Система анализа данных из научных источников на примере текстов по дендритноклеточным вакцинам*

А. А. Бойко, А. М. Кайдина, Я. С. Ким, А. Ю. Лупатов,
А. И. Панов, Р. Е. Суворов, А. В. Швец

Аннотация. Ускоряющееся увеличение количества публикаций приводит к практической невозможности проведения ручных полных обзоров результатов исследований в различных областях науки. Также в некоторых областях исследований достаточно сложно выявлять закономерности без привлечения специальных методов обработки данных. Лечение рака с помощью дендритноклеточных вакцин является одним из таких направлений исследований. В настоящей работе представлена система полуавтоматического извлечения и анализа информации из научных публикаций. Представлены экспериментальные результаты разработанных и примененных методов по разделению, классификации пациентов и извлечению причинно-следственных связей для задачи установления успешности лечения пациентов различными типами дендритноклеточных вакцин.

Ключевые слова: *анализ естественного языка, машинное обучение, анализ текста, причинно-следственные связи, ДСМ-метод, генетические алгоритмы, AQ-метод, дендритные клетки, противораковая вакцина, клеточная терапия.*

Введение

Исследования в области медицины в последнее время характеризуются большим количеством результатов клинических испытаний пациентов, которые представлены в том числе и в виде научных статей. Так в реестре клинических испытаний Clinicaltrials [1] и в системе поиска информации Medline [2] только по теме испытаний дендритноклеточных вакцин содержится 387 и 587 результатов соответственно. В связи с тем, что количество таких публикаций растет с каждым годом и уже существенно превысило возможности отдельного исследователя в их подробном ручном анализе даже в достаточно узком направлении исследований, большую роль приобретают системы по сбору и анализу научных статей с целью выделения наиболее ценной информации, включающей основные выводы и результаты. В настоящей работе представлен оригинальный метод построения системы анализа информации об объектах, выделяемых из научных статей на примере медицинских текстов с целевой информацией о клинических испытаниях пациентов. Используются как разработанные ранее методы выделения признаков и объектов из полных текстов научных статей, классические методы машинного обучения, так и новые методы полуавтоматического конструирования правил извлечения информации из текста и индуктивные методы анализа данных. Представлены результаты проведенного с помощью разработанной системы анализа источников в задаче оценки качества лечения пациентов различными типами дендритноклеточных вакцин.

Выбор данной предметной области в качестве тестовой для системы связан с высокой актуальностью иммунотерапии рака. Опухолевые клетки экспрессируют значительное количество белков, которые могут выступать в качестве антигенов. В тоже время, из-за низкой иммуногенности этих белков противоопухолевый иммунный ответ обычно не развивается. Эффективное развитие иммунного ответа связано с деятельностью антиген-представляющих клеток, которые после поглощения антигена представляют его на своей поверхности лимфоцитам. Наиболее активными антиген-представляющими клетками являются дендритные клетки (ДК). В случае опухолевого роста ДК практически лишены возможности поглощать опухолевые антигены, кроме того, из-за отсутствия необходимых молекулярных

* Исследование выполнено при финансовой поддержке РФФИ (проекты № 13–07–12127 офи_м и № 15–59–31516 РТ-оми).

сигналов не происходит созревание ДК, необходимое для индукции иммунного ответа. Эта проблема может быть решена за счет создания клеточных вакцин на основе ДК. Сейчас в мире активно разрабатываются и проходят клинические испытания клеточное вакцины против ряда онкологических заболеваний. В научной литературе описано большое количество различных типов клеточных вакцин, вариантов иммунизации и других особенностей иммунотерапии. При этом эффективность используемых подходов не всегда очевидна. Следовательно, возникает задача определения характерных признаков пациента, указывающих на целесообразность или нецелесообразность применения дорогостоящей вакцины, что и составило цель проводимого анализа источников с использованием разрабатываемой системы.

Далее в статье представлено описание основных шагов предлагаемого метода: сбор и структуризация исходных данных, анализ полученной информации с применением методов машинного обучения и индуктивных методов выявления новой информации. Приведены примеры результатов работы каждого этапа.

1. Сбор исходных данных

В качестве источников информации о результатах клинических исследований по применению ДК вакцин для лечения онкологических заболеваний была использована система поиска информации биомедицинского профиля Medline, являющаяся ресурсом Национального института здоровья США. Этот ресурс охватывает максимальное количество необходимых данных по всему миру. Основными критериями отбора статей было использование авторами английского языка и наличие результатов лечения. Фаза клинического исследования и его дизайн (рандомизация, использование слепого метода и т. д.) не учитывались при отборе.

Для извлечения данных при помощи ключевых слов были использованы сочетания слов «cancer» или «tumor» и «dendritic cells», а также ограничение «clinical trial». Из полученной выборки были исключены исследования, в которых использовались аллогенные или другие типы дендритных клеток, отличные от аутологичных. В результате было отобрано 468 статей. Среди них лечению карцином было посвящено 148 статей, меланом — 162, различных опухолей мозга — 38, сарком — 17. Кроме того, 103 статьи относились к лечению опухолей кроветворной и иммунной систем.

Анализ реестра клинических испытаний Clinicaltrials выявил 387 зарегистрированных исследований, однако только 20 из них содержали сведения о результатах. В связи с небольшим количеством пригодных для анализа данных в дальнейшем этот ресурс не использовался для обучения системы автоматизированного анализа.

2. Алгоритм численного построения функций Ляпунова

В результате была собрана и загружена в систему 71 научная статья. Основной целью структуризации данных являлось получение информации обо всех пациентах, участвовавших в исследованиях, в векторной форме, пригодной для анализа, построения автоматического классификатора пациентов и извлечения новой информации о причинно-следственных связях. Процесс структуризации данных состоял из следующих шагов:

- предварительный анализ данных и составление схемы разметки исходного материала;
- ручная разметка начального набора данных и дальнейшее полуавтоматическое извлечение информации;
- преобразование данных в векторную форму.

Рассмотрим отдельные этапы подробнее.

2.1. Предварительный анализ

Целью данного этапа является обзор состава и структуры основных элементов исследований, приведенных в отобранных материалах. На верхнем уровне к основным элементам исследований относится информация об исходных характеристиках пациентов (возраст, тип заболевания, стадия, результаты основных анализов), о проводимом лечении (способ подготовки вакцины, объем вакцинации), а также о результатах лечения (статус, результаты анализов после вакцинации).

На основании обзора статей был выбран следующий состав и структура извлекаемой информации, в основу которой было положено пять кластеров (фактически, это структура взаимосвязи аннотаций в соответствии с эталонной моделью TIPSTER [3]), представленных в табл. 1.

Данный набор атрибутов составляет тип объектов «Группа пациентов» и характеризует одного пациента или группу пациентов. В ходе формирования системы типов был исключен ряд параметров, редко встречающихся при описании клинических исследований в проанализированных статьях, например, размер опухоли, индекс Карновского и ECOG после лечения, а также гистологический тип опухоли. Кроме того, в ходе разработки системы типов из нее были исключены данные, представленные преимущественно в виде графиков и рисунков, которые не поддаются автоматическому анализу, в частности,

Таблица 1

Система типов

I) Данные о выборке: – Количество пациентов	II) Данные пациента: – Возраст – Пол – Раса – Гаплотип – Иммунный статус до иммунизации	III) Заболевание: – Диагноз – Стадия заболевания – Индекс ECOG до лечения – Индекс Карновского до лечения – Лечение до иммунизации
IV) Схема лечения: – Тип вакцины – Источник дендритных клеток – Индукторы созревания – Вакцинация – Количество вакцинаций – Общее количество введенных клеток – Адъювант в составе вакцины – Способ введения вакцины – Сопутствующая терапия	V) Результаты лечения: – Объективный клинический ответ – Выживаемость – Результат – Срок дожития/наблюдения – Единицы измерения – ELISPOT – Антиген – Количественные значения до и после иммунизации – Качественное значение – DTH – Антиген – Количественные значения до и после иммунизации – Качественное значение – Антиген-специфические лимфоциты in vitro – Антиген – Количественные значения до и после иммунизации – Качественное значение – Побочные эффекты – Опухолевые маркеры – Тип маркера – Количественные значения до и после иммунизации – Качественное значение	

результаты некоторых иммунологических тестов, характеризующих противоопухолевый иммунный ответ после проведенной иммунотерапии. Также были исключены параметры, значения которых одинаковы для подавляющего большинства исследований, такие, например, как индукторы дифференцировки моноцитов в дендритные клетки.

Важной особенностью исходных данных является неоднородность. В статьях используется два основных типа описаний: описание конкретных пациентов (при этом указывается, какие именно), описание некоторой группы пациентов (при этом не сообщается, о каких именно пациентах речь). В рамках данной работы предложено унифицировать схему разметки для обоих типов описаний. Это достигается посредством добавления для каждого простого атрибута верхнего уровня двух дополнительных атрибутов: количество пациентов и доля пациентов, т. к. в статьях встречаются оба варианта задания численности описываемой группы пациентов. К тому же, каждый из атрибутов верхнего уровня, кроме «Количества пациентов», имеет списочный тип (то есть может иметь несколько сложных значений). Более подробно процедура использования этих свойств описана в разделе 2.3.

2.2. Полуавтоматическая разметка начального набора данных

Программная система, в которой осуществлялась разметка статей, ранее поверхностно описывалась в [4]. К ключевым особенностям этой системы можно отнести:

- объектно-ориентированный подход к заданию схемы разметки. При этом процесс разметки заключается в создании объектов в соответствии с заранее описанной структурой типов и задании для каждого атрибута описывающих его фрагментов текста и извлечении нормализованного значения. В схеме разметки допускается агрегация и композиция объектов;
- индексирование всей имеющейся в системе информации в графовой базе данных. Это позволяет унифицировать алгоритмы для автоматизации процесса извлечения информации. Во время индексирования в базе данных создаются вершины для всех документов, типов, объектов, атрибутов, их нормализованных значений, графем, а также дуги между ними, имеющие смысл «быть частью», «иметь значение», «быть сопоставленным с» и т. п.

Информация в исходных материалах представляется как в виде обычного текста, так и в таблицах, причем информация об одних и тех же пациентах может приводиться в нескольких местах.

Цель разметки (извлечения) начального набора данных заключается в создании базовой выборки, подходящей для обучения методов классификации с целью автоматизированной разметки остальной части статей. Для этого вручную экспертами были размечены фрагменты обычного текста, содержащие релевантные описания, а также первые строчки таблиц.

Далее остальная часть статей была проанализирована в полуавтоматическом режиме. Автоматизированный анализ выполнялся в два этапа: поиск релевантных фрагментов текста и сопоставление их атрибутам объектов; нормализация значений атрибутов на основе сопоставленных фрагментов текста. При поиске релевантных фрагментов текста применялись особенности структуры, сохраненной в графовой базе данных. Использованный алгоритм является итеративным, каждая итерация которого состоит из следующих шагов:

1. Выбрать вершины (исходные вершины), соответствующие графемам, сопоставленным интересующим нас атрибутам объектов, для которых правильность сопоставления была подтверждена экспертом.

2. Начиная с выбранных вершин, сгенерировать заданное количество путей в графе, используя случайную политику обхода.

3. Преобразовать сгенерированные пути с целью их обобщения.

4. Для каждого сгенерированного пути построить его идентификатор, состоящий из меток дуг, входящих в него, расположенных в том же порядке, в котором дуги входят в путь.

5. Начиная с исходных вершин и переходя по дугам, имеющим метки в соответствии со сгенерированными идентификаторами путей, выбрать все вершины (контекстные вершины), находящиеся в конце каждого пути.

6. Начиная со всех контекстных вершин и переходя по дугам, имеющим метки в соответствии со сгенерированными идентификаторами путей в обратном порядке, выбрать все вершины, находящиеся в конце каждого инвертированного пути (целевые вершины), и при этом не являющиеся исходными.

7. Удалить из списка целевых вершин вершины, соответствующие графемам, находящимся близко по тексту к исходным, но которые при этом не были ранее размечены экспертом.

8. Найти непрерывные фрагменты текста, составленные из целевых вершин.

9. Для каждой целевой вершины вычислить оценку релевантности — количество раз сколько она участвовала в инвертированном пути в качестве конечной.

10. Для каждого найденного непрерывного фрагмента текста найти оценку релевантности как среднюю релевантность входящих в него вершин.

11. Вывести список найденных непрерывных фрагментов в порядке убывания рейтинга.

После каждой итерации эксперт проверяет найденные фрагменты и сопоставляет их атрибутам объектов.

В качестве необходимого в условиях данной работы дополнительного этапа предобработки выполнялось извлечение таблиц с помощью модифицированного алгоритма, изложенного в [5]. Каждая таблица представлялась в виде набора объектов, имеющих следующую структуру:

1. Таблица (Заголовок, Нижний колонтитул, Ячейки, Строки, Столбцы);

2. Ячейка (Текст, ССылкаНаСтроку, ССылкаНаСтолбец);

3. Строка (Порядковый номер, ЯчейкаЗаголовок, РодительскаяСтрока);

4. Столбец (Порядковый номер, ЯчейкаЗаголовок, РодительскийСтолбец).

Упомянутая система находится в стадии разработки, поэтому качество работы извлечения таблиц и поиска релевантных фрагментов текста формально не оценивалось и является предметом дельнейших исследований. Также в настоящее время не автоматизированы процессы поиска новых объектов и связывания объектов друг с другом, однако это планируется реализовать через поиск фрагментов текста, соответствующих атрибутам с множественностью «один к одному».

После извлечения и сопоставления всех релевантных фрагментов из загруженных в систему документов была применена процедура извлечения нормализованных значений (числовой или номинальной величины, представляющей информацию, содержащуюся в соответствующих фрагментах текста). Для этого применялась комбинация из метода ближайших соседей, основывающегося на N-граммах символов, и вручную составленных правил, основанных на регулярных выражениях. Средняя F1-мера работы метода ближайших соседей в зависимости от атрибута составляла от 0.6 до 1 (по результатам трехкратной перекрестной проверки на вручную размеченной части данных). Для обработки случаев, в которых метод ближайших соседей ошибался, были использованы вручную составленные правила, основанные на регулярных выражениях. Проводились эксперименты и с другими методами машинного обучения (SVM, Random Forest), но на имеющихся данных они показали себя хуже.

В результате этого этапа обработки исходных материалов было извлечено 927 объектов типа «Группа пациентов» соответствующих в общей сложности 1549 пациентам.

2.3. Преобразование в векторную форму

Цель данного этапа — преобразование гетерогенной информации, извлеченной из исходных материалов, в однородную выборку в некотором векторном пространстве, пригодную для применения современных методов анализа данных с целью поиска новых закономерностей.

Ключевой особенностью извлеченной информации является гетерогенность, проявляющаяся в следующих ситуациях:

1. В рамках одной статьи могут быть как объекты, характеризующие отдельных пациентов (явно указывается каких), так и объекты, характеризующие сразу нескольких пациентов (без указания конкретных пациентов). При этом в объектах этих двух типов могут быть заполнены как одни и те же атрибуты, так и различные.

2. В рамках одной статьи могут встречаться объекты, соответствующие как непересекающимся группам пациентов, так и частично пересекающимся.

3. Допускается вложение групп пациентов.

Эти свойства делают извлеченные данные в необработанном виде непригодными для дальнейшего использования.

Для решения этой проблемы в рамках данной работы предлагается рассматривать исходные материалы и извлеченную из них информацию как описание сложного совместного распределения вероятностей нескольких случайных величин (как дискретных, так и непрерывных). Такой подход позволяет применить процедуру генерации заданного количества реализаций этих случайных величин по имеющемуся описанию, общий алгоритм которой состоит из следующих основных шагов:

1. Выбрать очередную статью из загруженных.

2. Определить количество пациентов, участвовавших в описанном в статье исследовании, как наибольшее количество пациентов в какой-либо «Группе пациентов» из этой статьи.

3. В соответствии с определенным количеством пациентов сгенерировать набор векторов-заготовок для хранения реализаций случайных величин.

4. Определить группы пациентов, характеризующих отдельных пациентов («Количество пациентов» у которых равно 1), создать по ним генераторы и в равном количестве применить каждый генератор к некоторому подмножеству имеющихся векторов-заготовок.

5. Отсортировать остальные группы пациентов в порядке увеличения количества пациентов, создать по ним генераторы и последовательно применить каждый генератор.

Генератор — программный объект, для которого задана частота срабатывания (как отношение количества пациентов к количеству пациентов в родительском объекте) и который ответственен за заполнение элементов векторов-заготовок. Генераторы могут быть вложены друг в друга. Процедура построения генератора заключается в рекурсивном обходе дерева свойств, соответствующего каждому объекту типа «Группа пациентов», при этом для каждой вершины создается генератор. Для генерации простых свойств создаются атомарные генераторы, заполняющих только один элемент вектора-заготовки и не содержащие вложенных генераторов. Процедура применения генератора состоит из трех основных шагов:

1. Оценить возможность применения генератора: если какие-либо из вложенных генераторов могут присвоить новые, отличные от имеющихся, значения уже заполненным элементам вектора-заготовки, то генератор не применяется.

2. Если конфликты не обнаружены, недетерминировано определяется применять генератор или нет. Это осуществляется в соответствии с заданной частотой срабатывания.

3. Если генератор атомарный, заполнить соответствующий элемент вектора-заготовки.

4. Если генератор не атомарный, применить все вложенные генераторы.

В результате применения описанной процедуры к извлеченным данным была сгенерирована матрица 1549×46, содержащая 21 столбцов с числовыми значениями и 25 столбцов с номинальными значениями.

3. Анализ данных

В рамках работы анализ данных преследует следующие цели:

1. Построить правила для определения пациентов, которые поддаются лечению с помощью тех или иных типов вакцин.

2. Построить гипотезы о влиянии тех или иных признаков пациентов и подходов к лечению на исход применения вакцины.

3. Оценить качество выборки (целостность, непротиворечивость и т. п.).

Для достижения указанных целей рассмотрено несколько задач классификации и регрессии:

1. Предсказание объективного клинического ответа в зависимости от остальных признаков пациентов (классификация).

2. Предсказание характера объективного клинического ответа (классификация).

3. Предсказание ответа на вопрос «Превысит ли время дожития пациента некоторую заданную величину?» для различных порогов времени дожития (классификация).

4. Предсказание времени дожития (регрессия).

Таблица 2

Результаты теста разделимости пациентов с заданными объективными клиническими ответами

Объективный клинический ответ	Число положительных примеров	F1-мера		Точность		Полнота	
		Разделимость	3-КВ	Разделимость	3-КВ	Разделимость	3-КВ
Стабильная болезнь	314	0.63	0.60	0.96	0.88	0.47	0.46
Частичное выздоровление	201	0.71	0.69	0.77	0.74	0.66	0.65
Полное выздоровление	108	0.89	0.88	0.98	0.99	0.81	0.79
Прогрессирующая болезнь	480	0.97	0.82	0.96	0.79	0.98	0.85
Среднее		0.8	0.75	0.92	0.85	0.73	0.69
Положительный ответ (полное или частичное выздоровление)	309	0.8	0.71	0.85	0.88	0.75	0.61

При анализе данных использовались общеизвестные методы машинного обучения (в частности, метод ансамбля деревьев решений). Для построения гипотез о причинах значений тех или иных свойств пациентов применились авторские методы GAAQ+JSM и AQ+JSM, использующие ДСМ метод на предварительно оптимизированной базе фактов.

3.1. Классификация пациентов

В данном эксперименте участвовали только те пациенты, у которых было заполнено поле «Объективный клинический ответ» (1103 пациента). В качестве мер качества были выбраны традиционные для задач классификации точность, полнота и F1-мера.

Матрица входных данных обладает неравномерной заполненностью, наиболее часто встречаемые признаки: «Возраст» (80 %), «Объективный клинический ответ» (79 %), «Пол» (69 %), «Предыдущий диагноз» (64 %), «Количество вакцинаций» (62 %). Наиболее редко встречаемые признаки: «Раса» (< 1 %), «Количественные показатели опухолевого маркера до и после иммунизации» (около 1 %).

В связи с разреженностью и с требованиями современных методов машинного обучения входные данные подвергались следующей предобработке:

1. Отбрасывались столбцы, содержащие меньше, чем T% заполненных значений. Проводилось несколько экспериментов для разных T.

2. Столбцы с номинальными значениями преобразовывались в несколько числовых с помощью процедуры бинаризации (замены признака, который может принимать значения из некоторого множества размера k, k признаками, каждый из которых принимает значение из множества {0, 1}).

3. Все столбцы приводились к диапазону [0, 1] посредством вычитания минимального значения и деления на максимальное после вычитания. Столбцы, при нормализации которых приходилось делить на 0, отбрасывались (такие столбцы содержат только одинаковые значения, поэтому не информативны).

В качестве базового метода классификации использовался ансамбль решающих деревьев, построенных по принципу бэггинга. Использована реализация RandomForestClassifier [6] из пакета Scikit Learn [7].

Для оценки разделимости классов на имеющихся данных были проведены эксперименты с обучением и тестированием на одном и том же наборе данных. Для оценки предсказательной силы была проведена аналогичная серия экспериментов по методике трехкратной перекрестной проверки. Результаты приведены в табл. 2.

В данном эксперименте участвовали только пациенты, у которых было заполнено поле «Время дожития/наблюдения» (574 пациента). На рис. 1 приведено распределение количества пациентов в зависимости от их времени дожития.

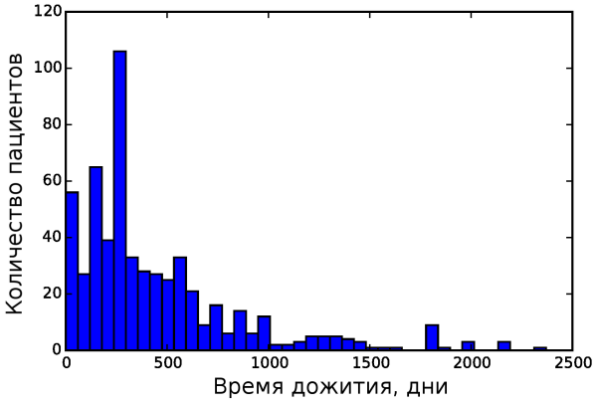

Рис. 1. Распределение количества пациентов в зависимости от времени дожития

Для оценки возможности предсказания ожидаемого времени дожития также были проведены две серии экспериментов: с проверкой на обучающей выборке (для оценки принципиальной возможности такого предсказания по имеющимся данным) и с трехкратной перекрестной проверкой. В качестве оценок качества регрессии использовались объясненная дисперсия, средняя абсолютная ошибка и коэффициент детерминации. Результаты экспериментов приведены в табл. 3.

Таблица 3

Результаты экспериментов по оценке возможности предсказания времени дожития

	Объясненная дисперсия	**Средняя абсолютная ошибка (в днях)**	**Коэффициент детерминации**
Проверка на обучающей выборке	0.8	96.7	0.8
Трехкратная перекрестная проверка	– 0.12	348.6	– 0.13

Эксперименты показали сложность предсказания точного времени дожития. Сравнение количества пациентов с различным временем дожития и установленным объективным клиническим ответов (рис. 2) показало, что эти признаки часто вступают в противоречие. В свою очередь, это говорит о зависимости времени дожития от множества других факторов, не учтенных в настоящем исследовании.

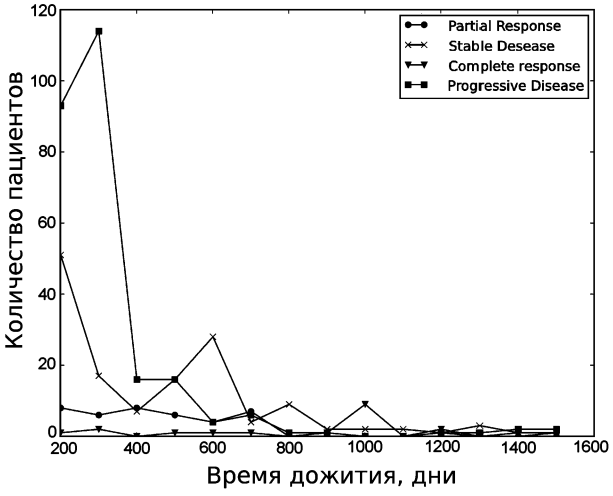

Рис. 2. Распределение количества пациентов с определенным временем дожития в зависимости от клинически установленного объективного ответа

3.3. Предсказание превышения времени дожития заданного порога

На следующем этапе анализа данных решалась задача предсказания превышения ожидаемого времени дожития заданного порога. В данном эксперименте участвовали только пациенты, у которых было заполнено поле «Время дожития/наблюдения» (574 пациента). На рис. 3, 4 и 5 соответственно приведены показатели полноты, точности и F1-меры при тестировании классификатора на обучающей выборке в зависимости от выбранного порога времени дожития (тест разделимости), а также те же показатели, полученные по методике трехкратной перекрестной проверки (тест предсказательного потенциала). На рис. 6 приведено количество положительных примеров для каждого использованного порога времени дожития.

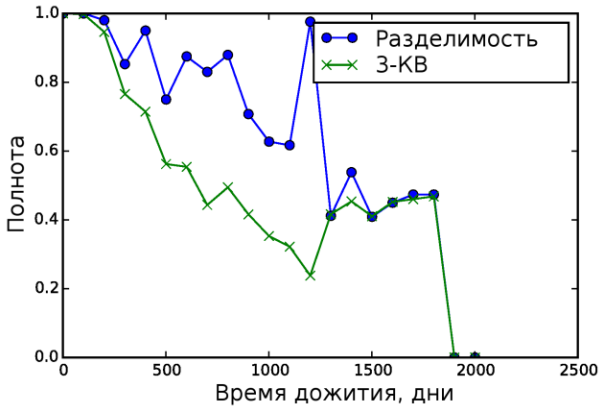

Рис. 3. Измеренная полнота классификации в зависимости от выбранного порога времени дожития (для теста разделимости и трёхкратной перекрёстной проверки)

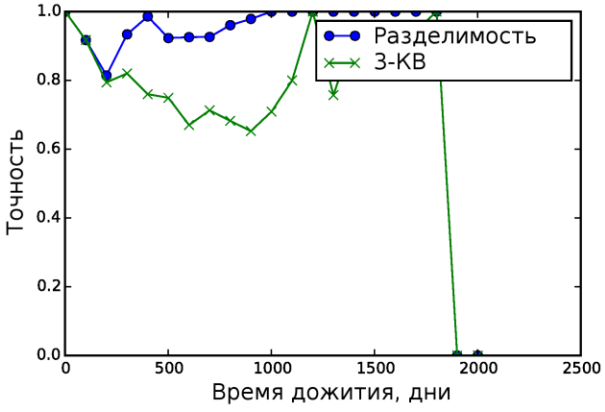

Рис. 4. Измеренная точность классификации в зависимости от выбранного порога времени дожития (для теста разделимости и трёхкратнойперекрёстной проверки)

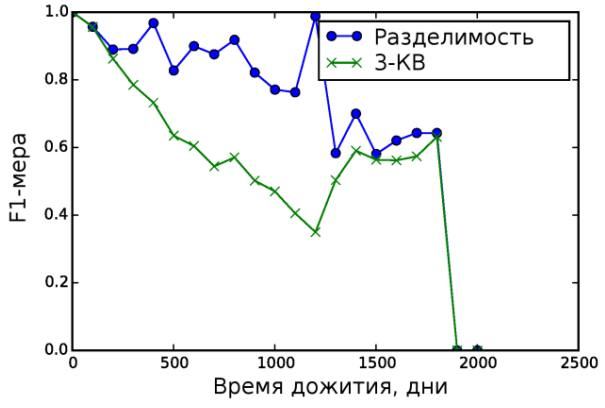

Рис. 5. Измеренная F1-мера классификации в зависимости от выбранного порога времени дожития (для теста разделимости и трёхкратной перекрёстной проверки)

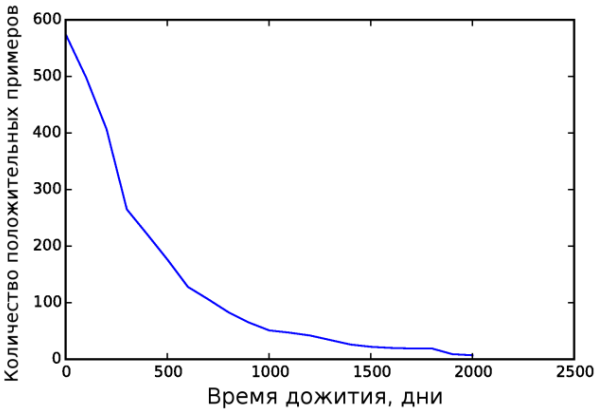

Рис. 6. Количество положительных примеров для каждой задачи классификации в зависимости от выбранного порога времени дожития

3.4. Извлечение каузальных связей

В связи с тем, что при работе с данными о пациентах встает задача не только предсказать некоторые характеристики, но и объяснить их, следующим этапом анализа данных было построение гипотез о причинно-следственных связях. Для этого применялся оригинальные методы GAAQ+JSM и AQ+JSM, включающиеся в себя предварительный этап сокращения пространства признаков.

В качестве метода для отбора признаков применялось сочетание логического метода индуктивного обучения AQ (quasi-minimal algorithm) с коэволюционным асимптотическим генетическим алгоритмом GAAQ [8]. Метод AQ заключается в формировании правил, описывающих группы объектов заданного класса. Каждое правило представляет собой конъюнкцию конкретных значений отличительных признаков этих объектов. Коэволюционный асимптотический генетический алгоритм позволяет решать задачу оптимизации, состоящей в максимизации числа покрываемых AQ-правилом объектов при минимизации количества признаков, включаемых в правило. Алгоритм GAAQ генерирует набор правил, покрывающих объекты заданного класса и не покрывающих ни одного объекта других классов.

Стоить отметить, что и стандартный алгоритм AQ, и его модификация GAAQ позволяют работать с неравномерно заполненными данными. Каждое пропущенное значение признака рассматривается как любое возможное значение этого признака. Считается, что правило покрывает объект, если для каждого значения, включенного в правило, соответствующее значение объекта совпадает с ним или является пропущенным, но, по крайней мере, одно из этих значений объекта заполнено.

Рассмотрим, какой из этих алгоритмов следует использовать для отбора признаков. В табл. 4 приведены результаты применения обоих алгоритмов ко всему набору данных.

Признаки, входящие в правило с максимальным покрытием для AQ:

1. В случае положительного результата лечения: «Выживаемость/Результат», «Индекс ECOG», «Антиген-специфические лимфоциты in vitro», «Вакцинация/Количество вакцинаций».

2. В случае негативного результата лечения: «Источник дендритных клеток», «Стадия заболевания», «Вакцинация/Общее количество введенных клеток», «Возраст», «Выживаемость/Результат», «Антиген-специфические лимфоциты in vitro», «Вакцинация/Количество вакцинаций».

Признаки, входящие в правило с максимальным покрытием для GAAQ:

1. В случае положительного результата лечения: «DTH», «ELISPOT/Антиген», «Адъювант в составе вакцины», «Антиген-специфические лимфоциты in vitro», «Выживаемость/Срок дожития/наблюдения», «Гаплотип», «Диагноз», «Иммунный статус до иммунизации», «Индукторы созревания ДК», «Лечение до иммунизации», «Опухолевые маркеры/Тип мар-

Таблица 4

Результаты работы алгоритмов AQ и GAAQ для отбора признаков

	Тип алгоритма отбора признаков	**Положительный результат лечения (Частичное или полное выздоровление)**	**Негативный результат лечения (Стабильная или прогрессирующая болезнь)**
Число правил	AQ	26	17
	GAAQ	75	11
Процент покрытых объектов, %	AQ	82	100
	GAAQ	98	39
Процент объектов, покрытых правилом с максимальным покрытием	AQ	23	44
	GAAQ	43	27
Среднее количество объектов, покрываемых одним правилом, %	AQ	9	9
	GAAQ	21	8
Средняя длина правила (количество признаков)	AQ	11	6
	GAAQ	14	16

кера», «Сопутствующая терапия», «Способ введения вакцины», «Тип вакцины».

2. В случае негативного результата лечения: «Адъювант в составе вакцины», «Вакцинация/Общее количество введенных клеток», «Возраст», «Выживаемость/Результат», «Выживаемость/Срок дожития/наблюдения», «Гаплотип», «Диагноз», «Иммунный статус до иммунизации», «Опухолевые маркеры», «Опухолевые маркеры/Тип маркера», «Пол», «Способ введения вакцины», «Стадия заболевания», «Тип вакцины».

Согласно табл. 4 для первого класса лучшие результаты показывает алгоритм GAAQ: почти все объекты покрыты сформированными правилами (98 %), правила в среднем покрывают большее число объектов (21 % против 9 %), лучшее найденное правило также покрывает большее число объектов (43 % против 23 %). Объекты второго класса лучше описываются правилами, найденными с помощью алгоритма AQ, однако среднее количество покрываемых объектов отличается незначительно.

Особенность правил, получаемых с помощью алгоритма AQ, заключается в том, что они опираются на признаки, включенные в правило с максимальным покрытием, которое меняется с каждым запуском алгоритма. Тогда как правила, формируемые с помощью GAAQ, не зависят друг от друга и позволяют обнаружить все признаки, характеризующие объекты заданного класса. В связи с этим использование этих правил и соответствующих им признаков более предпочтительно для поиска гипотез о причинно-следственных связях.

Полученные с помощью AQ и GAAQ подмножества признаков, использовались для составления базы фактов ДСМ метода порождения гипотез [9] о наличии причинно-следственных связей между значениями признаков. Корректность совместного использования ДСМ метода на основе описаний классов, составленных с помощью AQ покрытий опирается на гипотезу о том, что алгоритм построения описания классов путем путем поиска максимально обобщенных общих свойств использует тот же принцип, что и поиск максимальных пересечений при генерации гипотезы о причинно-следственной связи и поэтому не вносит искажений в каузальное пространство базы фактов. В сформированной базе фактов (подробнее алгоритм представлен в [8, 10–11]) проводилось устранение конфликтов и дубликатов: удалялись положительные и отрицательные примеры, дублирующие другие положительные или отрицательные примеры, если описание положительного примера совпадало с отрицательным — удалялся последний. В качестве алгоритма поиска минимальных пересечений использовался алгоритм Норриса [12]. Полученное множество гипотез редуцировалось путем удаления вложенных гипотез и гипотез длины, превышающей некоторое критическое значение. Количество найденных каузальных связей в зависимости от метода формирования базы фактов и некоторые причины для разных значений признаков «Выживаемость» и «Объективный клинический ответ» представлены в табл. 5.

Некоторые причины классового свойства, полученные методом AQ+JSM:

1. В случае, когда целевым свойством выступало свойство «Выживаемость — летальный исход»: «Кол-во клеток введенных за одну вакцинацию — высокое», «Индекс Карновского — низкий», «Возраст — выше среднего», «Возраст выше среднего и пол женский».

2. В случае, когда целевым свойством выступало свойство «Объективный клинический ответ — стабильная или прогрессирующая болезнь»: «Индекс Карновского — высокий», «/Кол-во клеток введенных за одну вакцинацию — высокое», «DTH — не проводилось и Возраст — низкий», «Возраст — низкий и Лечение до иммунизации — химио-, гормональная, иммуно- и радиотерапии».

3. В случае, когда целевым свойством выступало свойство «Объективный клинический ответ — частичное или полное выздоровление»: «Кол-во клеток введенных за одну вакцинацию — высокое».

Некоторые причины классового свойства, полученные методом GAAQ+JSM:

1. В случае, когда целевым свойством выступало свойство «Выживаемость — летальный исход»: «Количество вакцинаций — высокое», «Возраст — низкий или высокий», «Индекс ECOG — низкий или средний», «DTH — не проводилось и Антиген-специфические лимфоциты in vitro — раковый антиген».

2. В случае, когда целевым свойством выступало свойство «Объективный клинический ответ — ста-

Таблица 5

Количество найденных каузальных связей

		Классовое свойство «Выживаемость»		**Классовое свойство «Объективный клинический ответ»**	
		«Летальный исход»	**«Положительный исход»**	**«Неуспешное лечение»**	**«Успешное лечение»**
Количество найденных каузальных связей	AQ+JSM	27	12	20	37
	GAAQ+JSM (среднее значение)	82	55	81	42

бильная или прогрессирующая болезнь»: «Антиген-специфические лимфоциты in vitro — да и Общее количество введенных клеток — низкое», «Индукторы созревания ДК — все кроме Flt3L и GM – CSF».

Из табл. 5 видно, что применение метода GAAQ для отбора значимых признаков существенно увеличивает находимых причинно-следственных связей. Количество найденных причин также больше при использовании связки ДСМ+GAAQ, что повышает вероятность нахождения среди них полезных для специалистов в данной предметной области.

Заключение

В настоящей статье описан оригинальный подход к излечению и анализу информации из научных статей. Представлены результаты работы экспериментальной программной системы, реализующей основные этапы сбора и извлечения информации. Рассмотрена задача по определению успешности лечения пациентов различными типами дендритноклеточных вакцин. Была собрана тестовая коллекция документов из более чем 70 источников. Были проведены эксперименты по разделению и классификации пациентов, а также построению гипотез о наличии причинно-следственных связей между свойствами пациентов различных групп.

Литература

1. https://clinicaltrials.gov
2. http://www.ncbi.nlm.nih.gov/pubmed
3. *Grishman R.* // TIPSTER Text Architecture Design. Version 3.1. New York, NYU, 1998.
4. Assessment of Dendritic Cell Therapy Effectiveness Based on the Feature Extraction from Scientific Publications / A. Y. Lupatov [et al.] // Proceedings of the International Conference on Pattern Recognition Applications and Methods. 2015. V. 2. P. 270–276.
5. *Kieninger T. G.* Table structure recognition based on robust block segmentation // Photonics West'98 Electronic Imaging. 1998. P. 22–32.
6. *Breiman L. et al.* Arcing classifier (with discussion and a rejoinder by the author) // The annals of statistics. 1998. V. 6. № 3. P. 801–849.
7. *Pedregosa F. et al.* Scikit-learn: Machine learning in python // The Journal of Machine Learning Research. 2011. V. 12. P. 2825–2830.
8. *Панов А. И.*, *Швец А. В.*, *Волкова Г. Д.* Метод извлечения причинно – следственных связей с использованием оптимизированных баз фактов // Искусственный интеллект и принятие решений. 2015. № 1. С. 27–34.
9. *Финн В. К.* Об определении эмпирических закономерностей посредством ДСМ метода автоматического порождения гипотез // Искусственный интеллект и принятие решений. 2010. № 4. С. 41–48.
10. *Панов А. И.*, *Швец А. В.* Эволюционный метод покрытий для составления базы фактов ДСМ метода // Четырнадцатая национальная конференция по искусственному интеллекту с международным участием КИИ-2014 (24–27 сентября 2014 г., г. Казань, Россия): Труды конференции. 2014. С. 323–330.
11. *Панов А. И.* Выявление причинно-следственных связей в данных психологического тестирования логическими методами // Искусственный интеллект и принятие решений. 2013. № 1. С. 24–32.
12. *Kuznetsov S. O.*, *Obiedkov S. A.* Comparing performance of algorithms for generating concept lattices // Journal of Experimental and Theoretical Artificial Intelligence. 2002. V. 14. P. 189–216.

Бойко Анна Александровна. Н. с. ФГБНУ «Институт биоорганической химии им. академиков М. М. Шемякина и Ю. А. Овчинникова РАН». К. б. н. Окончила в 2006 г. ГБОУ ВПО РНИМУ им. Н. И. Пирогова. Количество печатных работ: 9. Область научных интересов: иммунология, клеточная биология, онкология. E-mail: boyko_anna@mail.ru

Кайдина Алиса Михайловна. М. н. с. ФГБНУ «НИИ биомедицинской химии имени В. Н. Ореховича». Окончила в 2008 г. ГБОУ ВПО РНИМУ им. Н. И. Пирогова. Количество печатных работ: 5. Область научных интересов: клеточная биология, раковые стволовые клетки, онкология. E-mail: alisa.kaydina@ibmc.msk.ru

Ким Ян Сергеевич. М. н. с. ФГБНУ «НИИ биомедицинской химии имени В. Н. Ореховича». Окончил в 2011 г. РХТУ им. Д. И. Менделеева. Количество печатных работ: 3. Область научных интересов: онкология, раковые стволовые клетки, иммунология. E-mail: yankimhcc@gmail.com

Лупатов Алексей Юрьевич. В. н. с. ФГБНУ «НИИ биомедицинской химии имени В. Н. Ореховича» РАМН. Окончил в 1989 г. МГУ имени М. В. Ломоносова. К. б. н. Количество печатных работ: 40. Область научных интересов: биоинформатика, онкология, иммунология. E-mail: alupatov@inbox.ru

Панов Александр Игоревич. Н. с. ФИЦ ИУ РАН. К. ф.-м. н. Окончил в 2009 г. Новосибирский ГУ, в 2011 г. МФТИ. Количество печатных работ: 28. Область научных интересов: методы машинного обучения, распознавание образов, когнитивное компьютерное моделирование, мультиагентные системы. E-mail: pan@isa.ru

Суворов Роман Евгеньевич. Аспирант ФИЦ ИУ РАН. Окончил в 2013 г. Рыбинский государственный авиационный технический университет. Количество печатных работ: 14. Область научных интересов: анализ данных, машинное обучение, распределенные вычисления. E-mail: rsuvorov@isa.ru.

Швец Александр Валерьевич. М. н. с. ФИЦ ИУ РАН. Окончил в 2011 г. Сибирский федеральный университет. Количество печатных работ: 16. Область научных интересов: компьютерная лингвистика, математическое моделирование, методы оптимизации, искусственный интеллект. E-mail: shvets@isa.ru