Verlag, 2008. — 375 p.
2. Box G. E., Jenkins G. Time series analysis: forecasting and control. — San Francisco : Holden-Day, 1970. — 553 p.
3. Franco C., Zakoian J. M. GARCH models: structure, statistical inference and financial applications. – John Wiley & Sons, 2019. — 485 p.
4. Tealab A. Time series forecasting using artificial neural networks methodologies: A systematic review // Future Computing and Informatics Journal. — 2018. — Vol. 3, no. 2. — P. 334—340.
5. Cheng C. et al. Time series forecasting for nonlinear and non-stationary processes: a review and comparative study // IIE Transactions. – 2015. – Vol. 47. – no. 10. – P. 1053-1071.
6. Рябко Б. Я. Прогноз случайных последовательностей и универсальное кодирование // Проблемы передачи информации. — 1988. — Т. 24, №. 2. — С. 3—14.
7. Ryabko B. Applications of Kolmogorov complexity and universal codes to nonparametric estimation of characteristics of time series // Fundamenta Informaticae. — 2008. — Vol. 83, no. 1/2. — P. 177—196.
8. Приставка П. А. Экспериментальное исследование метода прогнозирования, основанного на универсальных кодах // Вестник СибГУТИ. — 2010. — №. 4. — С. 26—35.
9. Лысяк А. С., Рябко Б. Я. Методы прогнозирования временных рядов с большим алфавитом на основе универсальной меры и деревьев принятия решений // Вычислительные технологии. — 2014. — Т. 19, №. 2. — С. 76—93.

3. Подробное описание работы, включая используемые алгоритмы.

Мы разрабатываем метод прогнозирования временных рядов, использующий набор алгоритмов сжатия данных для построения прогноза. В этот набор могут быть включены произвольные методы сжатия без потерь, в нашей программной реализации поддерживается работа с *zlib*, *bzip2*, *ppmd*, *rp*, *lcacomp*, *zstd*, *zpaq*. Если прогнозируемый временной ряд состоит из символов некоторого конченого множества (алфавита) A , метод непосредственно может быть применён для его прогнозирования. Если временной ряд

вещественный, предварительно используется процедура квантования – множество возможных значений ряда разбивается на конечное число интервалов и вместо прогнозирования исходных значений прогнозируются номера интервалов. Прогноз представляется в виде распределения вероятностей. Пусть $F=\{\phi_1, \phi_2, \ldots, \phi_k\}$ - некоторое конечное множество методов прогнозирования, $x_1, x_2, \ldots, x_t$ - прогнозируемый временной ряд. Мы находим оценку условной вероятности появления $a \in A$ в качестве следующего элемента временного ряда с помощью всех методов сжатия из F по следующей формуле:

$$P_\phi(x_{t+1}=a \mid x_1, x_2, \ldots, x_t) = \left(\sum_{\phi \in F} 2^{-|\phi(x_1, x_2, \ldots, x_t, a)|}\right) / \left(\sum_{\phi \in F} \sum_{b \in A} 2^{-|\phi(x_1, x_2, \ldots, x_t, b)|}\right), \quad (1)$$

где $|\phi(u)|$ - размер сжатого представления последовательности u в битах. Если требуется получить прогноз на более чем один шаг вперёд, то a и b нужно брать не из A, а из A^h, где h - количество шагов, на которое строится прогноз. В качестве точечных прогнозов мы использовали математические ожидания. Данный метод был обобщён для прогнозирования многомерных данных. Для сокращения объёма вычислений реализован адаптивный метод прогнозирования, в котором с помощью всех алгоритмов из F сжимается только небольшой фрагмент данных, и затем для прогнозирования используются только алгоритмы с наилучшей степенью сжатия на выбранном фрагменте.

Для прогнозирования последовательностей наподобие 010010001… был разработан метод на основе конечных автоматов, основанный на алгоритме из [1]. Данное слово относится к классу полилинейных слов, определённому в [2]. В [1] был предложен алгоритм для 10-головочного автомата, выполняя который автомат, начиная с некоторого t_0, начинает безошибочно прогнозировать x_{t+1} для любого полилинейного слова $x_1, \ldots, x_t$ (x_i принадлежит некоторому конечному алфавиту). В данной работе мы модифицировали этот алгоритм таким образом, что его можно рассматривать как метод сжатия данных и использовать совместно с архиваторами.

Для включения в формулу 1 произвольных методов прогнозирования, не основанных на сжатии данных, мы разработали модификацию, позволяющую рассматривать любой метод прогнозирования π, способный по последовательности целых или вещественных чисел $x_1, x_2, \ldots, x_t$ дать прогноз для x_{t+1}, как метод сжатия данных. Таким образом, мы можем использовать методы сжатия данных и алгоритм на основе автоматов совместно с любыми методами прогнозирования.

1. Smith T. Prediction of infinite words with automata // Theory of Computing Systems. — 2018. — Vol. 62, no. 3. — P. 653—681.
2. Smith T. On infinite words determined by stack automata // IARCS Annual Conference on Foundations of Software Technology and Theoretical Computer Science (FSTTCS 2013). — IIT Guwahati, India, 2013. — P. 413—424.

4. Полученные результаты.

Мы провели вычисления с использованием разработанного метода для временного ряда солнечных пятен, ряда планетарного K-индекса, а также некоторых социально-экономических показателей Новосибирской области. Результаты вычислений, на наш взгляд, показывают, что разработанные методы обладают высокой точностью. Например, для

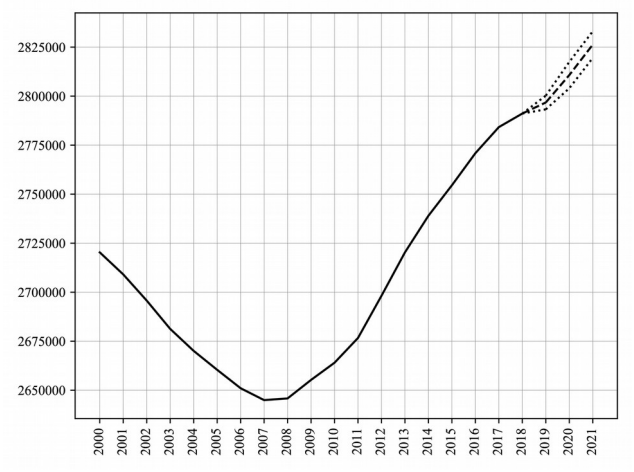
временного ряда среднемесячного количества солнечных пятен доступен архив прогнозов, выполненных Службой космической погоды (The Space Weather Services, SWS) Австралийского метеорологического бюро (http://listserver.ips.gov.au/pipermail/ips-ssn-predictions/). Мы построили прогнозы на 4 шага вперёд с совместным использованием разбиений области возможных значений ряда на 2, 4, 8 и 16 интервалов для каждого месяца, начиная с февраля 2016 года по май 2020 года. При использовании 50% значений ряда в адаптивном методе для выбора лучшего архиватора был выбран zstd. Далее приведём в таблице средние абсолютные ошибки прогнозов на шаги 1-4, где под комбинированным методом обозначена комбинация из всех 7 доступных архиваторов и автомата. Видно, что zstd обеспечивает такую же точность, что и комбинированный метод, при этом в среднем прогнозы на 1 шаг у нас оказались несколько точнее, чем у SWS.

Метод	Номер шага			
	1	2	3	4
zstd	8.1	10.3	11.8	13.3
Комбинированный метод	8.1	10.3	11.8	13.3
SWS	8.3	9.1	9.7	9.8

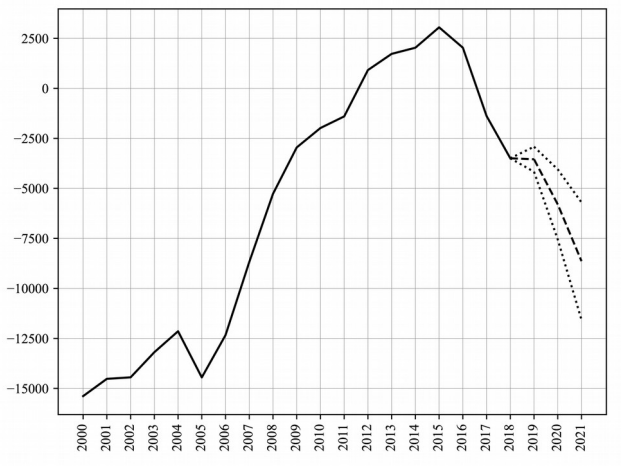
Использование комбинированного метода при построении одного прогноза на 4 шага вперёд позволило сократить время вычислений более чем в 17 раз по сравнению с комбинированным методом без потери точности.

5. Иллюстрации, визуализация результатов.

В качестве ещё одного примера приведём прогноз для временных рядов среднегодовой численности населения (а) и естественного прироста населения (б) в Новосибирской области, опубликованные в [2].



(а)



(б)

6. Эффект от использования кластера в достижении целей работы

Количество последовательностей, которые необходимо сжать всеми используемыми алгоритмами в нашем методе, определяется как $|A|^h$, где $|A|$ - количество символов в алфавите временного ряда, h - количество шагов, на которое строится прогноз. Для оценки точности нашего метода мы строили прогнозы для временных интервалов, за которые значения рассматриваемого процесса были уже известны. В связи с этим, число вычисляемых прогнозов было большим и использование кластера позволило проводить расчёты с алфавитами большей размерности и на большее количество шагов, чем если бы вычисления проводились на локальной машине.

Перечень публикаций, содержащих результаты работы.

1. Chirikhin K., Ryabko B. Compression-Based Methods of Time Series Forecasting // Mathematics. — 2021. — Vol. 9, no. 3. — P. 1—11. — URL: https://www.mdpi.com/2227-7390/9/3/284.
2. Чирихин К. С., Рябко Б. Я. Применение методов искусственного интеллекта и сжатия данных для прогнозирования социальных, экономических и демографических показателей Новосибирской области // Вычислительные технологии. — 2020. — Т. 25, No 5. — С. 80—90.