

МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное образовательное учреждение
высшего образования

**«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ИМЕНИ Н.Г. ЧЕРНЫШЕВСКОГО»**

Кафедра социальной информатики

АНАЛИЗ ВРЕМЕННЫХ РЯДОВ

(автореферат бакалаврской работы)

студента 4 курса 451 группы
направления 09.03.03 - Прикладная информатика
профиль Прикладная информатика в социологии
Социологического факультета
Никонорова Александра Анатольевича

Научный руководитель

кандидат физико-математических наук, доцент _____ М.Г. Плешаков
подпись, дата

Зав. кафедрой

кандидат социологических наук, доцент _____ И.Г. Малинский
подпись, дата

Саратов 2022

ВВЕДЕНИЕ

Актуальность проблемы: в связи с необходимостью анализа быстро растущих объемов данных, средства обработки временных рядов становятся повсеместной деталью нашей жизни, и грамотный анализ и интерпретация полученных данных становятся необходимостью.

Объект исследования: временные ряды.

Предмет исследования: анализ временных рядов.

Цель: целью выпускной квалификационной работы является разработка искусственной нейронной сети, решающую задачу анализа временного ряда, с дальнейшей интерпретацией полученных данных, а также рассмотрение социального явления, создающего аномалию в нем.

Задачи работы:

1. Определение временного ряда и его дальнейшего прогнозирования.
2. Определение основных проблем в прогнозировании временных рядов.
3. Изучение основных методов анализа временных рядов.
4. Проектирование ИНС для такого анализа.
5. Тестирование сети.
6. Анализ взаимосвязи социального явления и спрогнозированных значений временного ряда.

Теоретическая значимость исследования заключается в возможности использования основных положений и выводов данной работы для дальнейшего изучения выбранной проблематики.

Структура бакалаврской работы. Выпускная квалификационная работа состоит из введения, четырех глав по 2, 2, 4 и 2 параграфа соответственно, заключения, списка использованных источников и приложения.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

В первой главе «Временные ряды» вводятся основные определения временного ряда.

Определение «временные ряды» в статистике России пока непривычно. В зарубежной англоязычной литературе принят термин *time series*, в немецкой - *zeitreihen analyze*. Обе формулировки лучше всего передаются по-русски как временные ряды или анализ временных рядов.

Опасение, что студенты воспримут определение «временной ряд» как временный, заставило предпочесть новый для статистики и неточный, по существу, термин динамический ряд, ряд динамики.

Неточность последнего термина состоит в том, что не каждый ряд уровней за последовательные моменты или периоды времени отражает динамику какого-либо признака. Термин динамика правильнее относить к наличию тенденции рассматриваемых во времени показателей. Про ряд уровней, содержащих лишь колебания, но не имеющих надежно установленной тенденции, говорят: «В этом ряду, в данном процессе нет никакой динамики». Так можно характеризовать экономику застойного периода или же остановившееся производство из-за технических неполадок.

Следовательно, динамические ряды – определение, относящееся к тем рядам уровней, в которых находится тенденция, а временные ряды – более общее понятие, включающее как динамические, так и статические последовательности уровней какого-либо показателя.

Временной ряд — это последовательность упорядоченных во времени числовых показателей, характеризующих уровень состояния и изменения изучаемого явления.

Временные ряды различаются по следующим признакам:

1) по времени – моментные и интервальные. Интервальный ряд – последовательность, в которой уровень явления относят к результату, накопленному или вновь произведенному за определенный интервал времени.

2) по форме представления уровней – ряды абсолютных относительных и средних величин;

3) по расстоянию между датами или интервалами времени выделяют полные и неполные временные ряды. Полные ряды имеют место, когда даты

регистрации или окончания периодов следуют друг за другом с равными интервалами, неполные - когда принцип равных интервалов не соблюдается.

4) по содержанию показателей – ряды частных и агрегированных показателей. Частные показатели характеризуют изучаемое явление односторонне, изолированно. Например, среднесуточный объем выпуска промышленной продукции предоставляет возможность дать оценку динамике промышленного производства, численность граждан, состоящих на учете в службе занятости, наглядно демонстрирует эффективность социальной политики государства;

Познание закономерностей изменений во времени — сложный и трудоемкий процесс анализа, так как любое изучаемое явление образует множество факторов,двигающих систему в разных направлениях. По характеру непосредственного воздействия эти факторы можно разделить на две группы. К первой группе относятся факторы, определяющие основную тенденцию динамики (рост или снижение уровней).

Вторая группа факторов, вызывающая случайные колебания, отклоняет уровни от тенденции то в одном, то в другом направлении. Например, тенденция динамики урожайности связана с прогрессом агротехники, с укреплением экономики данной совокупности хозяйств, совершенствованием организации и управления производством. Колеблемость урожайности вызвана чередованием благоприятных по погоде и неблагоприятных лет, циклами солнечной активности, колебаниями в развитии вредных насекомых и болезней растений. При статистическом изучении динамики необходимо четко разделять два основных ее элемента - тенденцию и колеблемость, чтобы дать каждому из них количественную характеристику с помощью специальных показателей.

Статистические данные, представленные в виде временных рядов, должны быть сопоставимы по территории, кругу охватываемых объектов, единицам измерения, моменту регистрации, методике расчета, ценам, достоверности. Несопоставимость по территории возникает в результате изменений границ стран, регионов, хозяйств и т.п. Для приведения данных к сравнимому виду

производится пересчет прежних данных с учетом новых границ. Полнота охвата различных частей явления - важнейшее условие сопоставимости уровней ряда. Требование одинаковой полноты охвата разных частей изучаемого объекта означает, что уровни ряда за отдельные периоды должны характеризовать размер того или иного явления по одному и тому же кругу входящих в состав его частей. Например, при характеристике динамики урожайности овощных культур в регионе по годам нельзя в одни годы учитывать только сельскохозяйственные предприятия, а в другие - все категории хозяйств, в том числе приусадебные участки сельских жителей и сады, огороды горожан.

Среднее изменение уровней ряда за единицу времени – главный параметр и константа прямолинейного тренда. Следовательно, этот тип тренда подходит для отображения тенденции примерно равномерных изменений уровней: равных в среднем абсолютных приростов или абсолютных сокращений уровней за равные промежутки времени. Практика показывает, что такой характер динамики встречается достаточно часто. Причина близких к равномерным абсолютным изменениям уровней ряда состоит в следующем: многие явления, как, например, урожайность сельскохозяйственных культур, численность населения региона, города, сумма дохода населения, среднее потребление какого-либо продовольственного товара и др., зависят от большого числа различных факторов. Одни из них влияют в сторону ускоренного роста изучаемого явления, другие - в сторону замедленного роста, третьи - в направлении сокращения уровней и т.д. Влияние разнонаправленных и разноускоренных (замедленных) сил факторов взаимно усредняется, частично взаимно погашается, а равнодействующая их влияний приобретает характер, близкий к равномерной тенденции. Итак, равномерная тенденция динамики (или застоя) — это результат сложения влияния большого количества факторов на изменение изучаемого показателя.

Во второй главе «Основные методы анализа временных рядов» описываются главные методы анализа временных рядов.

Для начала нужно дать краткий обзор методов анализа данных, представленных в виде временных рядов, т.е. в виде последовательностей измерений, упорядоченные в моменты времени. Главное отличие от анализа случайных выборок в том, что в случае с временными рядами мы основываемся на предположении, что последовательные значения данных, которые мы наблюдаем, находятся друг от друга на равных временных промежутках.

Существуют две основные цели анализа временных рядов: определение природы ряда и прогнозирование (предсказание будущих значений временного ряда по настоящим и прошлым значениям). Обе эти цели требуют, чтобы модель ряда была идентифицирована и, более или менее, формально описана.

Итак, давайте рассмотрим основные модели прогнозирования временных рядов:

Наивное предсказание – простейший метод анализа временного ряда, который попросту предсказывает временной ряд основываясь на последних значениях.

Экспоненциальное сглаживание – классический метод анализа временного ряда и предсказания. По сути метод заключается в том что мы пропускаем определенный промежуток временного ряда через функцию сглаживания и на основе полученных параметров делает предсказание (прогнозы равны взвешенному среднему от старых наблюдений).

ARIMA, SARIMA – метод, также как и экспоненциальное сглаживание, часто используется для анализа временных рядов, более детальное название данного метода – интегрированной модель авторегрессии скользящего среднего.

NNETAR – модель, представляющая себя в виде авторегрессии нейронной сети (полносвязная нейронная сеть). Модель принимает на вход принимает на вход последние элементы последовательности и выводит прогноз на основе этих данных.

Garch – предыдущие модели анализа предполагали, что слагаемые ошибки в стохастическом процессе генерации временных рядов имели одинаковую дисперсию, в модели GARCH предполагается, что слагаемое ошибки следует

авторегрессионному скользящему среднему, следовательно слагаемое меняется по ходу времени.

Традиционные нейронные сети не обладают этим свойством, и в этом их главный недостаток. Представим, например, что мы хотим классифицировать события, происходящие в фильме. Непонятно, как традиционная нейронная сеть могла бы использовать рассуждения о предыдущих событиях фильма, чтобы получить информацию о последующих.

Основными проблемами предсказания грядущей динамики ряда, заключается во множестве таких факторов, как:

- 1) Временные ряды зачастую имеют ограниченное содержание информации.
- 2) Каждый временной ряд уникален.
- 3) Аномалии, для которых требуется проработка личного метода для анализа ряда.
- 4) Множество потенциальных подходов к анализу.
- 5) Невозможность точного предсказания на большое количество временных промежутков.

Основным фаворитом на данный момент времени являются методы анализа, построенные на ИНС. Они могут включать во внутренний слой довольно обширный спектр методов по анализу и предсказанию временных рядов, например регрессионный анализ.

В третьей главе «Взаимосвязь социальной аномалии и предсказания динамики временного ряда» описывается взаимосвязь социальной аномалии и предсказания динамики временного ряда.

Для начала введем определение аномалии во временных рядах:

Под аномалией будем подразумевать данные не похожие на остальные во временном ряду, то есть сильно отклоняющиеся от построенной тенденции и имеющее под собой внешний характер. В качестве аномалии может быть атака на информационную систему, отклонение медицинских показателей пациентов или даже финансового мошенничества.

Касательно анализа, существуют определенного вида практики, позволяющие обнаружить или предположить грядущую аномалию. В сущности, они представляют из себя объединение результатов работы нескольких методов, совершающие разные ошибки, которые взаимно корректируются при их комбинации.

Используемые для поиска аномалий во временных рядах методы принято разделять на группы:

Proximity-based: выявление аномалии на основе информации о близости параметров или последовательности параметров фиксированной длины, подходит для выявления точечных аномалий и выбросов.

Prediction-based: построение прогнозной модели и сравнение прогноза и фактической величины, лучше всего применимо ко временным рядам с выраженными периодами, циклами или сезонностью.

Reconstruction-based: методы, основанные на реконструкции фрагментов данных, используют восстановление фрагмента данных, поэтому может выявлять как точечные аномалии, так и групповые.

В мае 2021 года, основатель компании «Tesla» и «SpaceX», объявил, что Tesla приостанавливает прием оплаты за электромобили в криптовалюте «Bitcoin», а также дальнейшее сотрудничество. Причиной для отказа стала отрицательное влияние на экологию способа добычи криптовалюты. В результате стоимость биткоина упала на 15%. Через несколько дней в социальной сети «Twitter», после дискуссии Маск заявил, что в следующем квартале владельцы активов будут разочарованы, если знают, что компания «Tesla» избавилась от оставшихся у нее активов. После стоимость монеты опустилась еще на 5%.

Для анализа временного ряда, которым в данном случае у нас выступает курс стоимости крипто монеты «Bitcoin», был выбран метод анализа данных с помощью построения нейронной сети.

ИНС содержит в себе слой анализа данных в основе которого лежит линейный регрессионный анализ, после чего нейронная сеть на основе

полученных данных обучается и делает предсказания. Самый активное отклонение от тенденции наблюдается в промежутке между 12.05.2021 и 24.05.2021. В ходе работы мы сделаем предсказание на 14 дней вперед, учитывая каждый день, начиная с 4го и 10го мая 2021 года, а также предсказание на 30 дней вперед начиная с 24го апреля 2021 года. После, мы сравним полученные данные с реальными и между собой и сделаем вывод о достоверности предсказания ИНС. Также стоит учесть, что в качестве анализируемых данных для нейронной сети, будет взят ежедневный курс крипто монеты с 2017 года.

В ходе получения результатов анализа временного ряда, можно сделать вывод, что обученная ИНС предсказала социальную аномалию в данной сфере которая сильно обрушила курс стоимости крипто монеты, однако же, это не совсем так, стоит учесть что график крипто монеты «Bitcoin» часто выстраивает снижение тенденции когда его значение достигает определенного пика, и как следствие происходит дальнейшее снижение.

На нашем примере мы показали, что многие социальные явления или аномалии могут быть предсказаны и математически обусловлены, в данном случае не временной ряд изменился исходя из социальной аномалии, а наоборот, социальная аномалия возникла в тот же момент, когда график начал постепенное снижение после достижения пика. Случаи, когда аномалия случается единожды, предсказание практически невозможно определить точно, но можно предположить ее появление заранее исходя из анализа данных, к которому она принадлежит.

Погрешность, с которой ИНС предсказывает значения прямо пропорционально увеличивается относительно длины предсказания, а также становятся заметны резкие колебания, в связи с чем можно сделать вывод что чем меньше предсказываемый ряд тем оно более точное.

И так можно сделать вывод что на основе полученных данных мы можем предсказывать не только курс криптовалюты, но и также находить предпосылки для глобальных социальных явлений и аномалий, связанных непосредственно с этой темой.

В четвертой главе «Основной инструментарий для анализа временных рядов» описывается основной инструментарий для анализа временных рядов.

При решении задач прогнозирования аналитику приходится принять решения относительно таких характеристик временного ряда как тренд, сезонная и циклическая компоненты, делать предположения о модели временного ряда – аддитивной, мультипликативной и др. Автоматического способа обнаружения трендов во временных рядах не существует. В то же время при изучении кривой, отражающей результаты наблюдений, аналитику трудно делать предположения относительно повторяемости формы кривой через равные промежутки времени. Общим недостатком статистических моделей является сложность выбора типа модели и подбора ее параметров. Все это существенно увеличивает субъективный вклад участников процесса анализа и прогнозирования ВР. Таким образом, результат анализа и прогнозирования ВР зависит как от квалификации аналитика в предметной отрасли, так и от его квалификации в методах анализа.

Быстроту реакции модели на изменения в динамике процесса отображает так называемый параметр адаптации. Процесс обучения модели состоит в выборе наилучшего параметра адаптации на основе проб в ретроспективном материале. Например, воздействие мощного раздражителя в виде процесса «белого шума» с большей дисперсией вызывает у модели адекватную «оборонительную реакцию». Модель не адаптируется под к белому шуму, а наоборот, инертна по отношению к нему. При наличии тенденции в стохастическом процессе наилучшей реакцией модели является определенный компромисс между двумя крайними ситуациями, обеспечивающий отражение тенденции и одновременно фильтрацию случайных отклонений от нее. По тому, насколько хорошо модель поддается «обучению», можно судить ее способности адекватно отражать закономерности данного временного ряда.

Нейронные сети представляют собой новую и весьма перспективную вычислительную технологию, дающую новые подходы к исследованию динамических задач в финансовой области. Первоначально нейронные сети открыли новые возможности в области распознавания образов, затем к этому

прибавились статистические и основанные на методах искусственного интеллекта средства поддержки принятия решений и решения задач в сфере финансов.

Способность к моделированию нелинейных процессов, работе с зашумленными данными и адаптивность дают возможности применять нейронные сети для решения широкого класса финансовых задач.

Решение использования ИНС для анализа временного ряда обусловлено тем, что алгоритмы нейронных сетей лучше всего справляются с обработкой данных и благодаря современным методам имеют возможность обнаружить даже самые незначительные колебания ряда.

Первоначально алгоритм строится на вводе первичного временного ряда, где, мы убираем значение стоимости крипто монеты последние 14ти и 30ти дней, это делается для того чтобы в дальнейшем мы могли сравнить результаты, после алгоритм изучает оставшиеся данные и строит на их основе предсказание, на основе которого нейронная сеть обучается. Обученная искусственная нейронная сеть на основе обучения и циклического повторения делает предсказание на 14 или 30 дней вперед.

В процессе, после обучения, мы запрашиваем у нейронной сети насколько ее дальнейшее предсказание верно:

- 1) Для 14 дневного предсказания нейронная сеть выводит 95%
- 2) Для 30 дневного предсказания нейронная сеть выводит 91%

ЗАКЛЮЧЕНИЕ

С каждым днем в мире расширяется список возможностей для нас. Особенно в быстро растущем мире заметен рост IT технологий. С каждым днем объем информации в нашем мире увеличивается в геометрической прогрессии, такой большой объем требует сортировки и анализа для дальнейшего расширения IT сектора. Расширение приводит к социальным колебаниям и изменениям подхода в отношении человека и поставленного перед ним процесса, т.е. процесс становится проще и меняется алгоритм решения задачи, из-за чего

на рынок выходят новые компании и организации, а старые пытаются быстро нагнать технический прогресс.

В ходе процесса изучения методов анализа временных рядов, которыми мы можем представить поток информации практического любого объема, я столкнулся с интересной ситуацией. При ориентировании по графику динамики временного ряда, можно столкнуться с аномалией. В нашем случае аномалия имела социальный внешний характер, т.е. причиной такого резкого изменения динамики послужило социальное явление, а именно что внешний фактор повлиял на поведение и дальнейшую тенденцию.

ИНС обученная для предсказания дальнейшей динамики временного ряда, смогла обнаружить предпосылки резкой нисходящей тенденцией и построить предсказание практически точно к и исходным данным, хоть и предсказание не учитывает множества факторов, так как для любого временного ряда можно выстроить бесконечное число влияющих на него параметров.

Уже на данный момент существует довольно обширные модели анализа временных рядов и аномалий, содержащихся в них, практически все из них построены на нейронных сетях. Вычислительные мощности кластеров на базе которых работают самые сложные нейронные сети способны проводить анализ колоссальных предприятий и предсказывать взлеты и падения.

В конечном итоге можно сделать вывод что социальные явления, которые практически хаотически проявляются в нашем мире могут быть обнаружены первоначально из сфер, которые эти явления затрагивают.