УДК: 004.855.5 DOI: 10.26102/2310-6018/2019.27.4.039

ПРОГНОЗИРОВАНИЕ ВРЕМЕННЫХ РЯДОВ ПОСРЕДСТВОМ ПРИВЯЗКИ СОБЫТИЙ

И.Н. Колесников
ФГБОУ ВО «Пензенский государственный университет», Пенза, Российская Федерация
e-mail: iljakolesnikoff@yandex.ru

Резюме: В данной статье рассматривается концепция модификации метода анализа временных рядов, ориентированная на интеграцию с методами кластеризации в режиме обучения в реальном времени. Проанализированы различные методы прогнозирования временных рядов и машинного обучения. Описанный в статье метод дает прогноз поведения временного ряда на основе больших данных, полученных из различных источников, и связанных с существующими транзакциями временного ряда. Такой подход дает возможность находить зависимости изменения определенных показателей рассматриваемых систем в зависимости от различных событий. Выполненное исследование предлагает концепцию автоматизированного обучения системы в режиме реального времени с возможностью дальнейшей программной реализации. Рассматриваемая концепция позволяет строить прогнозы на любые временные ряды, зависимые от различных событий, новостей и данных, находящихся в открытом доступе. Предложен подход, который связывает события с графиком транзакций. Преимуществом подхода является возможность нахождения различных зависимостей между происходящими событиями и различными изменениями показателей, например: цен на биржах, значений социальных показателей и многих других.

Ключевые слова: анализ данных, прогнозирование, временные ряды, большие данные, кластерный анализ, интеллектуальный анализ данных.

Для цитирования: Колесников И.Н. Прогнозирование временных рядов посредством привязки событий. *Моделирование, оптимизация и информационные технологии*. 2019;7(4). Доступно по: https://moit.vivt.ru/wp-content/uploads/2019/11/Kolesnikov_4_19_1.pdf DOI: 10.26102/2310-6018/2019.27.4.039

FORECASTING TIME SERIES USING EVENT BINDING

I.N. Kolesnikov
Penza State University, Penza, Russian Federation

Abstract: This article discusses the concept of modification of the time series analysis method, focused on integration with clustering methods in real-time training mode. Various methods of forecasting time series and machine learning are analyzed. The method described in the article predicts the behavior of the time series based on large data obtained from various sources and associated with existing transactions in the time series. This approach makes it possible to find the dependence of changes in certain indicators of the considered systems depending on various events. The performed research offers the concept of automated system training in real time with the possibility of further software implementation. The concept under consideration allows you to build forecasts for any time series, depending on various events, news and data that are in the public domain. An approach is proposed that links events to a transaction chart. The advantage of this approach is the ability to find various dependencies between events and various changes in indicators, for example: prices on exchanges, values of social indicators and many others.

Keywords: data analysis, forecasting, time series, big data, cluster analysis, data mining.

For citation: Kolesnikov I.N. Forecasting time series using event binding. *Modeling, optimization and information technology*. 2019;7(4). Available by: https://moit.vivt.ru/wp-content/uploads/2019/11/Kolesnikov_4_19_1.pdf DOI: 10.26102/2310-6018/2019.27.4.039 (In Russ.).

Введение

В современном мире, в различных видах хозяйственной деятельности, одной из главных целей является развитие и улучшение результатов. Проведение различных научных исследований с целью разработки прогноза помогает в развитии процессов в любой области применения. Для составления прогнозов довольно часто используются исследования временных рядов. К примеру, исследования временных рядов в поведении различных экономических систем позволяет анализировать и прогнозировать их дальнейшее изменение и влиять на него.

Прогнозирование временных рядов является очень важной областью машинного обучения, поскольку существует множество задач прогнозирования, которые включают временную составляющую. Примерами являются прогноз цены закрытия акций или прогноз продаж компании.

Временной ряд – это последовательность значений, описывающих протекающий во времени процесс, измеренных в последовательные моменты времени, обычно через равные промежутки [1]. Для выявления структур временных рядов существует множество математических моделей. Регрессионный анализ – метод исследования влияния независимых переменных на зависимую переменную. Это самый распространенный метод прогнозирования временных рядов. Существуют методы прогнозирования рядов с помощью искусственных нейронных сетей, нейронных сетей RNN, LSTM, методом ARIMA [2] и многие другие.

Каждый из рассмотренных методов имеет свои достоинства и недостатки, они систематизированы в Таблице 1.

Таблица 1 – Сравнение моделей и методов прогнозирования
Table 1 – Comparison of models and forecasting methods

Модель и метод	Достоинства	Недостатки
Регрессионные модели и методы	простота, гибкость, прозрачность моделирования; единообразие анализа и проектирования	сложность определения функциональной зависимости; трудоемкость нахождения коэффициентов зависимости; отсутствие возможности моделирования нелинейных процессов (для нелинейной регрессии)
Авторегрессионные модели и методы	простота, прозрачность моделирования; единообразие анализа и проектирования; множество примеров применения	трудоемкость и ресурсоемкость идентификации моделей; невозможность моделирования нелинейностей; низкая адаптивность

Модель и метод	Достоинства	Недостатки
Модели и методы экспоненциального сглаживания	простота моделирования; единообразие анализа и проектирования	недостаточная гибкость; узкая применимость моделей
Нейросетевые модели и методы	нелинейность моделей; масштабируемость, высокая адаптивность; единообразие анализа и проектирования; множество примеров применения	отсутствие прозрачности; сложность выбора архитектуры; жесткие требования к обучающей выборке; сложность выбора алгоритма обучения; ресурсоемкость процесса обучения
Модели и методы на базе цепей Маркова	простота моделирования; единообразие анализа и проектирования	невозможность моделирования процессов с длинной памятью; узкая применимость моделей
Модели и методы на базе классификационно-регрессионных деревьев	масштабируемость; быстрота и простота процесса обучения; возможность учитывать категориальные переменные	неоднозначность алгоритма построения дерева; сложность вопроса останова

Точность модели не введена в достоинства или недостатки моделей, так как зависит не только от модели, но и от опыта исследователя, набора данных и других параметров.

Целью настоящего исследования является обзор концепции по созданию модифицированного метода прогнозирования временных рядов ориентированного на интеграцию с методами кластеризации в режиме обучения реального времени. Подобный подход дает возможность компенсировать недостатки одних моделей при помощи достоинств комбинации видов анализа больших данных, и направлен на повышение точности прогнозирования, как одного из главных критериев эффективности модели. Основой метода является привязка различных событий к временному ряду.

Материалы и методы

В качестве метода кластерного анализа для модифицированного метода выберем метод «К-ближних соседей». Этот метод является популярным и несложным в реализации, однако при большой разрежённости входных данных результат имеет повышенную погрешность и поведение метода становится нестабильным [7].

Для получения обучающих данных метод предполагает сбор данных с новостных сайтов, сайтов социальных сетей, экономических порталов, торговых площадок и прочих открытых ресурсов. Анализ финансовых данных – сложная задача, они имеют случайную природу и могут вести себя крайне непредсказуемо.

Методы кластеризации могут применяться, как на структурированных данных, так и на неструктурированных данных, таких как текстовые документы, графические изображения, аудио и видеозаписи. Одной из востребованных задач является

классификация текстовых документов и сообщений. Целью классификации текстовых сообщений является обработка больших неструктурированных данных, извлечение значимых семантических аспектов из текста и предоставление резюме текста в удобочитаемом формате.

Для нашего метода будем считать, что мы используем один источник данных (например: официальный сайт новостей) для получения данных о событиях, и одну торговую площадку для получения данных значений. В результате для каждого события образуется вектор ключевых слов.

Для разрабатываемой модели важны количественные признаки для определения типа события. Количественным является признак, отдельные варианты которого имеют числовое выражение и отражают размеры, масштабы изучаемого объекта или явления. К количественным признакам, например, относятся доход домохозяйства, площадь жилого помещения, цена товара, стаж работы. Количественные признаки в статистике преобладают над другими видами признаков, они наиболее информативны, именно на работу с данными признаками нацелена большая часть многообразного статистического инструментария [6].

Сначала проводим кластеризацию полученных сообщений с помощью метода «K-ближних соседей». Тип события – это один кластер. Количество типов событий указывается при запуске алгоритма кластеризации, либо вычисляется средствами глубокого обучения с помощью среднего семантического подобия текстовых векторов. Типов событий не может больше количества самих событий, это параметр, который нужно будет подобрать для получения наилучшего результата. Для любых окружающих нас объектов и явлений можно выделить достаточно большое число признаков, которые будут определять тип события.

Одной из задач является определение времени действия события. Нужно определить категорию важности события и в зависимости от нее выставлять конечную дату.

Далее по соответствующей каждой записи дате, нужно получить значение на графике временного ряда. Временной ряд состоит из точек – транзакций. Нужно считать начальное и конечное значение. Для определения восходящий это или нисходящий график, полезно будет считать максимальное и минимальное значение.

Примерный тип записи представлен в Таблице 2.

Таблица 2 – Тип данных в БД
Table 2 – the type of data in the database

BegDate	01.10.2018
EndDate	01.11.2018
BegPrice	80$
EndPrice	70$
Average	75$
Max	80$
Min	70$
Source1	[1,2,3]
Source2	[1,2,3]
SrcN	[1,2,3]
Result	-10$
KeyWords	[смена, управляющий, персонал]

BegDate – Дата начала события.
EndDate – Дата окончания события.
BegPrice – Показатель на начало даты.
EndPrice – Показатель на конец даты.
Average – Средний показатель.
Max – Максимальный показатель.
Min – Минимальный показатель.
Source 1,2,N – Набор событий за период.
Result – Изменение показателя.
KeyWords – Вектор ключевых слов.

Имея большие объемы данных, производим обучение системы. Для получения более точных результатов нужно максимально увеличить количество данных на этапе обучения, тем самым увеличиваем точность дальнейшей работы модели прогноза.

На Рисунке 1 приведен пример временного ряда – цены акций виртуальной компании. Есть информация, что в октябре 2018 года в этой компании произошел ряд изменений, одно из которых – смена генерального директора. На графике видно, что после таких изменений цена акции начала стремительно падать, и за 2 месяца упала на 34 %. Подобные ситуации наблюдаются во многих случаях, и идея в том, чтобы предугадывать поведение графика по случившимся событиям.

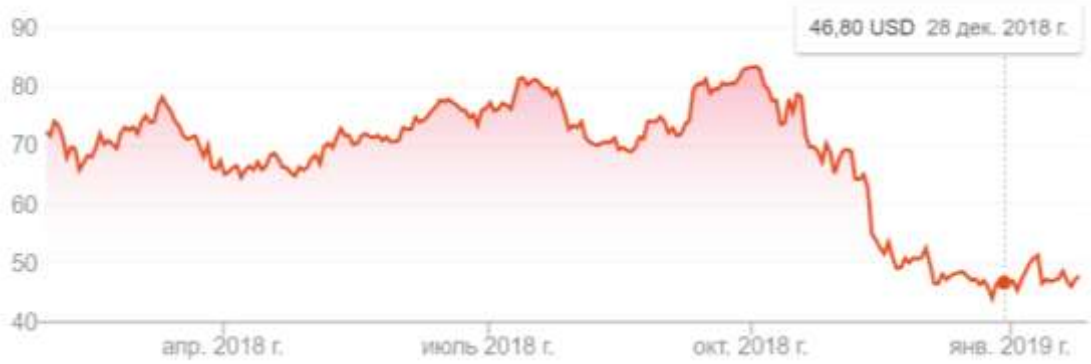

Рисунок 1 – Динамика цен на акции
Figure 1 – Dynamics of stock prices

Падение цены можно было спрогнозировать. Упрощенный алгоритм процесса прогнозирования выглядит следующим образом:

1. Получение текущих событий из открытых источников.
2. Решение задачи кластеризации методов «K-ближних соседей».
3. Получение данных об аналогичных событиях и вычисление средней статистической вероятности изменения показателя (в зависимости от заданных параметров расчетов).

Пусть временной ряд содержит дискретные значения, которые характеризуют спад, стабильное положение и подъем цены акций. Кластером будет являться последовательность:

$$Z_i^M = Z(i), Z(i+1), Z \dots, Z(i+M), \quad (1)$$

для i = 1,2,...,N-M, где N – число доступных отчетов временного ряда Z(t).

Для определения прогнозного значения рассмотрена последняя доступная информация, а именно последовательность:

$$Z(N,M) = Z(N-M+1), Z(N-M+2), \dots, Z(N), \quad (2)$$

для которой определена ближайшая похожая

$$Z(Q,M) = Z(Q+1), Z(Q+2), \dots, Z(Q+M), \quad (3)$$

При этом функция, определяющая близость, имеет вид

$$F(N-M,Q) = \sum_{j=1}^{M} |Z(N-M+1) - Z(Q+1)| \quad (4)$$

Далее вычисляется прогнозное значение:

$$Z(N+1) = Z(Q+M+1) \quad (5)$$

Схема метода сбора и кластеризации событий отображена на Рисунке 2.

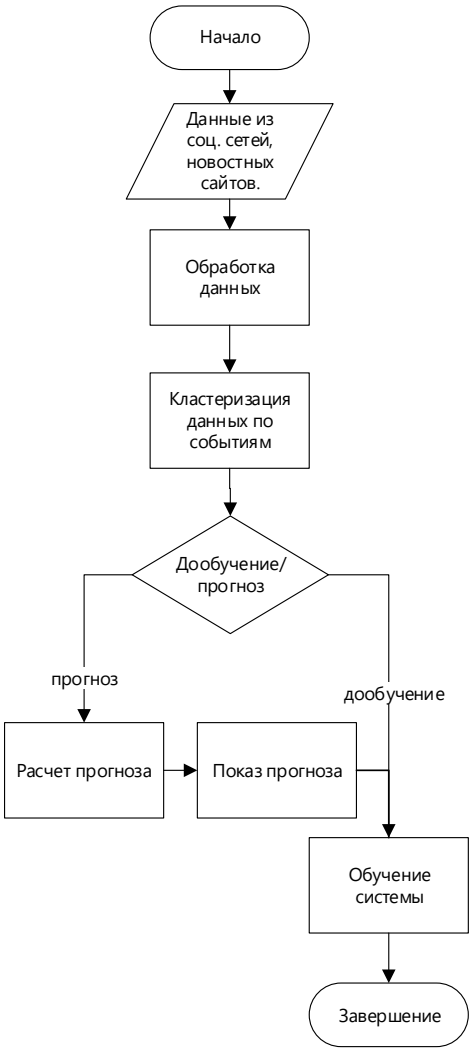

Рисунок 2 – Схема работы метода
Figure 2 – Scheme of the method

Таким образом, модифицированная модель использует прогнозирование временных рядов с использованием алгоритма кластерного анализа.

Результаты и обсуждения

Описанная методика упрощает нахождение зависимостей изменения показателей от случившихся событий. Алгоритм показывает наилучшие результаты обучения, на больших массивах данных обучения. На Рисунке 3 показан ожидаемый результат, зеленый график – оригинальные транзакции, синий график – примерные транзакции после использования системы прогнозирования. Разброс цены стабилизируется, исключаются сильные перепады.

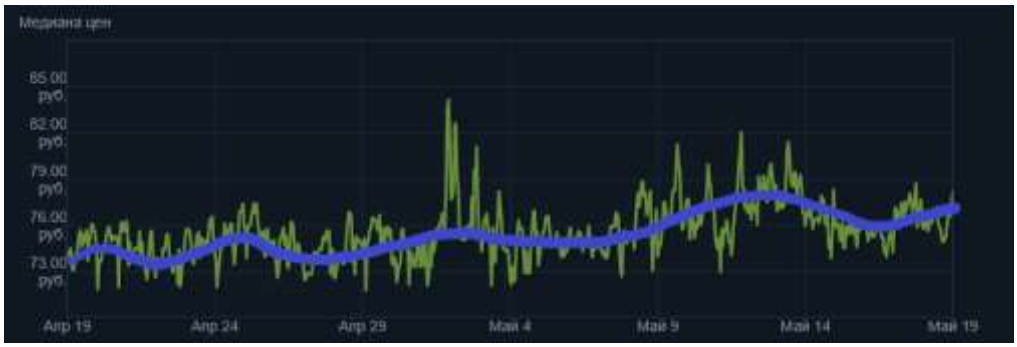

Рисунок 3 – Ожидаемые результаты поведения графиков
Figure 3 – The expected results of the behavior of the graphs

У данного метода довольно широкий круг применения, он позволяет находить зависимости между различными событиями без участия пользователя или эксперта. В концепции метода описанной в статье для построения прогноза используется только текстовая информация. В качестве расширения типов входных данных можно использовать данные с видеокамер с привязкой ко времени, графическую информацию с привязкой ко времени или с координатами геолокации. Данные типы информации обработать будет сложнее, чем текст, но теоретически это возможно.

Положительной стороной данного метода является то, что благодаря привязке к реальному времени, его повторное обучение будет происходить непрерывно на массивах актуальных данных.

Метод можно использовать для анализа тренда и сезонности. Большинство регулярных составляющих временных рядов принадлежит к двум классам: они являются либо трендом, либо сезонной составляющей. Тренд представляет собой общую систематическую линейную или нелинейную компоненту, которая может изменяться во времени. Сезонная составляющая – это периодически повторяющаяся компонента. Оба эти вида регулярных компонент часто присутствуют в ряде одновременно. Например, продажи компании могут возрастать из года в год, но они также содержат сезонную составляющую (как правило, 25 % годовых продаж приходится на декабрь и только 4 % на август) [9].

Заключение

Задача прогнозирования временных рядов имеет высокую актуальность для многих предметных областей и является неотъемлемой частью повседневной работы многих компаний. Результатом работы системы прогнозирования временных рядов будет возможность стабилизация показателей в сфере маркетинга. Система может найти

свое применение в финансовых учреждениях и банках, в различных торговых компаниях для анализа спроса на товары и услуги и анализ возможностей изменения их цены. Система позволит производить прогнозирование показателей, получаемых с помощью социологических исследований, которые позволят изучать поведение людей в обществе. Одной из сложных прикладных проблем, для решения которой можно использовать данную систему, является проблема банкротства предприятий. Данная проблема имеет высокий уровень несбалансированности распределения данных, с чем прекрасно справляются метод прогнозирования временных рядов.

Для реализации концепции данной системы одним из ключевых моментов будет являться подбор параметров для обучения системы. Подбор должен осуществляться опытным путем, что позволит увеличить точность прогнозирования и избежать серьезных ошибок. В системе требуется предусмотреть защита от переобучения. При переобучении система в определенных ситуациях будет предоставлять не правильный прогноз, из-за наличия аналогичного события.

ЛИТЕРАТУРА

1. Халафян А.А. *STATISTICA 6. Статистический анализ данных*. М.: Бином, 2010:512.
2. Трофимов П.Ю., Носков В.Ю. Прогнозирование временных рядов методом ARIMA. *Теплотехника и информатика в образовании, науке и производстве: сборник докладов VI Всероссийской научно-практической конференции студентов, аспирантов и молодых учёных (ТИМ'2017) с международным*. 2017:260–262.
3. Безручко В.П., Смирнов Д.А. *Математическое моделирование хаотических временных рядов* Саратов: Гос УНЦ «Колледж». 2005:532.
4. Manyika J., Chui M., Brown B. Big data: *The next frontier for innovation, competition, and productivity*. McKinsey Global Institute, 2011:156.
5. Шмойлова Р.А. *Общая теория статистики*: Учебник. М.: Финансы и статистика, 2008:296.
6. Замятин А.В. *Введение в интеллектуальный анализ*. Учебное пособие. 2016:120.
7. Анализ временных рядов. *Электронный учебник по статистике*. [Электронный ресурс] StatSoft – Москва, 2009 – Режим доступа: http://statsoft.ru/home/textbook/modules/sttimser.html.
8. Cao L., Soofi A. Nonlinear deterministic forecasting of daily dollar exchange rates. *International journal of forecasting.* 1999;15:421−430.
9. Meese R., Rogoff K. Empirical exchange rate models of the seventies: do they fit out-of-sample? *Journal of international economics*. 1983;14:3−24.

REFERENCES

1. Halafyan A.A. *STATISTICA 6. Statistical analysis of data*. M .: Binom. 2010:512.
2. Trofimov P.Yu., Noskov V.Yu. Forecasting time series by the ARIMA method. *Heat engineering and computer science in education, science and production: a collection of reports of the VI All-Russian Scientific and Practical Conference of students, graduate students and young scientists (TIM'2017) with international* 2017:260–262.
3. Bezruchko V.P., *Smirnov D.A. Mathematical modeling of chaotic time series Saratov*: State Scientific Center «College». 2005:532.
4. Manyika J., Chui M., Brown B. *Big data: The next frontier for innovation, competition, and productivity*. McKinsey Global Institute, 2011:156.
5. Shmoilova R.A. *General theory of statistics: Textbook.* M.: Finance and Statistics. 2008:296.
6. Zamyatin A.V. *Introduction to Mining. Tutorial*. 2016:120.

7. Time series analysis. *Electronic textbook on statistics.* [Electronic resource] StatSoft - Moscow, 2009. Access mode: http://statsoft.ru/home/textbook/modules/sttimser.html.
8. Cao L., Soofi A. Nonlinear deterministic forecasting of daily dollar exchange rates. *International journal of forecasting.* 1999;15:421−430.
9. Meese R., Rogoff K. Empirical exchange rate models of the seventies: do they fit out-of-sample? *Journal of international economics*. 1983;14:3−24.

ИНФОРМАЦИЯ ОБ АВТОРЕ / INFORMATIONS ABOUT AUTHORS

Колесников Илья Николаевич, ООО «КСК Технологии», младший разработчик, аспирант, Пензенский государственный университет, Пенза, Российская Федерация.

Ilya N. Kolesnikov, OOO «KSK technology», junior developer, graduate student, Penza State University, Penza, Russian Federation.