Методы комплексного интеллектуального анализа клинических данных

А.О. Шелманов, Д.А. Девяткин

shelmanov@isa.ru,

devyatkin@isa.ru

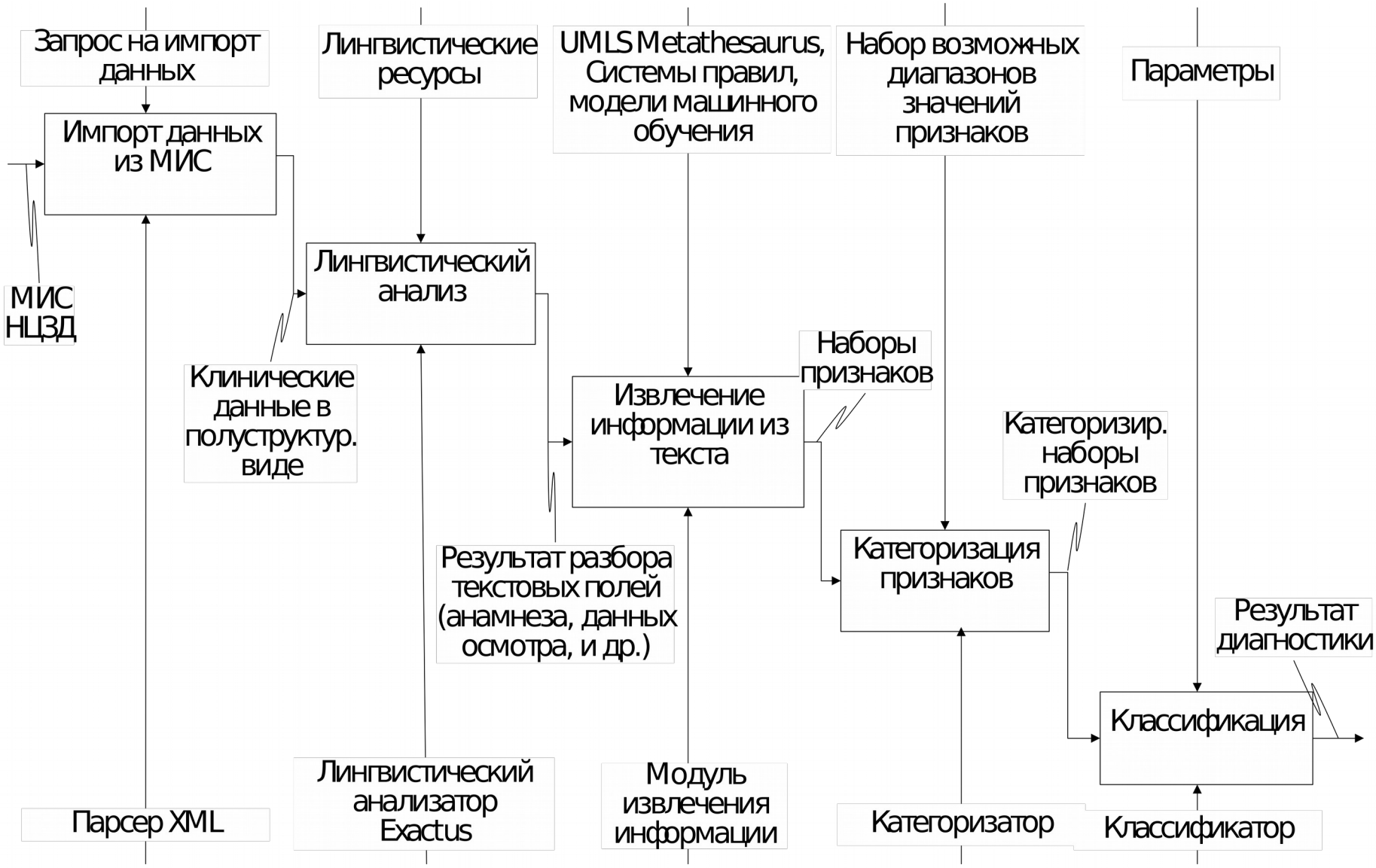
ИСА РАН ФИЦ ИУ РАН

2017,
Москва

Задача интеллектуального анализа клинических данных

- Клиники генерируют **большие объемы данных о пациентах**
- Автоматический или автоматизированный анализ этих данных может существенно **улучшить качество** предоставления **медицинских услуг**
- Большая часть этих данных представлена в виде слабоструктурированных **текстов на естественном языке**:
  - эпикризы, результаты осмотров, результаты рентгеновских обследований, ЭКГ, УЗИ, рекомендации врачей
- Необходимы **методы для совместной обработки** как структурированных данных (например, результатов анализов), так и слабоструктурированных данных, содержащихся в текстах
- Автоматический анализ текстов может помочь:
  - **Структурировать клинические записи**
  - Упростить поиск схожих случаев заболеваний
  - Упростить обработку историй болезней пациентов
  - Упростить обмен клиническими записями между медицинскими организациями
  - **Получить данные для методов принятия решений**
  - и др.

Схема анализа медицинских данных и текстов



Методы анализа слабоструктурирован-ных клинических текстов

Задачи, связанные с извлечением информации из клинических текстов

- Разработка анализатора для:
  - Нахождения в текстах медицинских терминов: **заболеваний**, **частей тела**, **лекарств** и проч.
  - Нормализация терминов: **сопоставить термины** в тексте **концептам** медицинского тезауруса
  - Выявить атрибуты терминов (в настоящий момент только для заболеваний и симптомов): **тяжесть заболевания**, **течение заболеваний**, **часть тела**, с которой связано заболеваний, **отрицание**, **относится ли заболевание к пациенту** или нет (атрибут «НеПациент»)
  - Нормализация атрибутов
- Для разработки и тестирования анализатора необходимо было **разметить корпус клинических текстов**

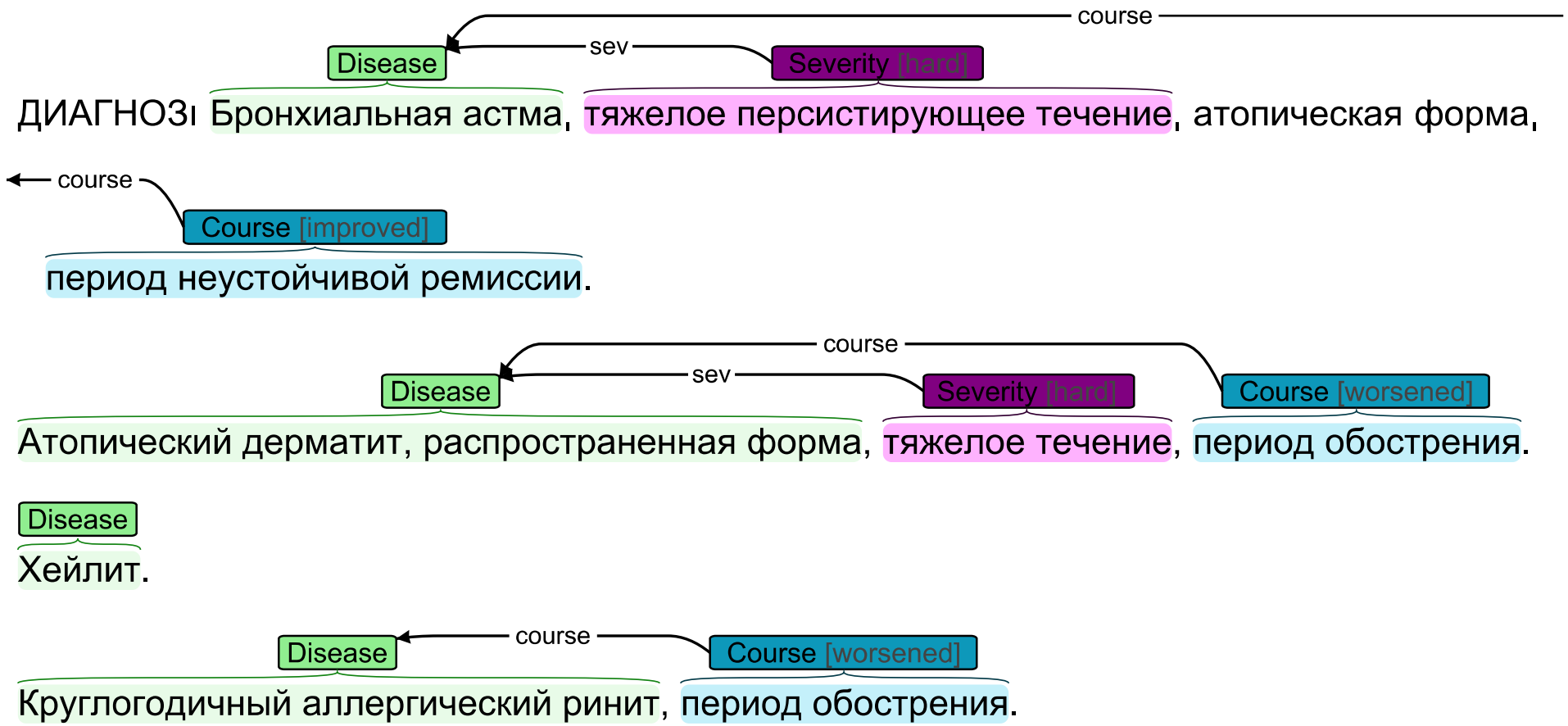
Работы, связанные с анализом клинических текстов

- Извлечение информации из медицинских текстов – **актуальная, быстро развивающаяся область**
- По этой теме проводится **большое количество соревнований** и семинаров:
  - CLEF eHEALTH (2013 – 2017)
  - SemEval (2014 – 2017)
  - i2b2 (2008 – 2014)
  - BioNLP-ST (2009 – 2015)
- Разработано множество систем обработки медицинских текстов на английском языке:
  - MedLEE (Friedman C., 2000)
  - HiTEX (Qing T Zeng et al., 2006)
  - **cTAKES** (Mayo Clinic) (Savova et al., 2010)
- Другие работы:
  - Извлечение из текстов упоминаний частей тела, с которым связано заболеваний и тяжести заболевания (Dligach D. et al., 2014)
  - Извлечение и нормализация упоминаний заболеваний (Pradhan S., 2014)

Корпус клинических текстов на русском языке (1)

- Корпус **размечен с привлечением специалистов** Научного центра здоровья детей (НЦЗД)
- Корпус состоит из 60 историй болезней пациентов НЦЗД с аллергическими заболеваниями и заболеваниями легких
- В историях болезни содержатся эпикризы, результаты рентгеновских обследований, ЭКГ, УЗИ, результаты осмотров, рекомендации врачей
- Размечены следующие аннотации: заболевания “disease”, симптомы “symptom”, лекарства “drug”, методы лечения “treatment”, результат применения методов лечения, части тела, атрибут тяжесть заболевания “severity”, атрибут течение заболевания “course”, отрицание “negation”, атрибут, определяющий относится ли заболевание к пациенту “НеПациент” + нормализованные значения атрибутов и связи

Пример разметки корпуса



Корпус клинических текстов на русском языке (2)

- Около 45 000 токенов. Более чем 7 600 размеченных сущностей и более 4,000 размеченных атрибутов и связей
- Схожие размеченные корпуса:
  - ShARe (Mowery D. L., 2014)
  - SHARPn (Pradhan S., 2015)
- Корпус анонимизирован: изменены имена и даты
- Доступен для исследовательских целей: http://nlp.isa.ru/datasets/clinical
- Для доступа к корпусу необходим сертификат, разрешающий работу с данными пациентов

Извлечение медицинских терминов из текстов и сопоставление их с тезаурусом (1)

- Был разработан анализатор на основе подхода, предложенного в MetaMap (Aronson et al., 2010), - система для поиска в текстах медицинских терминов и их нормализации
- Алгоритм анализа следующий:
  - По тексту **сгенерировать** большое количество различных **вариантов терминов**
  - Осуществить **нежесткое сравнение** сгенерированных вариантов с терминами из медицинского тезауруса, оценить их сходство
  - **Отранжировать** варианты по оценке сходства
  - Выбрать варианты, **наиболее похожие** на термины в тезаурусе

Тезаурусы для анализа клинических тексов на русском языке

- UMLS Метатезаурус + UMLS Семантическая сеть
  - Метатезаурус **сопоставляет концепты** различных других медицинских тезаурусов с **единым кодом** (уникальным идентификатором концепта, **УИК**)
  - Объединяет в себе MeSH, SNOMED-CT, ICD-10, и др.
  - Единственный ресурс на русском – **MeSHRUS** ~ 27 тыс. концептов; 85 тыс. терминов
  - Семантическая сеть **сопоставляет каждому УИК семантический тип**: болезнь, симптом, микроорганизм, хим. вещество, и др.
  - Этот ресурс использовался для извлечения из текстов упоминаний **болезней, симптомов и частей тела**
- Государственный реестр лекарственных средств (ГРЛС)
  - База данных **всех лекарственных препаратов,** официально зарегистрированных и разрешенных к продаже в РФ
  - Мы **сгруппировали** препараты со **схожими активными веществами** и на этой основе сформировали **концепты тезауруса**
  - Получено 3 600 уникальных концептов и около 12 000 терминов
  - Этот ресурс используется для извлечения упоминаний **лекарственных препаратов** из текстов клинических записей

Алгоритм извлечения медицинских терминов из текстов и сопоставления их с концептами тезауруса

- Построить **индекс ключевых слов терминов** из тезауруса
- **Сгенерировать варианты** терминов, используя за основу ключевые слова, а также: **синтаксические связи**, **линейный контекст** вокруг ключевого слова
- **Сравнить варианты** с терминами из тезауруса по след. параметрам:
  - **Лексическая похожесть** – взвешенное гармоническое среднее ():
    - взвешенного покрытия токенов термина варианта токенами термина в тезаурусе
    - взвешенного покрытия токенов термина в теазурусе токенами варианта
  - **Центральность**:
    - 1 если лемма синтаксической вершины наиболее крупной фразы варианта присутствует в термине тезауруса
    - 0 иначе
  - **Похожесть по связности** – аналогично “лексической похожести”, однако применяется к синтаксически связным фразам
- Выбрать пары вариант – термин в тезаурусе, оценка близости которых **выше заданного порога** (выбирается эмпирически / эвристически)

Определение отрицаний при упоминаниях заболеваний / симптомов

- Использован подход, основанный на **вручную построенных правилах**
- Подход, основанный на правилах, широко используется в современных системах, например, алгоритм **NegEx** (Chapman W. W., 2013) осуществляет поиск шаблонов в окне токенов вокруг упоминания заболевания
- В нашей работе реализован **поиск шаблонов в синтаксическом дереве**
- Шаблоны для поиска отрицаний:
  - “не” синтаксически зависит от токена заболевания/симптома
  - “не” синтаксически зависит от предикатного слова, с которым связаны токены заболевания/симптома
  - “нет” синтаксически управляет токеном заболевания/симптома
  - токен заболевания/симптома управляется глаголом со смыслом отрицания, например, “отсутствует”
  - “нет” следует за упоминанием заболевания/симптома

Извлечения атрибута “НеПациент”, определяющего относится ли заболевание к пациенту

- Использован подход, **основанный на вручную построенных правилах**
- Поиск **упоминаний родственников** в рамках предложения, содержащего упоминание заболевания (в педиатрических записях большое внимание уделяется наследственности)
- Правила для извлечения атрибута «НеПациент»:
  - фразы вида “у” + “упоминание родственника” синтаксически связаны или предшествуют упоминанию заболевания в предложении
  - слово “наследственность” предшествует упоминанию заболевания в предложении

Извлечение атрибутов тяжесть и течение заболевания

- **Машинное обучение** на разработанном корпусе, применены: лин. SVM, RBF SVM, случайный лес, AdaBoost
- Разработано два отдельных подмодуля:
  - Для определения **отрезка текста**, соответствующего атрибуту
  - Для **нормализация** атрибута
- Осуществлялась **классификация каждого токена** отдельно, определялось является ли он частью атрибута, связанного с заданным заболеванием
- Признаки: леммы и части речи слов в заданном окне вокруг классифицируемого токена; проверка того, что классиф. токен синтаксически зависит от токена заболевания; расстояние в токенах между классиф. токеном и токенами упоминания заболевания; позиция токена относительно упоминания заболевания; количество упоминаний заболеваний между классиф. токеном и тем упоминанием заболевания, с которым осуществляется попытка его связать

Нормализация атрибутов тяжести и течения заболевания

- **Машинное обучение** на разработанном корпусе, применено: линейный SVM, RBF SVM, случайный лес, AdaBoost
- Осуществлялась **классификация отрезков текста, помеченных на предыдущем этапе** как атрибут
- Признаки включают в себя **«мешок слов» (лемм)**, находящихся в отрезке текста, помеченном как соответствующий атрибут

Связывание упоминаний частей тела с упоминаниями заболеваний

- Упоминания частей тела и заболеваний определяются в тексте с помощью анализатора на основе тезаурусов
- Для их связывания используются **методы машинного обучения**: линейный SVM, RBF SVM, случайный лес, AdaBoost
- Признаки:
  - Расстояние в токенах между упоминанием заболевания и упоминанием части тела
  - Наличие синтаксической связи между упоминаниями
  - Проверка того, что они синтаксически подчинены одному и тому же слову
  - Часть речи этого слова
  - Количество всех упоминаний заболеваний между данным упоминанием заболевания и данным упоминанием части тела

Оценка качества извлечения упоминаний заболеваний

- Для оценки качества было предложено два «базовых» подхода:
  - В «Баз. подходе 1» помечаются все слова в тексте, которые присутствуют в терминах типа «заболевание» из тезауруса (максимальная полнота)
  - В «Баз. подходе 2» помечаются только те отрезки текста, которые полностью соответствуют терминам из тезауруса (максимальная точность)

Метод	**Полнота, %**	**Точность, %**	**F_1-мера,%**
Предлож. метод	72,8	95,1	**82,4**
Баз. подход 1	**84,9**	9,3	16,7
Баз. подход 2	69,8	**99,2**	81,9

Оценка качества извлечения упоминаний лекарственных препаратов

- Для оценки использовался тот же подход, что и для оценки качества извлечения заболеваний
- Результаты:
  - Точность = 84,3 %
  - Полнота = 74,6 %
  - F_1-мера = 79,2 %
- Вывод: применение ГРЛС для извлечения упоминаний лекарственных препаратов оправдано
- Ошибки:
  - При создании корпуса разметчики помимо лекарственных препаратов помечали также упоминания лечебной косметики
  - Пример: неофициальное название “пенициллин” отсутствует в ГРЛС, но “бензилпенициллин” присутсвует
  - Проблемы с нормализацией

Оценка качества извлечения атрибутов отрицаний и «НеПациент» при упоминаниях симптомов и заболеваний

- Извлечение этих атрибутов симптомов и заболеваний оценивалось совместно

Атрибут	Полнота, %	Точность, %	F_1-мера,%
Отрицание	98,7	95,3	97,0
«НеПациент»	90,9	96,8	93,8

- Простой подход на основе правил показывает хорошие результаты
- Однако количество размеченных атрибутов в корпусе мало для объективной оценки метода (около 100 примеров)

Оценка качества извлечения атрибутов тяжесть и течение заболевания

- Пятикратная перекрестная проверка на размеченном корпусе
- Протестированы различные методы классификации
- Атрибут «тяжесть заболевания»

Классификатор	**Полнота,%**	**Точность,%**	**F_1-мера,%**
Лин. SVM	99,2	41,7	58,6
RBF SVM	95,0	80,8	87,1
Случайный лес	93,6	82,6	**87,5**
AdaBoost	97,3	75,2	84,7

- Атрибут «течение заболевания»

Классификатор	**Полнота,%**	**Точность,%**	**F_1-мера,%**
Лин. SVM	**92,3**	99,2	**95,7**
RBF SVM	88,3	**99,3**	93,4
Случайный лес	88,3	**99,3**	93,4
AdaBoost	90,0	98,4	93,9

Оценка качества нормализации атрибутов и связывания упоминаний частей тела и заболеваний

Задача	Классификатор	Точность (Acc.),%
Нормализация атрибута «тяжесть заболевания»	Лин. SVM	88.4
	RBF SVM	88.0
	Случайный лес	89.3
	AdaBoost	**89.8**
Нормализация атрибута «течение заболевания»	Лин. SVM	89.4
	RBF SVM	91.4
	Случайный лес	**92.7**
	AdaBoost	91.4

Связывание упоминаний частей тела и заболеваний

Классификатор	Точность, %	Полнота, %	F_1-мера, %
Лин. SVM	85.4	**77.5**	81.0
RBF SVM	**91.4**	76.6	**83.3**
Случайный лес	86.6	75.8	80.8
AdaBoost	84.0	76.6	79.9

Методы анализа структурированных клинических данных

Задачи анализа

- Выявление влияния качественных признаков, извлеченных из текстов, на результаты классификации
- Выявление признаков заболеваний
- Построение шаблонных комбинаций симптомов и лекарственных средств, связанных с заболеваниями

Состав анализируемого набора данных (1)

Заболевание	Похожие заболевания
Бронхиальная астма	Бронхит Аллергический ринит Муковисцидоз
IgA-нефропатия	Широкий спектр гломерулярных заболеваний
Юношеский артрит	Спондилит

Состав анализируемого набора данных (2)

- В наборе данных представлено более 120 деперсонализированных историй болезни. Они содержат:
- анамнез;
- результаты анализов (анализы мочи, крови, посевы микрофлоры) в полуструктурированном виде;
- результаты проведения дополнительных исследований (томографии, рентгенографии, кожные пробы) в текстовой форме (заключения)

Набор ассоциативных правил, выражающих шаблонные комбинации признаков для болезней верхних дыхательных путей

Apriori algorithm [Agrawal R. et al., 1994]

- *('ГИПЕРСЕНСИБИЛИЗАЦИЯ', 'АСТМА БРОНХИАЛЬНАЯ', 'РИНИТ') => ('Монтелукаст')*
- *('ГИПЕРСЕНСИБИЛИЗАЦИЯ', 'АСТМА БРОНХИАЛЬНАЯ') => ('Будесонид')*
- *('ГИПЕРСЕНСИБИЛИЗАЦИЯ', 'РИНИТ') => ('АСТМА БРОНХИАЛЬНАЯ')*
- *('КАШЕЛЬ', 'АСТМА БРОНХИАЛЬНАЯ') => ('РИНИТ')*
- *('АСТМА БРОНХИАЛЬНАЯ') => ('Флутиказон’)*
- *('IGE повышен', 'РИНИТ') => ('КАШЕЛЬ')*
- *('IGE повышен') => ('Уровень базофилов повышен')*
- *('IGE повышен', 'ГИПЕРСЕНСИБИЛИЗАЦИЯ ') => ('РИНИТ')*

Влияние качественных признаков на результаты классификации (5-кратный перекрестный скользящий контроль)

Нозология	Категори-зация численных признаков	Метод	F_1	σ	F_1	σ
Бронхиальная астма	Нет	Деревья решений	0,82	0,23	0,69	0,22
		Случайный лес	0,95	0,07	0,82	0,22
		Градиентный бустинг на деревьях решений	0,78	0,23	0,70	0,20
	Да	Деревья решений	0,86	0,19	0,61	0,25
		Случайный лес	**0,98**	**0,04**	0,73	0,19
		Градиентный бустинг на деревьях решений	0,81	0,20	0,65	0,24
IgA-нефропатия	Нет	Деревья решений	0,87	0,21	0,63	0,24
		Случайный лес	0,74	0,24	0,75	0,25
		Градиентный бустинг на деревьях решений	0,88	0,19	0,69	0,24
	Да	Деревья решений	**0,92**	**0,15**	0,64	0,24
		Случайный лес	0,74	0,24	0,75	0,24
		Градиентный бустинг на деревьях решений	0,90	0,16	0,69	0,24
Юношеский артрит	Нет	Деревья решений	0,91	0,07	0,79	0,16
		Случайный лес	0,94	0,05	0,87	0,10
		Градиентный бустинг на деревьях решений	0,96	0,03	0,83	0,14
	Да	Деревья решений	0,90	0,06	0,76	0,14
		Случайный лес	0,95	0,05	0,89	0,10
		Градиентный бустинг на деревьях решений	**0,97**	**0,03**	0,84	0,12

Признаки заболеваний

Gini importance, Gini impurity [Breiman L., 1996]

Нозология	Тип признаков	Наиболее значимые признаки
Юношеский артрит	Качественные признаки	Локализация: ягодичная область. Симптом: лейкоцитоз.
	Численные признаки	Удельное количество лейкоцитов в крови, удельное количество тромбоцитов в крови, уровень креатинина в крови, уровень гемоглобина в крови, СОЭ
IgA-нефропатия	Качественные признаки	Симптом: васкулит, повышен уровень IgA, сердечный тон, гиперемия. Локализация: спина.
	Качественные признаки	Уровень креатинина в крови, удельное количество цилиндров в моче, удельный вес мочи, уровень холестерина в крови, наличие слизи в моче.
Бронхиальная астма	Качественные признаки	Симптом: сенсибилизация подтверждена пробами, бронхитические изменения в легких
	Численные признаки	Уровень эозинофилов в крови, уровень лимфоцитов в крови, уровень IgE в крови, СОЭ, уровень базофилов в крови

Заключение

- Разработан анализатор для извлечения комплексной информации из клинических текстов на естественном языке
- В предложенной схеме анализа решен ряд проблем: организация первичной обработки разнородных данных о пациентах, неравномерное распределение заболеваний в обучающей выборке и интеграция различных методов интеллектуального анализа данных и текстов в рамках единой процедуры
- Совместное использование методов анализа структурированных и текстовых данных в рамках единой процедуры, а также выполнение предварительной категоризации числовых показателей здоровья пациентов позволило значительно повысить качество классификации
- В рамках предложенной схемы можно анализировать разнородные данные как структурированного, так и неструктурированного характера, что в перспективе может позволить создать средства автоматизации различных мероприятий в ходе лечебного процесса при ряде хронических нозологических форм в педиатрической практике

Литература (1)

- Friedman C. A broad-coverage natural language processing system //Proceedings of the AMIA Symposium. – American Medical Informatics Association, 2000.
- Extracting principal diagnosis, co-morbidity and smoking status for asthma research: Evalua-tion of a natural language processing system / Qing T Zeng, Sergey Goryachev, Scott Weiss et al. // BMC medical informatics and decision making. — 2006. — Vol. 6, no. 30.
- Dligach, D., Bethard, S., Becker, L., Miller, T. A. and Savova, G. K. (2014), Discovering body site and severity modifiers in clinical texts, Journal of the American Medical Informatics Association (JAMIA), pp. 448–454
- Pradhan, S., Elhadad, N., South, B. R., Martinez, D., Christensen, L., Vogel, A., Suominen, H., Chapman, W. W. and Savova, G. (2015), Evaluating the state of the art in disorder recognition and normalization of the clinical narrative, Journal of the American Medical Informatics Association (JAMIA), (1), Vol. 22, pp. 143–154
- 2001Sokirko, A. (2001), A short description of Dialing Project, available at: http://www.aot.ru/

Литература (2)

- Apresjan J. D., Boguslavskij I. M., Iomdin B. L., Iomdin L. L., Sannikov A. V., Sannikov V. G. and Sizov L. L. (2005), Syntactically and semantically annotated corpus of Russian language: Present state and perspectives [Sintaksicheski i semanticheski annotirovannyj korpus russkogo jazyka: sovremennoe sostojanie i perspektivy], National Corpus of Russian Language: 2003–2005 [Natsional'nyj korpus russkogo jazyka: 2003–2005], pp. 193–214, (in Russian)
- Aronson, A. R. and Lang, F.-M. (2010), An overview of MetaMap: historical perspective and recent advances, Journal of the American Medical Informatics Association (JAMIA), (3), Vol. 17, pp. 229–236
- Chapman, W. W., Hilert, D., Velupillai, S., Kvist, M., Skeppstedt, M., Chapman, B. E., Conway, M., Tharp, M., Mowery, D. L. and Deleger, L. (2013), Extending the NegEx lexicon for multiple languages, Studies in health technology and informatics, Vol. 192, pp. 677–681
- Mowery D. L. et al. Task 2: Share/clef ehealth evaluation lab 2014 //Proceedings of CLEF 2014. – 2014.

Сравнение результатов анализа и разметки корпуса

Annotation		Corpus	Parser
ДИАГНОЗ: Бронхиальная астма, тяжелое персистирующее течение, атопическая форма, период неустойчивой ремиссии. Атопический дерматит, распространенная форма, тяжелое течение, период обострения. Хейлит. Круглогодичный аллергический ринит, период обострения. ... Пролапс митрального клапана с регургитацией 3-4 мм. Грудной сколиоз 1 степени. Вегетососудистая дисфункция по гипотоническому типу.			
Disease		Бронхиальная астма	Бронхиальная астма (C0004096)
	Severity	тяжелое персистирующее течение (hard)	тяжелое персистирующее течение (hard)
	Course	период неустойчивой ремиссии (improved)	период неустойчивой ремиссии (improved)
Disease		Атопический дерматит распространенная форма	дерматит (C0011603) | Атопический дерматит (C0011615)
	Severity	тяжелое течение (hard)	Течение (medium)
	Course	период обострения (worsened)	период обострения (worsened)
Disease		Круглогодичный аллергический ринит	ринит (C0035455) | аллергический ринит (C0018621, C0035457)
	Course	период обострения (worsened)	период обострения (worsened)
Disease		Пролапс митрального клапана с регургитацией 3-4 мм	Пролапс (C0033377) | Пролапс митрального клапана (C0026267, C0003505, C0040962, C0079485)
	Body location	митрального клапана	митрального клапана
Disease		Грудной сколиоз	Сколиоз (C0036439)
	Severity	1 степени (light)	1 степени (light)