**УДК 004.942**

# МЕТОДЫ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА ДАННЫХ И ПРОЦЕССОВ В ЗАДАЧАХ КЛАСТЕРИЗАЦИИ И КЛАССИФИКАЦИИ

**Елховская Л.О.** (Университет ИТМО, г. Санкт-Петербург)
**Научный руководитель – к.т.н., доцент Ковальчук С.В.**
(Университет ИТМО, г. Санкт-Петербург)

Применение методов интеллектуального анализа данных и процессов дало развитие перспективным подходам в задачах моделирования. В работе представлен аналитический обзор таких методов как интегрированных решений, а также предложены новые подходы к их совместному применению, продемонстрированные на примере дистанционного мониторинга пациентов с артериальной гипертензией.

**Введение.** Осведомленность о том, что может произойти в краткосрочной и долгосрочной перспективе посредством предсказательного моделирования, стала ценной возможностью для большинства компаний. Однако существует потребность в лучшем понимании того, что происходит здесь и сейчас. Интеллектуальный анализ процессов (англ. Process Mining, сокр. PM) стал перспективным подходом к анализу внутриорганизационных рабочих процессов. Кроме того, техники PM могут использовать и быть адаптированы в алгоритмах кластеризации, деревьях решений, глубоком обучении, рекомендательных и экспертных системах и т.д. Например, модели процессов, извлеченные из журналов событий с помощью техник PM, могут быть использованы для прогнозного моделирования или симуляции процессов с использованием методов машинного обучения и статистики, и наоборот. Часто алгоритмы кластеризации применяются для нахождения основных шаблонов реализации процесса или абстракции событий в лог-данных, что позволяет строить более качественные модели процессов. Активное обучение в сочетании с PM используется для выявления и классификации отклонений в реализации процесса. Задача предсказания следующего действия в незаконченной цепочке событий породила отдельную ветвь в PM. Здесь применение методов машинного обучения и нейронных сетей может быть использовано для поддержки принятия решений и предсказания ресурсов для каждого действия. Потенциал применения PM не ограничивается приведенными примерами.

**Основная часть.** В данной работе предложены подходы к извлечению признаков из структуры моделей процессов, построенных с помощью техник PM, для прогнозного моделирования и кластеризации. Первый подход основан на концепции мета-состояний. Идея концепции возникла из области здравоохранения, где пациент участвует в реализации процесса. Циклическое поведение в реализации процесса может представлять собой рутинный комплекс процедур или повторяющиеся медицинские события, то есть нахождение пациента на некоторой стадии лечения или в мета-состоянии. События, составляющие цикл в модели, являются мета-состоянием, если частота циклического поведения в журнале событий превышает указанный порог. Идентифицированные таким образом мета-состояния используются в дальнейшем для обогащения признакового пространства для задачи прогнозирования. В рамках примера из здравоохранения новые мета-характеристики вводятся как относительная продолжительность пребывания пациента в конкретном мета-состоянии. Идея состоит в том, что такие скрытые состояния могут коррелировать с общим состоянием здоровья пациента. Структура модели может быть непосредственно использована для извлечения признаков в задаче кластеризации клинических путей. Независимые переменные могут быть описаны оценочными вероятностями событий и переходов между ними в модели процесса, а также числом элементов в ней. Такой подход может помочь определить кластеры схожих реализаций процессов вычислительно недорогим путем.

**Выводы.** Применение предложенных подходов продемонстрировано на двух взаимосвязанных наборах данных в рамках дистанционного мониторинга пациентов с артериальной гипертензией. Набор данных с клиническими и неклиническими событиями, инициированными измерениями давлений, используется для формирования журнала событий и для построения модели процесса реализации программы мониторинга с помощью модифицированного Fuzzy алгоритма. Данные о пациентах с измерениями давлений предназначены для прогнозного моделирования состояния пациента (задача бинарной классификации — контролируемые или неконтролируемые показатели давлений). С помощью подхода, основанного на концепции мета-состояний, удалось улучшить предсказательную способность модели логистической регрессии. Кластеризация с учетом структуры моделей процессов, построенных для каждого пациента, помогла выявить основные шаблоны реализации программы мониторинга.

Елховская Л.О. (автор) Подпись

Ковальчук С.В. (научный руководитель) Подпись