Технологии Data Mining в Медицине

Рамиз Алыгулиев
Институт Информационных Технологий НАНА, Баку, Азербайджан
r.aliguliyev@gmail.com

***Аннотация*– В статье описаны цели, задачи и этапы обнаружения знаний в базах данных (KDD) и интеллектуального анализа данных (data mining). Исследована роль технологии data mining в анализе медицинских данных и перечислены ее этапы. Указаны проблемы, ограничения и тенденции применения технологии data mining к медицинским данным.**

Ключевые слова*– *Knowledge Discovery in Database, data mining, medical data mining.

I. ВВЕДЕНИЕ

Современные больницы хорошо оснащены информационно-коммуникационными технологиями и другими средствами по сбору данных, которые обеспечивают сравнительно недорогое средство для сбора и хранения данных в меж- и внутри- госпитальных информационных системах. Обширное количество данных, собранных в медицинских базах, требует специальных инструментов для хранения и доступа к ним, а также для анализа и эффективного использования. Такое увеличение объема данных вызывает большие трудности при извлечении полезной информации для поддержки принятия решений из-за того, что традиционные методы не в состоянии обрабатывать такой объем данных. В таком случае методы эффективного компьютерного анализа незаменимы. Чтобы удовлетворить эту потребность, медицинская информатика использует технологии KDD (Knowledge Discovery in Database – обнаружение знаний в базах данных), разработанные в новой междисциплинарной области, охватывающей статистику, распознавание образов, машинное обучение, а также инструменты визуализации для анализа данных и обнаружения закономерностей в сырых данных. Термин KDD впервые был введен в середине 1990-х годов с целью извлечения знаний из базы данных [3].

II. ОБНАРУЖЕНИЕ ЗНАНИЙ В БАЗАХ ДАННЫХ

“KDD – это нетривиальный процесс выявления действительных, новых, потенциально полезных, в конечном счете, понятных закономерностей в данных”. KDD представляет собой автоматический, поисковый анализ и моделирование больших хранилищ данных.

Процесс KDD обычно состоит из следующих этапов [3, 9–11]:

1) **Понимание предметной области и постановка цели.** Этот шаг готовит почву для понимания того, что должно быть сделано с многочисленными решениями (о преобразовании, алгоритме, представлении и т.д.). Людям, которые отвечают за проект KDD, необходимо понять и определить цели конечного пользователя и окружающей среды, в которой процесс обнаружения знаний будет происходить.

2) **Выбор и создание набора данных**. Имея определенную цель, должны быть определены данные, которые будут использованы для обнаружения знаний. Этот процесс включает в себя выбор способов поиска, выявление доступных данных и интегрирование всех данных в один набор для обнаружения знаний, включая также те атрибуты, которые будут рассмотрены.

3) **Очистка и предварительная обработка данных**. Этап предназначен для повышения надежности данных. Он включает в себя очистку данных, таких как обработка пропущенных (недостающих) значений, удаление шумов и выбросов.

4) **Преобразование данных**. Вычислительная эффективность является одной из проблем анализа данных. Исследователи находятся под влиянием принципа Оккама, который может быть интерпретирован как “чем проще, тем лучше”. Методы преобразования данных включают уменьшение размерности (например, выбор и извлечение признаков, выборка записей), а также преобразование атрибутов (например, дискретизация численных атрибутов и функциональное преобразование). Этот шаг часто имеет решающее значение для успеха всего процесса KDD, но, как правило, зависит от конкретного случая. Например, в медицинских обследованиях отношение атрибутов часто может быть более важным фактором, чем каждый из них сам по себе. Предложены четыре подхода для уменьшения размерности набора данных. Первый подход заключается в использовании соответствующего источника знаний, таких как онтология или терминология, например SNOMED. Второй подход является субъективным, где эксперты в предметной области выполняют функцию выбора признаков. Третий подход использует алгоритмы интеллектуального анализа данных для оценки прогностической силы каждого атрибута или их комбинации. Последний – четвертый подход использует информационные меры, например, прироста информации для выбора соответствующих признаков. Этот подход измеряет важность каждой переменной по отношению к конкретной цели, но он не может обнаружить эффект комбинации подмножества переменных.

5) **Data mining**

5.1) *Выбор подходящей задачи data mining.* Принимается решение, какую задачу data mining использовать - классификацию, регрессию или кластеризацию? Это во многом зависит от целей KDD, а также от предыдущих этапов. Data mining предназначена

для достижения двух основных целей: прогнозирования и описания. Прогноз часто упоминается как контролируемый интеллектуальный анализ данных, в то время как описательный data mining включает в себя аспекты неуправляемости и визуализации.

5.2) *Выбор алгоритма(ов) data mining*. Имея стратегию, должно быть принято решение о тактике. Этап включает в себя выбор конкретного метода, который будет использован для поиска закономерностей. Этот подход пытается понять условия, при которых алгоритм data mining является наиболее подходящим. Каждый алгоритм имеет параметры и тактику обучения.

5.3) *Использование алгоритмов data mining*. Путем настройки параметров управления алгоритмов необходимо использовать алгоритмы несколько раз, пока не будет получен удовлетворительный результат.

6) **Оценка / интерпретация выявленных закономерностей.** На данном этапе в отношении целей, определенных на первом этапе, проводятся оценка и интерпретация выявленных закономерностей (шаблонов). При выявлении закономерностей во внимание принимаются шаги предварительной обработки в отношении их влияния на результаты алгоритма(ов) data mining. С целью дальнейшего использования, на этом этапе, а также ведется документация обнаруженных знаний.

7) **Использование обнаруженных знаний.** Теперь мы готовы включить обнаруженные знания в другую систему для дальнейших действий. Знание становится активным в том случае, если мы можем внести изменения в систему и измерить эффекты. На самом деле успех этого шага определяет эффективность всего процесса KDD. Есть много проблем на этой стадии, такие как потери в “лабораторных условиях”. Например, знание было извлечено из некоторых статических данных, но теперь данные становятся динамическими. Структуры данных могут меняться (некоторые атрибуты становятся недоступными), область данных может изменяться (например, атрибут может иметь значение, которое ранее не предполагалось). Следует отметить, что уровень использования обнаруженного знания в медицинской практике все еще остается низким.

Если исходные данные качественные, алгоритм и конечная цель известны, в процессе KDD некоторые из этих шагов (3, 4, 7) необязательны; то важно отметить промышленные инициативы в направлении стандартизации процесса KDD. Например, модель 5A (**A**ssess – оценка, **A**ccess – доступ, **A**nalyze – анализировать, **A**ct – действовать и **A**utomate – автоматизировать) компании SPSS и SEMMA (**S**ample – образец, **E**xplore – исследовать, **M**odify – изменять, **M**odel – модель, **A**ssess – оценка) из SAS [7].

III. ИНТЕЛЛЕКТУАЛЬНЫЙ АНАЛИЗ ДАННЫХ

Data Mining – это процесс обнаружения в “сырых” данных ранее неизвестных нетривиальных практически полезных и доступных интерпретации знаний, необходимых для принятия решений в различных сферах человеческой деятельности. Data mining представляет собой мультидисциплинарную область на пересечении технологии базы данных, математической статистики, машинного обучения и распознавания образов [9–11]. Методы data mining можно рассматриваться как развитие методов традиционной математической статистики. Однако между ними имеется существенное различие. В отличие от современных методов data mining традиционные методы математической статистики используются в основном для заранее сформулированных гипотез. А в основу технологии data mining положена концепция шаблонов, которые отражают фрагменты многоаспектных взаимоотношений в сырых данных.

Data mining является одним из шагов KDD, который состоит из нескольких этапов. В литературе определены следующие основные этапы data mining [2]:

1) ***Предварительная обработка данных***. Целью этого этапа является подготовка исходных данных для получения общего обзора данных и их последующего анализа. На практике от 60% до 90% времени, как правило, тратится на понимание данных и их подготовку. Этот этап состоит из следующих шагов:

1.1) *Описание данных и абстракция*. Статистические методы, такие как мода, медиана, среднее значение и т.д., гибридные методы – генетический алгоритм + корреляция могут рассматриваться в этом подходе.

1.2) *Очистка данных*. Под этим подразумевается очистка данных от шума и выбросов. Шум является случайной частью ошибки и должен быть устранен. Выбросы имеют отличающееся от других данных поведение и должны быть обнаружены. В статистике выбросы – это значения, резко отличающиеся от других значений в собранном наборе данных. Пропущенное значение означает, что некоторые ячейки данных пустые. Исключение, оценка и игнорирование могут решить эту проблему.

1.3) *Интеграция данных.* Объединение двух или более наборов данных для создания единого набора данных.

1.4) *Преобразование атрибутов.* Преобразование всех значений специальной переменной до требуемого масштаба или значения.

1.5) *Дискретизация и бинаризация.* Подготовка непрерывных данных для классификации и алгоритмов ассоциации.

1.6) *Сжатие данных*. Сокращение записей или столбцов, чтобы достичь более простых и интерпретируемых моделей, сократить время и память, устранить ненужные признаки и избежать “проклятия размерности”. В литературе использовано несколько методов для сокращения данных, таких как выбор подмножества признаков.

2) ***Моделирование данных***. На этапе моделирования выявляются отношения между данными для извлечения закономерностей. На этом этапе все задачи можно разделить на прогностические и описательные категории. Прогностические алгоритмы делятся на классификации и регрессии в зависимости от типа целевого переменного (дискретного или непрерывного). Описательные алгоритмы классифицируются как кластеризация и извлечение ассоциативных правил.

3) ***Постобработка данных***. На данном этапе визуализируются и оцениваются извлеченные знания, т.е. интерпретируются результаты. Визуализация является существенным этапом в data mining. Здесь также принимается во внимание применение полученных знаний в бизнес-приложениях.

Три основные категории стратегий data mining представлены в литературе: обучение с учителем (контролируемое обучение), обучение без учителя (неконтролируемое обучение) и обучение с частичным привлечением учителя (полуконтролируемое обучение).

Основные задачи data mining следующие [5, 11]*:*

1. Кластеризация.
2. Классификация.
3. Извлечение ассоциативных правил.
4. Обнаружение аномалий.
5. Регрессия.
6. Абстракция.

Широко распространенные алгоритмы data mining. Для решения вышеперечисленных задач предложено множество новых алгоритмов и модификаций существующих алгоритмов data mining. На международной конференции по интеллектуальному анализу данных (IEEE International Conference on Data Mining), которая проходила в декабре 2006 года в Гонконге, экспертами были определены 10 основных алгоритмов data mining [13]: 1) C4.5; 2) K-Means; 3) SVM; 4) Apriori; 5) EM; 6) PageRank; 7) AdaBoost; 8) *k*NN; 9) Naive Bayes и 10) CART.

IV. ИНТЕЛЛЕКТУАЛЬНЫЙ АНАЛИЗ МЕДИЦИНСКИХ ДАННЫХ

Обнаружение знаний в медицинских данных и здравоохранении сложнейшая и критически важная задача. Обнаружение знаний описывает процесс автоматического поиска больших объемов данных для обнаружения среди них закономерностей, которые могут быть рассмотрены как дополнительные знания о данных. Обнаруженные знания могут быть использованы для дальнейших манипуляций и открытий [12].

Технология data mining добилась значительных успехов в области медицинских исследований и клинической практике [8]. Об этом свидетельствует тот факт, что в последние годы термин data mining становится все более популярным в биомедицинских исследованиях. Например, в течение последних 10 лет число работ, имеющих в своем названии и ссылках “data mining”, в MEDLINE увеличилось в 10 раз. Деятельность междисциплинарных исследователей с целью поощрения методов интеллектуального анализа клинических данных, вероятно, будет одной из причин этого взрыва [**7**].

Medical data mining выявляет эффективное знание среди сырых данных, которое является необходимым условием для точного принятия решений. Различные типы моделей data mining использовались в прошлом для представления интересных фактов и скрытых закономерностей, а также тенденции в наборе медицинских данных с широким применением в медицинской практике [8].

В зависимости от точки зрения авторов в литературе можно встретить несколько определений medical data mining. Некоторые из этих определений сосредоточены на алгоритмах data mining, в то время как другие сосредоточены на областях медицины и болезней. Несмотря на такое разногласие, в литературе в основном принято следующее определение: “**Medical data mining** – это процесс извлечения неявной, потенциально полезной и новой информации из медицинских данных для повышения точности, уменьшения времени и затрат, построения системы поддержки принятия решений с целью пропаганды здорового образа жизни” [2, 10].

Как видно, это определение состоит из трех частей:

(1) **Data mining**
(2) **Медицинский характер**: использование медицинских данных и применение извлеченной модели к медицине.
(3) **Цель.** Определены четыре цели:
(3a) повышение эффективности и уменьшение человеческих ошибок (28%);
(3b) уменьшение времени и затрат (17%);
(3c) система поддержки медицинских решений (27%);
(3d) извлечение скрытых знаний (28%).

Числа в скобках указывают процентное соотношение целей применения data mining в медицине, т.е. с какой целью методы data mining применяются в медицине.

Medical data mining состоит из шести шагов [2]:

1. **Понимание проблемы**: на начальном этапе, с медицинской точки зрения, определяются проектные задачи, требования и ограничения, затем они преобразуются в задачу data mining. План проекта и стратегия проверки производятся как последний этап в этом шаге.

2. **Понимание данных**: данные собираются из обширных источников, таким образом достигается их обзор.

3. **Предварительная обработка данных**: на этом этапе выполняются описание и абстракция, очистка данных, интеграция, преобразование атрибутов, дискриминация, бинаризация и сокращение объема данных.

4. **Обработка данных**: алгоритмы, которые решают задачу data mining, генерируют различные модели. На этом этапе выбирается лучшая модель. Может потребоваться возврат к предыдущему шагу, если некоторые алгоритмы нуждаются в специальной предварительной обработке данных.

5. **Постобработка данных**: достоверность модели оценивается проверкой удовлетворения медицинским целям, которые могут быть необходимы для переоценки всего процесса извлечения знаний. Визуализации и

интерпретируемые модели наряду с экспертной помощью способствуют верификации модели.

6. **Развертывание**: подготовленная модель должна быть полезной для медицинских целей. Например, если мы рассматриваем систему data mining как встроенную, то она должна быть встроена в основную систему.

Data mining играет важную роль и в здравоохранении. Это обусловлено следующими причинами:

- генерируются огромные и сложные объемы данных в здравоохранении;
- неавтоматизированный анализ становится сложным и нецелесообразным;
- data mining позволяет генерировать информацию, которая может оказаться полезной для всех заинтересованных сторон, в том числе и пациентов путем предоставления эффективных методов лечения и передового опыта;
- наличие страхового мошенничества и злоупотреблений побуждает страховщиков использовать data mining.

Области применения data mining в здравоохранении можно сгруппировать следующим образом:

- эффективность лечения;
- управление здравоохранением;
- совершенствование управления взаимоотношениями с клиентами;
- мошенничество и обнаружение злоупотреблений.

Медицинские данные являются одним из самых сложных типов информации для анализа. Этот тип данных генерируется несколькими источниками: медицинские обследования, визуализация и испытания. Согласно [9] медицинские данные могут быть разделены на клинические и временные.

Ниже представлены различные типы медицинских данных:

1) клинические данные текстового характера и качественного формата;

2) пробные данные с числовым характером и количественным форматом;

3) данные изображений, такие как МРТ (магнитно-резонансная томография) и радиология;

4) данные УЗИ (ультразвуковое исследование): эхо и сонография;

5) последовательность или временные ряды данных;

6) данные сигнала, такие как ЭЭГ (электроэнцефалограмма) и ЭКГ (электрокардиограмма), которые также имеют характеристики, в виде данных временных рядов;

7) генетические, микромассивы белка данных, которые имеют маленькие записи и большие переменные.

Ниже приведены основные моменты уникальности медицинских данных [1, 4, 6]:

- **Неоднородность медицинских данных**
  - объем и сложность медицинских данных
  - интерпретация врача
  - чувствительность и специфичность анализа
  - плохая математическая характеристика
  - каноническая форма
  - решение: стандартные словари, интерфейсы между различными источниками интеграции данных; дизайн электронных историй болезни
- **Этические, правовые и социальные вопросы**
  - право собственности на данные
  - страх исков
  - конфиденциальность и безопасность данных человека
  - ожидаемые выгоды
  - административные вопросы
- **Статистическая философия**
  - засада в статистике
  - интеллектуальный анализ данных как надмножество статистики
  - интеллектуальный анализ данных и процесс обнаружения знаний
- **Особый статус медицины**

Медицина имеет особый статус в науке, философии и повседневной жизни. Медицина является необходимостью, а не просто необязательной роскошью, удовольствием или удобством.

Несмотря на то, что интеллектуальный анализ данных широко применяется в медицине, однако, из-за ее специфических особенностей применение технологии data mining в указанной сфере сталкивается с некоторыми ограничениями [1, 2, 8, 12]:

- из-за распространённости данных (клинические, административные, страховые компании, лаборатории и т.д.) доступ технология data mining к ним может быть ограничен;
- данные могут быть неполными, поврежденными, зашумленными или противоречивыми;
- проблемы конфиденциальности данных, а также этические, правовые и социальные вопросы;
- многие закономерности, найденные в data mining, могут быть результатом случайных флуктуаций, поэтому много таких закономерностей может оказаться бесполезным;
- medical data mining требует не только специфических медицинских знаний, но также знания в области технологии data mining.
- data mining требует институциональных обязательств и финансирования.

V. ТЕНДЕНЦИИ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА МЕДИЦИНСКИХ ДАННЫХ

Всю деятельность в области медицины можно разделить на шесть областей, известных как “медицинские задачи”: скрининг, диагностика, лечение, прогноз, мониторинг и менеджмент. Анализ литературы показал, что среди этих задач в последнее время большее внимание уделено диагностике, лечению и скринингу. Как следствие, лечение, диагностика и скрининг могут рассматриваться как популярные задачи в будущем [2].

Согласно данным Всемирной организации здравоохранения, сердечно-сосудистые заболевания, рак и диабет являются заболеваниями высокого риска для

жизни. В последние годы уровень смертности от этих заболеваний растет, что сделало их глобальной проблемой. Таким образом, в будущем больше внимания будет уделено этим заболеваниям. Следовательно, в настоящее время применение технологии data mining к этим заболеваниям является более распространенным явлением.

Как уже сказано выше, процесс data mining включает в себя три этапа. Большинство работ на этапе предварительной обработки направлено на сокращение данных. На этапе моделирования данных среди различных алгоритмов из-за своей простоты и интерпретируемости наиболее популярно дерево решений. В последнее время более эффективные алгоритмы, такие как SVM, также приобретают большую популярность. Растет также интерес к применению SVM, нечетких систем и регрессий. Методы оценки производительности на этапе постобработки данных должны быть выбраны с учетом типа извлеченных знаний и данных.

VI. ЗАКЛЮЧЕНИЕ

Анализ показал, что в области medical data mining основные проблемы связаны с данными. Эти проблемы можно группировать в три класса [2]:

1. **Проблемы предметной области:**
   - математическое описание
   - отличие от статистики
   - этические и правовые
   - конфиденциальность
2. **Общие проблемы:**
   - неоднородность
   - высокая размерность
   - скошенность (асимметричность)
   - пропущенное значение
   - интеграция
   - большие данные
   - шум, выбросы
3. **Проблемы сбора данных:**
   - достоверность
   - несоответствие
   - стандарт сбора
   - стандарт передачи данных

Другая не менее важная проблема связана с тем, что, в отличие от других предметных областей, в medical data mining медицинский эксперт играет важную роль в понимании проблемы и данных и в постобработке данных:

- определение медицинских целей;
- определение проблемы с медицинской точки зрения;
- представление основной информации о проблеме;
- представление метрик;
- разъяснение данных и их значение;
- оценка достоверности данных;
- сбор данных для финального тестирования;
- проверка добытых знаний с жизненным опытом;
- внесение предложений по исправлению шагов и деятельности.

ЛИТЕРАТУРА

[1] K.J. Cios & G.W. Moore, “Uniqueness of medical data mining”, Artificial Intelligence in Medicine, vol.26, nos.1–2, pp.1–24, 2002.

[2] N. Esfandiari, M.R. Babavalian, M.E. Moghadam, & V.K. Tabar, “Knowledge discovery in medicine: current issue and future trend”, Expert Systems with Applications, vol.41, pp.4434–4463, 2014.

[3] U.M. Fayyad, G. Piatetsky-Shapiro, & P. Smyth, “The KDD process for extracting useful knowledge from volumes of data”, Communications of the ACM, vol.39, no.11, pp.27–41, 1996.

[4] P.R. Harper, “A review and comparison of classification algorithms for medical decision making”, Health Policy, vol.71, pp.315–331, 2005.

[5] T. Hays, Medical data mining, pp.1–33, 2012. http://www.nist.gov/healthcare/upload/Hays-Medical-Data-Mining-slides-for-web.pdf

[6] M. Holena, A. Sochorova, & J. Zvarova, “Increasing the diversity of medical data mining through distributed object technology”, Studies in Health Technology and Informatics, vol.68, pp.442–477, 1999.

[7] J. Iavindrasana, G. Cohen, A. Depeursinge, H. Müller, R. Meyer, & A. Geissbuhler, “Clinical data mining: a review”, IMIA Yearbook of Medical Informatics, pp.121–133, 2009.

[8] S.G. Jacob & R.G. Ramani, “Data mining in clinical data sets: a review”, International Journal of Applied Information Systems, vol.4, no6, pp.15–26, 2012.

[9] N. Lavrac, “Machine learning for data mining in medicine”, Lecture Notes in Computer Science, vol.1620, pp.47–62, 1999.

[10] N. Lavrac, “Selected techniques for data mining in medicine”, Artificial Intelligence in Medicine vol.16, no.1, pp.3–23, 1999.

[11] N. Lavrac & B. Zupan, “Data mining in medicine”, Data Mining and Knowledge Discovery Handbook, 2nd ed. (eds. O. Maimon & L. Rokach) (pp.1111–1136), Springer, 2010.

[12] V. Paramasivam, T.S. Yee, S.K. Dhillon, & A.S. Sidhu, “A methodological review of data mining techniques in predictive medicine: an application in hemodynamic prediction for abdominal aortic aneurysm disease”, Biocybernetics and Biomedical Engineering, vol.34, pp.139–145, 2014.

[13] X. Wu, V. Kumar, J.R. Quinlan et al., “Top 10 algorithms in data mining”, Knowledge and Information Systems, vol.14, no.1, pp.1–37, 2008.

Приложение

Журналы с самой высокой частотой публикации в medical data mining:

1. Applied Soft Computing
2. Artificial Intelligence in Medicine
3. Computer Methods and Programs in Biomedicine
4. Computers in Biology and Medicine
5. Decision Support Systems
6. Expert Systems
7. Expert Systems with Applications
8. IEEE Engineering in Medicine and Biology
9. IEEE Transactions on Biomedical Engineering
10. IEEE Transactions on Information Technology in Biomedicine
11. IEEE/ACM Transactions on Computational Biology and Bioinformatics
12. Information Sciences
13. Journal of Biomedical Informatics
14. Journal of Medical Systems
15. Knowledge-Based Systems