DOI 10.37882/2223–2966.2020.11.06

ОБРАБОТКА ТРАФИКА ЧИСЛОВЫХ ДАННЫХ С ЦЕЛЬЮ ПРОГНОЗА ЕГО РАЗВИТИЯ

Батурин Дмитрий Сергеевич
Аспирант, Амурский государственный университет, г. Благовещенск
dbat2@mail.ru

Аннотация. Рассматриваются способы прогнозирования развития временных рядов данных связанных с реальными процессами. Выдвигается гипотеза о возможности повышения качества прогнозирования сложных процессов при совместном использовании численных методов и методов искусственного интеллекта.

Ключевые слова: гибридные интеллектуальные системы, численные методы, сетевые атаки, обнаружение аномалий, обнаружение вторжений, обнаружение злоупотреблений, сетевой трафик.

PROCESSING OF NUMERICAL DATA TRAFFIC FOR THE PURPOSE OF FORECASTING ITS DEVELOPMENT

D. Baturin

Summary. Methods of forecasting the development of time series of data associated with real processes are considered. A hypothesis is put forward about the possibility of predicting complex processes when combining the recommended numerical methods and artificial intelligence methods.

Keywords: hybrid intelligent systems, numerical methods, network attacks, anomaly detection, intrusion detection, abuse detection, network traffic.

Введение

При рассмотрении трафика числовых данных подразумевается, что такой трафик генерируется каким-либо объектом (процессом) интересующим исследователя. В качестве примеров исследуемых объектов (ИО) можно привести финансовые показатели [1], [2], трафик информационных сетей [3], [4], [5] или транспортный трафик [6].

Основной целью моделирования и идентификации временных рядов является прогнозирование развития состояния наблюдаемого объекта для принятия решения о необходимых действиях в текущий момент времени направленных либо на стабилизацию состояния и предотвращение нежелательных последствий, либо на увеличение эффективности от функционирования наблюдаемого объекта [5], [7], [8]. Необходимость решения подобных задач существует в любой сфере деятельности в постоянном режиме. При этом моделирование применяется к потоку информации, обладающей характеристиками «понятными» для системы прогнозирования. Такой поток структурированной информации в случае компьютерной системы представляет собой набор цифровых значений параметров выделенных в качестве важных параметров, описывающих состояние наблюдаемого объекта [3]. При этом процесс выделения и синтеза таких параметров является дополнительной сложной проблемой при решении основной задачи прогнозирования.

Модели и методы прогнозирования

Для рассмотрения обработки трафика числовых данных с целью прогноза его развития разделим понятия модели прогнозирования и метода прогнозирования [5].

Модель прогнозирования (МоП) в общем виде это представление об исследуемом объекте, согласно которому можно описать все значимые для исследователя состояния объекта. На основании такого представления можно сделать прогноз о будущих состояниях ИО. Описание состояния ИО основано на наборе параметров значимых с точки зрения исследователя, исходные параметры представляют собой трафик, который генерируется объектом (датчиками, контролирующими его текущее состояние) [9].

Метод прогнозирования (МеП) состоит из последовательности действий, которые позволяют на основа-

нии МоП (в некоторых случаях нескольких МоП) сделать прогноз об будущих состояниях ИО [10].

В объективной реальности создание односложной МоП в отношении реальных неуправляемых исследователем ИО на неограниченном промежутке времени с приемлемой точностью относится к нереальной задаче. В противном случае можно было бы говорить о полной предсказуемости окружающего мира. Несмотря, на отсутствие возможностей полного и всеобъемлющего прогнозирования для большинства реальных объектов исследования существует необходимость такого прогнозирования по следующим причинам: предотвращения нежелательных последствий, повышение эффективности деятельности. Объекты исследования (ОИ) можно градировать по степени предсказуемости по сложности построения моделей прогнозирования и длительности их справедливости. Ограниченность верности конкретной модели прогнозирования приводит к необходимости коррекции с течением времени отдельных показателей и коэффициентов в конкретной модели, а также часто к необходимости смены модели прогнозирования для одного и того же объекта при изменениях обстоятельств.

Существующие математические модели прогнозирования

К основным моделям прогнозирования относят [3], [9], [11]: регрессионные модели (regression model), авторегрессионные модели (auto regressive model, AR), нейросетевые модели (artificial neural network, ANN), модели экспоненциального сглаживания (exponential smoothing, ES), модели на базе цепей Маркова (Markov chain), регрессионные деревья (classification and regression trees, CART), метод опорных векторов (support vector machine, SVM), генетические алгоритмы (genetic algorithm, GA), модель на основе передаточных функций (transfer function, TF), формализованная нечеткая логика (fuzzy logic, FL), фундаментальные модели.

Примем прошлые и доступные значения временного ряда доступны в дискретные моменты времени $t = 1,2,...T$. В момент времени T необходимо определить значения процесса $Z(t)$ в моменты времени $T+1,...,T+P$. Обозначим временной ряд значений параметра $Z(t) = Z(1), Z(2),...,Z(T+P)$. Известными (доступными) значениями являются $Z(t) = Z(1), Z(2),...,Z(T)$. Момент времени T называется моментом прогноза, а величина P — временем упреждения [8].

Общая для всех моделей постановка задачи прогнозирования имеет следующий вид (1):

$$Z(t)=F(Z(t\text{-}k\text{-}n),\ ...Z(t\text{-}k),\ X_1(t\text{-}k\text{-}n),\ ...X_1(t\text{-}k),\ ...X_s(t\text{-}k\text{-}n),\ ...X_s(t\text{-}k)\ )+ \quad (1)$$

где $Z(t)$ — это набор прогнозируемых значений на основании предыдущих фактических значений прогнозируемых параметров $Z(t\text{-}k\text{-}n),\ ...Z(t\text{-}k)$, а также на основании измеримых значениях параметров внешних воздействий $X_1(t\text{-}k\text{-}n),\ ...X_1(t\text{-}k),\ ...X_s(t\text{-}k\text{-}n),\ ...X_s(t\text{-}k)$; ε_t — это разница между расчётными прогнозными показателями и фактическими измеренными после наступления момента времени соответствующего $Z(t)$; n — количество измеренных значений используемых для прогноза; k — значение, определяемое либо из степени сходства динамики трафика либо из соображений выявления значимой динамики, при этом часто $k = 0$.

В некоторых случаях внешние воздействия не учитываются и тогда задача прогнозирования приобретет следующий вид:

$$Z(t)=F(Z(t\text{-}k\text{-}n),\ ...Z(t\text{-}k))+\varepsilon_t \quad (2)$$

При таких исходных условиях и постановке задачи, самой простой и очевидной функцией, которую необходимо оптимизировать, является выражение, которое вычисляет среднее абсолютное отклонение истинного значения от прогнозируемого, и результат такого выражения стремится к минимуму при заданном P (3).

$$E=\frac{1}{P}\sum_{t=T+1}^{T+P}|\varepsilon_t| \to \min \quad (3)$$

Значение ε_t при качественном прогнозе не должно выходить за размеры приемлемого диапазона до достижения времени t, то есть фактическое $Z(t) \in [Z(t)-a, Z(t)+b]$, где a и b приемлемые значения заданные исследователем, при этом $\varepsilon_t < a$, $\varepsilon_t < b$. Промежуток времени от $T-1$ до T не обязательно равен промежутку от T до $T+1$, то же справедливо и для других дискретных значений моментов времени. В общем случае неравенство промежутков времени объясняется разной интенсивностью трафика. Интенсивность трафика характеризуется объемом событий формирующих показатели и размерами изменений в показателе и другими. Поэтому одной из задач является задача определения размеров таких промежутков.

Согласно общей постановке интересно выделить регрессионные и авторегрессионные модели прогнозирования [12].

В регрессионных моделях будущие значения параметра связаны только с внешними факторами $X(t)$, в этом случае крайне важна предсказуемость внешних факторов или высокая корреляция прогнозных значений прогнозируемого параметра с прошлыми известными значениями внешних факторов. В противном случае ис-

пользование таких моделей прогнозирования в чистом виде, когда значения $X(t)$ неизвестны, не имеет смысла.

Авторегрессионные модели предполагают, что значение процесса линейно зависит от некоторого количества предыдущих значений того же процесса $Z(t)=F(Z(t-k-n), ...Z(t-k))$. Часто именно в такой ситуации необходимо принимать решения, например финансовые рынки в отсутствии важных новостей.

Согласно видам авторегриессионных моделей самый простой способ прогноза состояния трафика — это предположение что будущие значения трафика будут близки к средним показателям прошлых периодов. В этом случае достаточно выбрать период усреднения, при этом необходимо учитывать, что увеличение такого периода увеличит разницу между текущими и будущими ближайшими значениями параметров, а уменьшение периода усреднения внесет неприемлемую случайность в результат усреднения. При этом, реальные процессы всегда приобретают направленные изменения под воздействием внешней среды. Направление таких изменений сопровождаются исходными изменениями на начальном этапе, которые затем переходят в более значительные изменения в соответствии с начальными изменениями. Такие исходные изменения могут быть малозаметными по сравнению с будущим развитием, в противном случае, когда изменения будут видны, прогноз этих изменений уже устареет. Таким образом, необходимо идентифицировать в малозаметных изменениях прогнозируемых параметрах «предвестников» будущих значительных изменений, которые необходимо прогнозировать. Кроме того, изменению прогнозируемых параметров может предшествовать значительные изменения второстепенных показателей, прогнозирование, которых не является целью обработки трафика (объемы торгов, например), но по изменению которых можно предсказать, например, не направление, но размер изменений прогнозируемых параметров.

Как можно заметить во всех рассуждениях встречается понятие изменений, то есть первая производная от скользящей средней. Именно величина и направленность предшествующих изменений часто помогает при прогнозировании определить величину и направление будущих изменений.

Таким образом, из, например, скользящей средней [8] можно извлечь ряд дополнительных показателей, которые помогут спрогнозировать дальнейшие изменения основного наблюдаемого и исследуемого показателя. Определение последовательности и условий использования тех или иных вычисляемых на основе моделей показателей осуществляется с помощью методов.

Общепринятые численные методы (обзор поиска коэффициентов)

Понятие «*метод прогнозирования*» гораздо шире понятия «модель прогнозирования». Можно провести следующую классификацию методов прогнозирования: интуитивные методы (основываются на мнении эксперта), формализованные методы (основываются на математических моделях).

Интуитивные методы прогнозирования основываются на суждениях и оценках экспертов. В электронном компьютерном виде могут быть реализованы в виде экспертных систем, содержащих базы данных со значениями параметров и выводами экспертов на основании конкретных значений или их диапазонов. В свойствах таких методов отсутствует возможность математического описания и, поэтому, без участия эксперта полностью отсутствует способность к адаптации к неизвестным ситуациям [2].

Формализованные методы основаны на использовании известных моделей прогнозирования с дополнительной обработкой полученных результатов. Дополнительная обработка результатов может осуществляться с применение других математических моделей, дополняющих модель, используемую на первом этапе обработки. Иначе говоря, это сложный прием, упорядоченная совокупность простых приемов, направленных на разработку прогноза в целом; путь, способ достижения цели, исходящий из знания наиболее общих закономерностей.

Для целей данной статьи интерес представляют формализованные методы, которые можно реализовать в виде программного обеспечения.

В формализованных методах можно выделить следующие: метод экстраполяции; тренд-анализ; интерполяция; сценарии; «прогнозы до абсурда»; факторный анализ; распознавание образов; вариационное исчисление; спектральный анализ; алгебра логики; теория игр; другие.

Далее классифицируем методы прогнозирования. Разделение на классы формализованных методов, основанных на математических моделях, можно провести по степени универсальности в различных предметных областях: модели предметной области; модели временных рядов.

В моделях предметной области используются зависимости, свойственные конкретной предметной области. Такого рода моделям свойственен индивидуальный подход в разработке и неприменимость к другим областям.

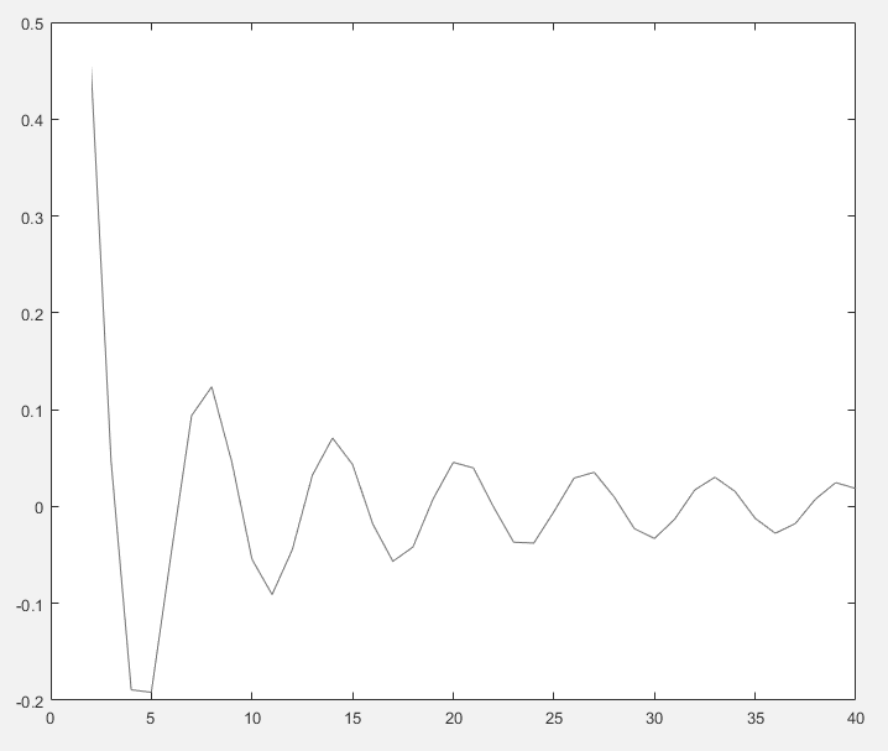

Рис. 1. Принятая истинная зависимость временного ряда.

В модели временных рядов используются универсальные математические модели прогнозирования, которые стремятся найти зависимость будущих значений от известных показателей, в основном от прошлых значений прогнозируемого показателя.

Предполагается, что существует функциональная связь между признаками и основным свойством (неизвестная пользователю).

В указанных классах рассмотрим модели временных рядов, и дополнительно их классифицируем в связи с тем, что они лишены несопоставимых характеристик, связанных с конкретными сферами их применения, в отличие от моделей предметной области.

В моделях временных рядов можно выделить: статистические модели и структурные модели прогнозирования.

В статистических моделях используются хорошо зарекомендованные методы статистической обработки данных

Структурные модели основаны на выражении зависимости будущих значений от прошлых через структуру и правил перехода по ней. К структурным моделям из перечисленных ранее относятся такие как нейросетевые модели, генетические алгоритмы и т.д.

Прогнозирование — это процесс построение предсказания будущего на основе исторических данных, текущих данных (текущей ситуации) и на основе анализа внешних воздействий. Риск и неопределенность являются центральными факторами для прогнозирования, поэтому в соответствии с лучшими практиками, необходимо указывать степень неопределенности по отношению к прогнозам.

Корректный подход к оценке метода прогнозирования включает несколько этапов. Следует выделить пять важных этапов: изучение природы исследуемого объекта или процесса для выбора адекватного метода прогнозирования; выделение двух групп среди доступных данных — для разработки прогнозов и для проверки полученных результатов; уточнение исходных данных с целью обнаружения ошибок; разработка прогнозов и оценка достоверности полученных результатов; использование (интерпретация) полученных результатов и выполнение, при необходимости, уточнения и дополнения прогнозов [13], [14].

Таким образом, методы прогнозирования можно характеризовать по следующим признакам: временной охват (горизонт прогнозирования) — краткосрочные, среднесрочные, долгосрочные; типы прогнозирования — экстраполятивное, альтернативное; степень вероятности будущих событий — вариантные, инвариантные; способ представления результатов прогноза — точечные, интервальные.

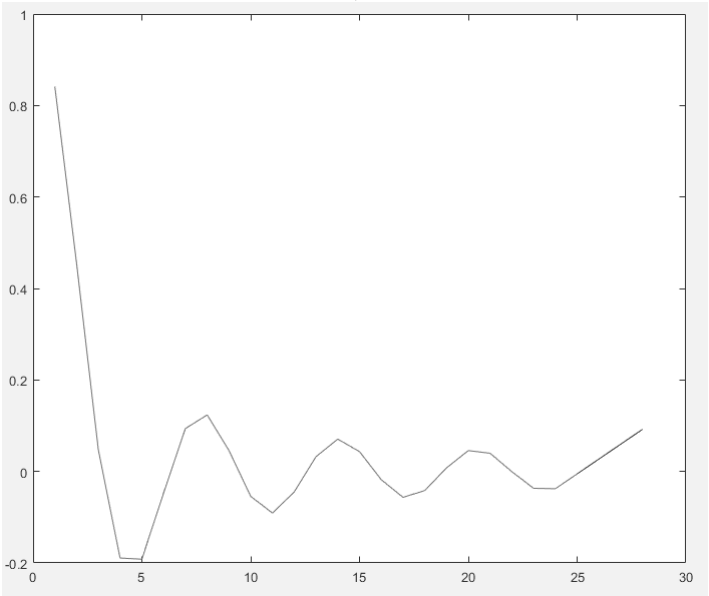

Рис. 2. Линейная интерполяция

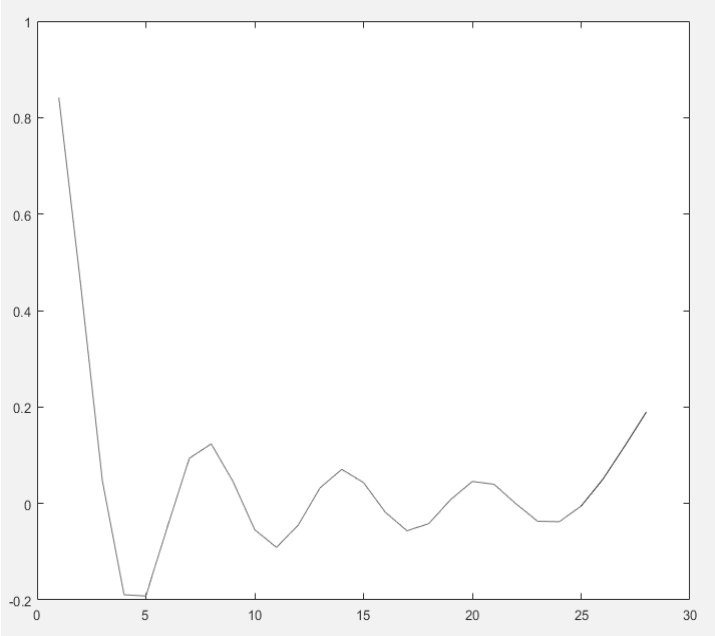

Рис. 3. Интерполяция кубического сплайна.

Специализированные программные продукты

Удовлетворительная точность прогнозирования может быть достигнута при использовании принятых моделей и универсальных аналитических программных пакетов, таких как MatLab, MathCad, Statistica, языки программирования R и Python.

Можно выделить следующее ПО использующее в своей работе прогнозирование: Sales-Forecast, Metatrader 4, Metatrader 5, системы обнаружения вторжений в информационных сетях, и многие другие.

В перечисленном ПО создаются программные модули, алгоритмы которых предварительно будут отработаны с использованием специализированного для конкретной сферы применения ПО.

Для примера рассмотрим прогнозирование в уже ставшим классическом пакете MatLab. В среде MatLab прогнозирование развития временного ряда интер-

Таблица 1. Результаты прогнозирования

Типы временного ряда	Номера моментов времени				
	25	26	27	28	E
Истинные значения	-0,00529	0,029329	0,035421	0,009675	
Прогноз в соответствии с линейной интерполяцией	-0,00529	0,027144	0,059583	0,092021	0,027173
Прогноз в соответствии с интерполяцией кубического сплайна	-0,00529	0,050289	0,118781	0,18995	0,071148
Прогноз в соответствии с кусочно-кубической интерполяцией Эрмита	-0,00529	0,029025	-0,02927	-0,27467	0,087336

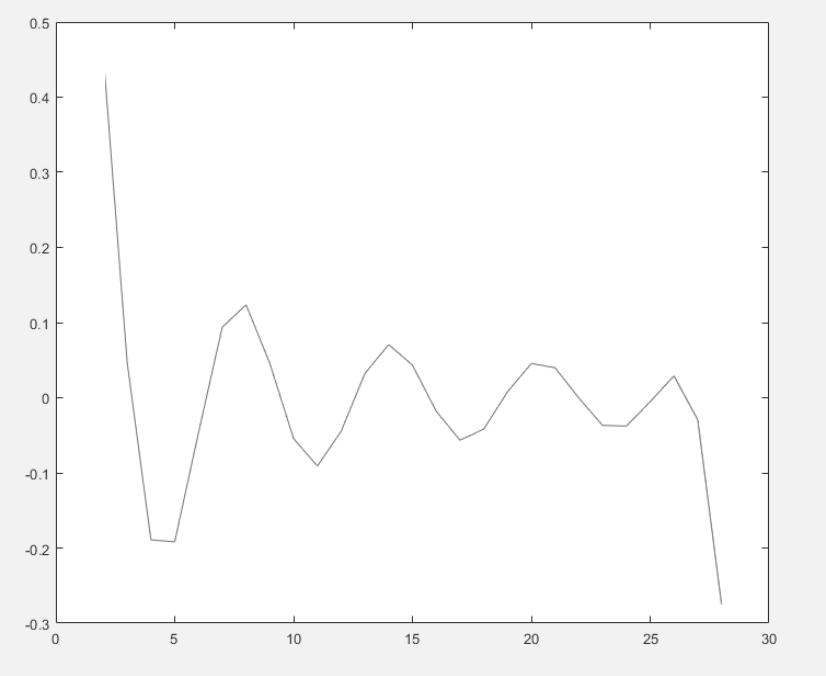

Рис. 4. Кусочно-кубическая интерполяция Эрмита

претируется как экстраполяция данных. Специализированная функция экстраполяции в среде MatLab отсутствует, экстраполяция данных на стандартной основе в выбранном инструменте осуществляется с использованием функции интерполяции. Методы интерполяции (экстраполяции) указываются в качестве параметров функции.

В качестве временного ряда для примера возьмем ряд образуемый синусоидой с уменьшением ее амплитуды (рис. 1).

Для оценки качества прогноза, который может быть обеспечен с использованием принятого в примере инструмента и его функции экстраполяции. В используемой функции интерполяции используем параметр, который задействует модели линейной интерполяции, интерполяции кубического сплайна и кусочно-кубической интерполяции Эрмита. Прогнозируемый ряд ограничим значениями, соответствующих отрезку моментов времени с индексами 1–25. Прогноз сделаем для отрезка моментов времени 26–28. Результаты представлены на рисунках 2–4.

Из графиков можно сделать вывод, что наиболее качественно прогноз сделан по модели кусочно-кубической интерполяции Эрмита. Для более точной оценки проведем расчет средних абсолютных отклонений истинного значения от прогнозируемого E. Результаты прогнозного расчета представлены в табл. 1.

Рассчитанные значения E показали, что наиболее приближенные значения на выбранном отрезке показал метод линейной интерполяции. Визуальная ошибка объясняется тем, что направленность изменений от пред-

сказанного значения 26 к 27 соответствует направлению изменений истинных данных.

Заключение

В ходе исследования были рассмотрены различные модели и методы для решения задачи прогнозирования развития временных рядов. В результате анализа были классифицированы методы и модели прогнозирования, а также сделан вывод об отсутствии отдельного приемлемого метода или модели для решения задачи прогнозирования развития временного ряда на достаточно высоком уровне в различных предметных областях.

Также были рассмотрены различные программные инструменты, реализующие прогнозирование развитие временного ряда. В результате проверки возможностей прогноза программных инструментов можно сделать вывод о том, что даже для небольших периодов прогнозирования и хорошо прогнозируемого временного ряда простых решений нет.

Следующими шагами исследования должны стать: исследования временных рядов с использованием программных инструментов с целью разработки эффективного гибридного метода прогнозирования, включающего несколько реализованных отдельных методов прогнозирования.

ЛИТЕРАТУРА

1. Тихонов Э. Е. Прогнозирование в условиях рынка. Невинномысск, 2006. 221 с.
2. Бурда А.Г., Бурда Г. П. Экономико-математические методы и модели: учеб. пособие (курс лекций); Кубан. гос. аграр. ун-т.— Краснодар, 2015. 178 с.
3. Батурин Д. С. Классификация параметров используемых для прогнозирования временных рядов в гибридных интеллектуальных системах // Modern Science, Издательство: научно-информационный издательский центр «Институт стратегических исследований» (Москва). 2019. № 5–2, С. 179–182.
4. Батурин Д. С., Анализ методов обнаружения атак в информационных сетях // Вестник АмГУ. 2019. N87, с. 54–59
5. Шелухин О. И., Филинова А.С, Васина А. В. Обнаружение аномальных вторжений а компьютерные сети статистическими методами // T-Comm: Телекоммуникации и транспорт. 2015. Том 9. N»10, С. 42–49.
6. Жанказиев С. В., Воробьев А. И., Шадрин А. В., Гаврилюк М. В. Имитационное моделирование в проектах ИТС: учебное пособие: под ред. д-ра техн. наук, проф. С. В. Жанказиева. М.: МАДИ, 2016. 92 с.
7. Яркова Т. М. Макроэкономическое планирование и прогнозирование: учебное пособие; М-во с.-х. РФ; «Пермский гос. аграрно-технолог. ун-т им. акад. Д. Н. Прянишникова». Пермь: ИПЦ «Прокростъ», 2018. 292 с.
8. Кувайскова Ю. Е. Статистические методы прогнозирования: учебное пособие / Ю. Е. Кувайскова, В. Н. Клячкин. Ульяновск: УлГТУ, 2019. 197 с.
9. Афанасьев В. Н., Юзбашев М. М. Анализ временных рядов и прогнозирование: Инфра-М, 2010.
10. Сухарев М. Г. МЕТОДЫ ПРОГНОЗИРОВАНИЯ: Учебное пособие для студентов специальности «Прикладная математика», РГУ нефти и газа им. И. М. Губкина Москва, 2009 г.
11. Judith Hurwitz, Daniel Kirsch, Machine Learning IBM Limited Edition,— Published by John Wiley & Sons, Inc. 111 River St. Hoboken, 2018. [Электронный ресурс]. Режим доступа: www.wiley.com.
12. Кизбикенов К. О. Прогнозирование и временные ряды; Учебное пособие Барнаул, ФГБОУ ВО «АлтГПУ», 2017.
13. Батурин Д. С. Необходимость использования методов статистики для обнаружения вторжений // Сборник избранных статей по материалам научных конференций ГНИИ «Нацразвитие» (Санкт-Петербург, Август 2019). СПб, ГНИИ «Нацразвитие», 2019. С. 84–86.
14. Батурин Д. С., Организация гибридной интеллектуальной системы для обнаружения вторжений в информационную сеть // Сборник избранных статей по материалам научных конференций ГНИИ «Нацразвитие» (Санкт-Петербург, Август 2019). СПб, ГНИИ «Нацразвитие», 2019. С. 130–132.

Журнал «Современная наука: актуальные проблемы теории и практики»