Интеллектуальная интеграция разнородных источников данных в задачах медицины и здравоохранения

М. А. Балахонцева[1], С. В. Ковальчук[2]
Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики
ФГБУ «НМИЦ им. В. А. Алмазова» Минздрава России
[1]mbalakhontceva@corp.ifmo.ru, [2]kovalchuk@corp.ifmo.ru

М. А. Ховричев[1], И. О. Кисляковский[2]
Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики
[1]mikhovr@gmail.com, [2]kisliakovskiii@yandex.ru

***Аннотация*. В работе освещено текущее состояние методологии интеграции данных и извлечения знаний из разнородных источников. Подробно описаны проблемы и вызовы, возникающие в рамках данной задачи и возможные пути их решения. Кроме того, авторами разработан тестовый пример интеграции нескольких источников данных, основанный на данных о работе кардиологического отделения специализированного медицинского центра.**

Ключевые слова: медицинские информационные системы; персонализированная медицина; интеграция данных и знаний; поддержка принятия решений

I. Введение

Информационные технологии на сегодняшний день получили широкое распространение в областях медицины и здравоохранения. Генерируемый объём данных неуклонно растёт, как и количество источников этих данных. Электронные медицинские карты (ЭМК), медицинские изображения, данные с носимых устройств и датчиков, опросники – это неполный перечень той информации, которая хранится в медицинских информационных системах (МИС) и сторонних сервисах по каждому из миллионов пациентов. В связи с этим, крайне актуальным является вопрос интеграции разнородных источников данных в рамках задачи персонификации процессов медицинской помощи. Другими словами, важно осуществить профилирование пациента на основе всех его взаимодействий как с медицинскими организациями, так и с любыми системами, данные из которых могут помочь при воссоздании полной клинической картины. Разрабатываемые методы и подходы должны отвечать современным вызовам предметной области.

II. Текущее состояние области и актуальные вызовы

В России активно реализуется проект Единой государственной системы в сфере здравоохранения (ЕГИСЗ) [1]. Система содержит в себе множество компонентов – ЭМК, электронную регистратуру, подсистему ведения административно-хозяйственной деятельности и многие другие компоненты для обмена медицинскими данными в масштабах страны. В настоящий момент к системе уже подключено 7 тыс. медицинских организаций и обрабатывается информация более чем 1,5 млн медицинских работников. С помощью Федеральной Электронной Регистратуры ежемесячно регистрируется более 7 млн. заявок на приём к врачу [2].

Прорыв в области наук о данных не мог не повлиять на рассматриваемую область. В современных интеллектуальных МИС данные представляют ценность не только для оперирующих ими людьми, но и для самих систем: информация, полученная из данных может использоваться МИС для улучшения собственной функциональности. Особенно это касается клинических систем поддержки принятия решений (КСППР): адаптируясь и приобретая опыт с каждым случаем, КСППР могут перейти из роли «ИИ-ассистента» и стать «ИИ-коллегой» [3].

Накапливаемый объём разнородной информации о пациенте требует поиска скрытых закономерностей методами Big Data [4]. Современные решения в этой области включают в себя использование стандартных технологий, таких как Apache Hadoop или Apache Hive. Для вычислений на видеокартах необходима довольно сложная архитектура, однако их стоимость и производительность подходят для биомедицинских вычислений. Похожим образом улучшают производительность процессоры Xeon Phi, позволяющие запускать рекомпилированный под них код почти в два раза быстрее, чем без их использования. При использовании облачных технологий как альтернативы стоит обратить внимание на то, что переносить в облако стоит код, а не данные.

Большой объём данных – это только одних из аспектов, которые следует учитывать при последующем анализе и интеграции. Без обработки и унификации разнородных данных невозможно сформировать входной поток для алгоритмов машинного обучения. При построении

Работа выполнена при финансовой поддержке Российского научного фонда, проект №17-71-10259

интеллектуальной модели для интеграции разнородных источников данных и знаний могут возникнуть следующие проблемы на этапе предобработки [4]: (а) зашумлённость данных: необходимо уделять отдельное внимание выбросам на этапе очистки набора данных, так как их удаление может полностью изменить качество модели; (б) наличие в данных пропущенных или отсутствующих значений: необходимо классифицировать пропуски по степени случайности возникновения; (в) вариативность терминологии (множество схем кодирования и вариативность международных стандартов): производить соответствие между классификациями отдельно от процесса обучения модели. Кроме того, одним из основных вызовов в процессе моделирования является разнородность источников данных: разные стандарты кодирования в МИС различных медицинских учреждений, необходимость получения дополнительных данных о социальных связях, паттернах потребительского поведения, демографической статистики. Из-за трудностей интеграции дополнительных источников данных с ЭМК осложняется и персонификация процессов оказания медицинской помощи [5]. По причине разнородности данных увеличивается количество выделяемых признаков. Высокая размерность в таком случае ведёт к разреженности: лишь небольшое подмножество признаков каждого пациента является актуальным и обновляемым, а в дальнейшем – и к несбалансированности данных.

Также открытой проблемой является работа с временными рядами, составляющими большинство медицинских данных. Попытка переформатировать ЭМК в табличный вид может привести к потере важной информации, поскольку каждый признак в данном случае может быть описан не одним значением, а последовательностью. Для решения данной проблемы используются методы, основанные на генерации символьных последовательностей и их дальнейшей кластеризации [6].

Работа с медицинскими данными с точки зрения семантики – один из основных вызовов в развитие инженерии знаний. Примерами медицинских онтологий могут служить Unified Medical Language System (UMLS) [7] и Medical Subject Headings (MeSH) [8] Тем не менее, онтологические базы знаний неудобны при использовании на практике в ЭМК, хотя и являются при этом основой для построения КСППР на знаниях [9]. Использования одних только онтологий для интеграции разнородных данных со сложными взаимосвязями недостаточно. Одним из расширений может служить методика построения баз данных по модели связанных данных (Linked Data Model, LDM) [10]. Анализ знаний статистическими методами также демонстрирует успешное применение в данной области.

Приватность данных, выражаемая в первую очередь, как врачебная тайна – самый сложный момент в проблеме персонификации медицинской помощи, препятствующий интеграции данных и знаний из разнородных источников. Использование блокчейна связывают с возможностью решить такие задачи как медленный доступ к медицинским данным; совместимость разнородных источников, контроль пациента над своими данными, приватность и безопасность. Первым рабочим прототипом функционирующей блокчейн-ЭМК стал MedRec [11]. MedRec – неавторизованная система, предполагающая включение «майнеров» (пользователей, предоставляющих вычислительные ресурсы для генерации новых блоков). Неавторизованность системы является серьёзным препятствием для использования в реальной медицинской практике. Тем не менее, первый пример блокчейн-ЭМК вдохновил исследователей на улучшение применимости технологии. Авторы [12], сравнивая свою сеть с Bitcoin предлагают более легковесный, а самое главное – масштабируемый блокчейн.

Преимущества применения блокчейн в медицине: (а) связывание всех взаимодействий пациента с разными ЛПУ в единой истории решит проблему персонификации; (б) синхронизация данных решит проблему дублирования записей и проведения исследований; (в) управление пользователем над своими данными.

III. Тестовый пример: подготовка данных для оценки нагрузки на младший медицинский персонал

В качестве тестового примера интеграции данных из разнородных источников были собраны данные ЭМК из МИС, выгрузка деперсонализированных данных из системы контроля управления доступом (СКУД) ФГБУ «НМИЦ им. В. А. Алмазова» (центр Алмазова), а также информация о штатном составе коллектива кардиологического отделения центра.

Так как данные МИС за определенный период выгружены в хранилище на PostgreSQL, то для работы с этим хранилищем выбраны соответствующие инструмент: psql 9.6 – инструмент командной строки для прямой работы с PostgreSQL; pgAdmin3 1.22.2 – визуальная среда для администрирования и разработки взаимодействия с postgres-совместимыми базами данных. Большинство процессов обработки, интеграции и анализа данных проводилось на базе Python 3.6.3 средствами библиотек: py-postgresql 1.2.1, pandas 0.21.1, numpy 1.13, datetime.

Для осуществления интеграции данных в рамках тестового примера была выбрана одна группа персонала – суточные медсестры, поскольку они наиболее мобильная и загруженная группа по числу событий в МИС и СКУД. Для визуализации данных была выбрана библиотека Seaborn. Средства данной библиотеки сориентированы на создание статистических графиков, в дополнение к базовым возможностям универсальной библиотеки matplotlib. Это позволяет визуально оценить полученные данные в связи с имеющейся возможностью составления выборок и отображения данных на конкретный временной период.

Помимо разнородности самих источников данных, существуют отличия и в формате представления одних и тех же данных в разных источниках. При выгрузке данных из МИС центра Алмазова формат временных меток выражается колонками event_date и event_time, который имеют тип varchar длин 300 и 500 соответственно. Для совместимости временные метки как МИС, так и СКУД были сконвертированы в объекты pandas.Timestamp.

В таблице событий МИС за относительно небольшой промежуток времени (за неделю) набирается значительное

количество событий, отнесённой к группе суточных медсестёр, но не имеющих отношения к их прямым действиям и обязанностям. К примеру, медсёстры заполняют ЭМК амбулаторных пациентов по назначениям врача в один момент времени (чаще всего в конце смены). Однако, сами эти события выполняются другими участниками рабочего процесса. Такого рода данные на начальном этапе включаются в рассмотрение в формате единого события: заполнение ЭМК пациента. Кроме того, такой формат ведения записей вносит дополнительную неопределенность в соответствии событий из таблицы МИС и их временных меток. Для преодоления этих сложностей необходимо иметь реестр действий, по метке которых однозначно определяется, было ли действие выполнено медсестрой или нет. Для поиска таких однозначных соответствий необходим детальный анализ цепочек событий из СКУД. Ряд обязанностей включен в деятельность медсестры (однако, нет чёткого нормативно определённого перечня), соответственно в данном тестовом примере интересны события по услугам, назначенным на исполнение медсестрам. Согласно расписанию коечных отделений центра Алмазова, в обязанности суточных медсестер входит приёмка новых и выписка прошедших лечение пациентов. В связи с этим, на данном этапе был осуществлён анализ загруженности кардиологического отделения по числу амбулаторных пациентов по дням недели (выгрузка данных из МИС) и анализ загрузки СКУД по числу событий, связанных с выбранной группой персонала. На рисунках представлены результаты анализа данных. На рисунке представлены результаты анализа данных за пять месяцев, сгруппированные по дням недели. Графики показывают относительную нагрузку на суточных медсестёр, согласно данным о количестве поступивших пациентов, выписывающихся пациентов и событий в СКУД. Такого рода оценки могут помочь сформировать штатное расписание более оптимально, а также при интеграции в них дополнительных данных о результатах опросников и описаний популяции пациентов, требующих ухода, полученные оценки могут быть уточнены и использованы при построении предсказательных моделей.

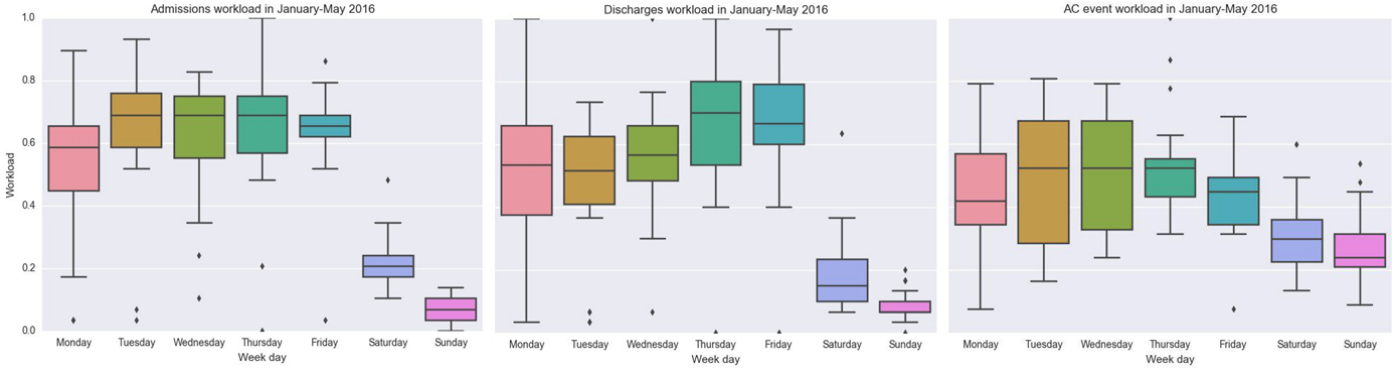

Рис. 1. Относительное распределение нагрузки по дням недели по количеству: поступающих пациентов, выписывающихся пациентов, событий в СКУД по определенной группе персонала

IV. Заключение

В рамках данной работы были подробно рассмотрены и проанализированы методы, технологии и подходы к интеллектуальной интеграции разнородных источников данных и знаний в медицине и здравоохранении. Кроме того, сформулированы актуальные вызовы и нерешённые задачи предметной области. Разрабатываемый тестовый пример основан на реальных источниках данных центра Алмазова и позволит интегрировать сразу несколько баз данных и дополнительных источников на примере задачи по оценке нагрузки на младший медицинский персонал.

Список литературы

[1] Портал оперативного взаимодействия участников ЕГИСЗ [Электронный ресурс]. [Б.м.: б.и.]. – Режим доступа: https://portal.egisz.rosminzdrav.ru/

[2] Ростех презентовал промежуточные итоги внедрения ЕГИСЗ [Электронный ресурс]. [Б.м.: б.и.]. – Режим доступа: http://rostec.ru/news/4521133/

[3] Baig M. M., Hosseini H. G., Lindén M. Machine learning-based clinical decision support system for early diagnosis from real-time physiological data //Region 10 Conference (TENCON), 2016 IEEE. IEEE, 2016. С. 2943-2946.

[4] Weber, G. M., Mandl, K. D., & Kohane, I. S. (2014). Finding the missing link for big biomedical data. Jama, 311(24), 2479-2480.

[5] Merelli, I., Pérez-Sánchez, H., Gesing, S., & D'Agostino, D. (2014). Managing, analysing, and integrating big data in medical bioinformatics: open problems and future perspectives. BioMed research international, 2014.

[6] Zhao, J., Papapetrou, P., Asker, L., & Boström, H. (2017). Learning from heterogeneous temporal data in electronic health records. Journal of biomedical informatics, 65, 105-119.

[7] Unified Medical Language System [Text]. [S. l.: s.n.]. – Access mode: https://www.nlm.nih.gov/research/umls/

[8] Medical Subject Heading [Text]. [S. l.: s.n.]. – Access mode: https://www.nlm.nih.gov/mesh/meshhome.html

[9] Mate, S., Köpcke, F., Toddenroth, D., Martin, M., Prokosch, H. U., Bürkle, T., & Ganslandt, T. (2015). Ontology-based data integration between clinical and research systems. PloS one, 10(1), e0116656.

[10] Xu, B., You, Y., Cheng, H., Gu, Y., & Cai, H. (2014, November). Personal healthcare record integration method based on linked data model. In e-Business Engineering (ICEBE), 2014 IEEE 11th International Conference on (pp. 38-43). IEEE.

[11] Azaria, A., Ekblaw, A., Vieira, T., & Lippman, A. (2016, August). Medrec: Using blockchain for medical data access and permission management. In Open and Big Data (OBD), International Conference on (pp. 25-30). IEEE.

[12] Xia, Q., Sifah, E. B., Smahi, A., Amofa, S., & Zhang, X. (2017). BBDS: Blockchain-based data sharing for electronic medical records in cloud environments. Information, 8(2), 4