Заключение

В этой статье представлена методика прогнозирования модели процесса — описания будущих процессов. Для этого мы разработали алгоритм, основанный на представлении данных о событиях в виде нескольких временных рядов. Каждый из этих временных рядов отражает эволюцию поведенческого аспекта модели процесса в виде прямых зависимостей (DF), что позволяет применять соответствующие методы прогнозирования временных рядов. DFs широко используются в интеллектуальном анализе процессов в качестве компонентов графов непосредственных зависимостей (DFG), которые являются полуформальными моделями процессов и часто привлекательны для специалистов-практиков. Последние широко используются в интеллектуальном анализе процессов в качестве представления процессов и, следовательно, являются своего рода моделью процесса. Реализация показывает, что прогнозируемые модели со средним размером алфавита (10–30 действий) имеют среднюю процентную ошибку менее 15% с точки зрения соответствия прогнозируемым процессам, что превосходит предложенные базовые модели.

Понимание бизнеса

Цель этой статьи – разработать методику прогнозирования всей модели процесса на основе исторических данных о событиях. Наша реализация демонстрирует осуществимость   
прогнозирования модели процесса с использованием реальных данных о событиях.

Аналитика процессов — это подраздел, который включает в себя предиктивный мониторинг процессов (PredictivePM), направленный на прогнозирование отдельных процессов или моделей процессов в целом.

Ключевой задачей в этом контексте является учёт эволюции, поскольку известно, что процессы подвержены изменениям. Прогноз может сообщить аналитику процессов, чем ожидаемые процессы будут отличаться от текущих, тем самым предоставив информацию для принятия решений по улучшению будущих процессов.

**Похожие работы и мотивация.**

Методы PPM разрабатываются с определённой целью: прогнозирования следующего действия, прогнозирования оставшегося времени или прогнозирования результатов. Переход от методов PPM с высокой детализацией к прогнозированию на основе моделей позволяет получить новые сведения о глобальном развитии процесса.

Так как горизонт прогнозирования значительно больше по сравнению с тем, что могут обеспечить существующие модели прогнозирования следующей операции, то прогноз на уровне модели даёт более полное представление о будущем развитии процесса.

**Понимание данных**

Описание выбранного датасета

**3 Подготовка данных**

**Прогнозирование на основе модели процесса.**

В этом разделе описывается, как извлекаются временные ряды прямых зависимостей из журналов событий, а также как они используются для получения прогнозов на основе модели процесса с помощью ряда широко распространённых методов прогнозирования.

Каждый журнал событий содержит записи путей, которые представляют собой последовательности событий, создаваемых информационной системой во время её работы. Отношения «непосредственно следует» между действиями в журнале событий можно выразить в виде функций подсчёта для пар действий >L:Σ×Σ→N, где >L(a1, a2) Эта функция подсчитывает количество раз, когда действие a1 непосредственно следует за действием a2 в журнале событий L.

Наконец, граф непосредственных зависимостей (DFG) процесса — это взвешенный ориентированный граф, в котором действия являются узлами, а отношения DF — взвешенными рёбрами, то есть DF G = (Σ, >L). Чтобы получить прогнозы относительно эволюции DFG, мы строим DFG для подмножеств журнала. Существует множество методов агрегирования и группирования для прогнозирования следующего шага, производительности и ориентированных на цель результатов.

В предложенном подходе к прогнозированию мы объединяем концепции анализа временных рядов. Для этого, эволюция DFG отслеживается на интервалах журнала, где возможна равновременная и равноразмерная агрегация. Равномерная временная агрегация: каждый интервал (пожурнал) журнала событий имеет равную продолжительность. Равноразмерная агрегация: каждый интервал (поджурнал) событий содержит в себе равное количество пар DF.

(Как будто бы важный абзац, но ничего не понятно) Эти агрегирования полезны по следующим причинам. Во-первых, агрегирование с равными интервалами, как правило, с большей вероятностью приближает исходные данные к временному ряду с белым шумом, что требуется для широкого спектра методов прогнозирования временных рядов [25]. Во-вторых, оба метода предлагают разные пороговые значения, при которых можно применять прогнозирование. В случае равномерной агрегации проще быстро создать нужное количество интервалов, просто разделив журнал событий на равномерные интервалы. Однако большинство методов прогнозирования временных рядов основаны на равномерных временных интервалах.

**4 Моделирование**

**От временных рядов DFG к прогнозам по процессной модели**

Целью прогнозирования по процессной модели является получение прогноза для будущих DFG путём объединения прогнозов по всем временным рядам DFG. Для этого мы предлагаем использовать методы прогнозирования временных рядов.

Например, наивный прогноз просто использует последнее значение временного ряда T в качестве прогноза. Альтернативный наивный прогноз использует среднее значение временного ряда T в качестве прогноза. Можно рассматривать DFG как многомерную совокупность временных рядов DFG или рассматривать каждый DFG по отдельности. Традиционные методы работы с временными рядами используют одномерные данные в отличие от многомерных подходов, таких как векторные авторегрессионные модели (VAR), а также методы машинного обучения, такие как нейронные сети или регрессоры случайного леса.

Однако для большого количества наборов данных, и авторы отмечают, что алгоритмы машинного обучения требуют значительно больше вычислительных мощностей. Этот результат был позднее подтверждён, хотя и отмечается, что гибридные решения эффективны [28]. Для более длительных периодов традиционные подходы к временным рядам по-прежнему превосходят модели на основе машинного обучения.

Поэтому в этой статье мы выбрали традиционные подходы к временным рядам и применили их к одномерным временным рядам Df, в которых присутствует хотя бы одно наблюдение за каждый сублог/временной интервал.

**Методы построения временных рядов**

Авторегрессия, скользящие средние, авторегрессионные интегрирующие скользящие средние (ARIMA) и модели переменной дисперсии составляют основные семейства традиционных методов прогнозирования временных рядов.

**Исследование изменений в процессах**

Описание инструмента визуализации

**Реализация**

Вся предварительная обработка была выполнена на Python с использованием комбинации pm4py7 и пакета statsmodels.

Чтобы получить представление об эффективности прогнозирования в сочетании с фактическим использованием DFG (которые редко используются в неагрегированном виде [12]), мы приводим среднюю абсолютную процентную ошибку (MAPE) между энтропийной релевантностью фактических и прогнозируемых DFG при полном размере, 50%-ном и 75%-ном сокращении, основанном на узлах (т. е. сохраняется только 25-й/75-й процентиль узлов по частоте). Таким образом, мы получаем показатель точности с точки зрения расхождения между фактическим и прогнозируемым поведением модели.

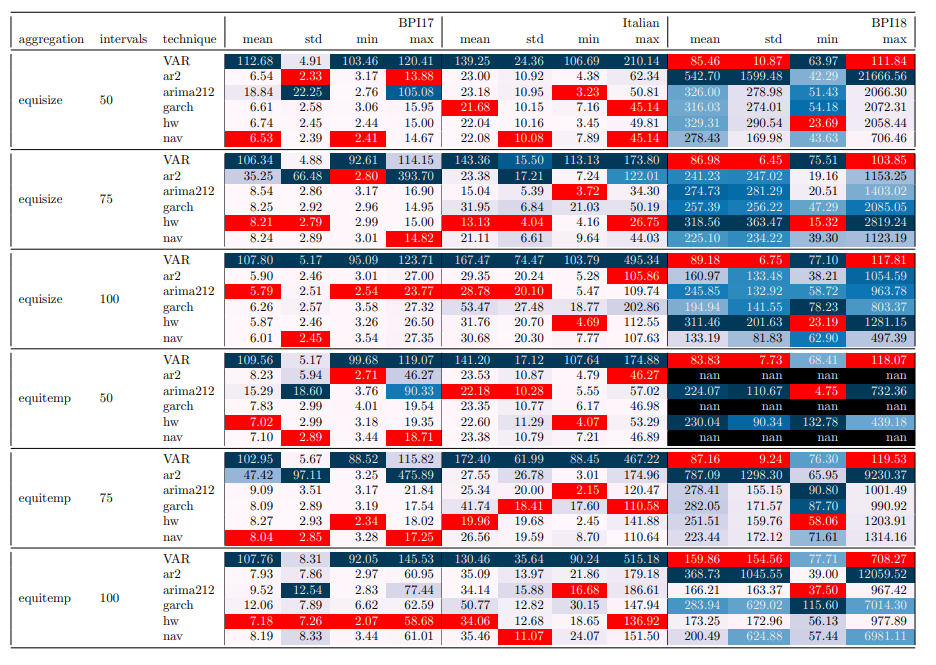


Таблица 5: Обзор средней процентной ошибки с точки зрения энтропийной релевантности для нередуцированных DFG для наборов данных BPI 17, итальянской службы поддержки и BPI 18. Цветовая кодировка аналогична таблице 4.

Модели, которые в среднем показывают наилучшие результаты, включают модель Холта-Уинтерса, модель авторегрессии и иногда наивный прогноз. Однако в последнем случае результаты других моделей схожи, чего нельзя сказать о первом случае.

Средние показатели ошибок после 50-процентного сокращения DFG приведены в таблице 6. Частота ошибок снижается до 5-20% для BPI12, BPI17 и сепсиса (за исключением 100 интервалов/эквивалентов) и до 11-20% для итальянского языка. Аналогичные показатели 100%+ для RTFMP с гораздо худшими результатами для агрегации эквивалентов зафиксированы, наряду с повышением до 145-282% для BPI18. Для BPI17 записывается много значений NaN, что связано с тем, что вычисление ER затруднено из-за неподходящих моделей, которые сокращаются до такой степени, что воспроизведение трассировок по ним невозможно. Однако в этом случае модели VAR работают лучше всего. Для других настроек хорошо работают модели AR/HW с хорошими результатами.

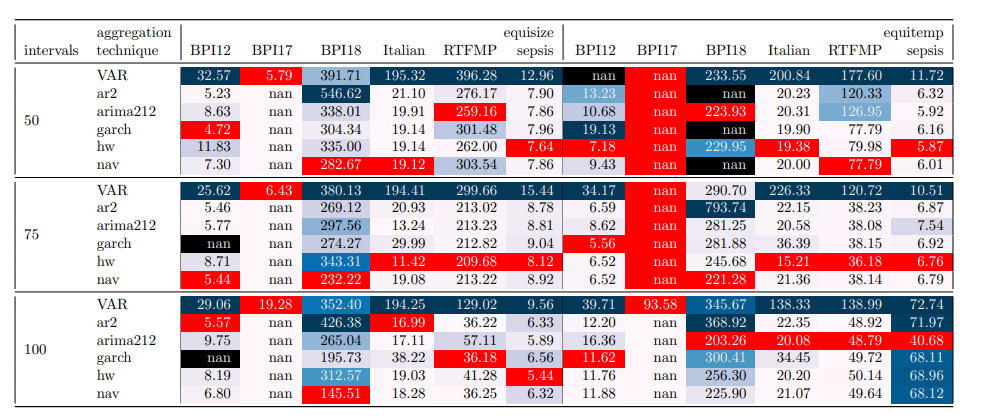


Таблица 6: Обзор средней процентной погрешности с точки зрения энтропийной значимости для DFG с 50-процентным сокращением.

При уменьшении на 75%, как показано в таблице 7, результаты для BPI12 и итальянских журналов службы поддержки ухудшаются, для BPI18, RTFMP и журналов событий, связанных с сепсисом, остаются прежними, а для BPI17 улучшаются/снова становятся вычислимыми.

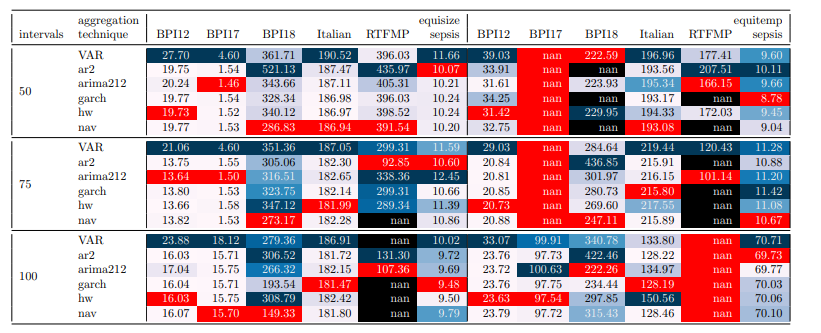


Таблица 7: Обзор средней процентной погрешности с точки зрения энтропийной значимости для DFG с сокращением на 75%.

Улучшение показателей в некоторых журналах событий после 50-процентного сокращения может быть связано с тем, что меньшее количество вариантов приводит к более сложным DFG, а прогнозы лучше предсказывают наиболее распространённые пути в графе активности, не наказывая за ошибки в расчётах ER из-за пропущенных действий и путей. Однако при очень сильном сокращении на 75% оказывается, что большинство методов с трудом создают воспроизводимую и совместимую с расчётами ER модель. Тем не менее, даже при таком сценарии для BPI12, BPI17 и сепсиса можно обнаружить 10-20% ошибок для обоих типов агрегации.

**5 Оценка**

В некоторых случаях для получения низкого уровня ошибок достаточно даже наивного прогноза. Однако в большинстве случаев модели HW, AR и ARIMA показывают наилучшие результаты по количеству ошибок.

Наконец, становится ясно, что нет очевидной пользы от использования какого-либо продвинутого статистического подхода к прогнозированию, учитывающего такие свойства, как сезонность, (долгосрочная) автокорреляция, сильные тенденции, корреляция между рядами и другие концепции, которые моделируются с помощью подходов. Это означает, что либо агрегированные данные не отражают должным образом эти концепции, либо они не присутствуют во (всех) временных рядах DF, либо подходы не способны извлекать их должным образом.

В этом смысле основной вклад данной статьи заключается в том, что она представляет собой план того, чем может быть прогнозирование на основе модели процесса, предлагая перейти от точки зрения, характерной для предиктивного мониторинга процессов, к анализу абстракций на уровне всей модели в виде одного экземпляра, который представляет собой использование временных рядов DF для прогнозирования DFG.