

# ИСПОЛЬЗОВАНИЕ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА ДАННЫХ В МЕДИЦИНСКИХ УЧРЕЖДЕНИЯХ.

Ербол Нисанбаев

## ВСТУПЛЕНИЕ

В этом докладе изложены возможные перспективы применения Интеллектуального анализа данных (ИАД) в лечебных учреждениях. Также рассматриваются необходимые условия для внедрения ИАД и трудности, которые при этом могут возникнуть.

В наше время огромное количество информации хранится и обрабатывается в электронном виде. В случае с лечебными учреждениями информация об анализах, процедурах, назначениях и ходе лечения больных образует огромные массивы данных, которые зачастую попадают в архив после выписки больного, где и хранятся до уничтожения. Таким образом, эта информация практически никак не используется и не анализируется. Исключение составляют лишь статистические подсчеты количества больных и их распределение по видам заболеваний.

В то же время информация, хранящаяся в тысячах историях болезней, может быть использована для нахождения новых взаимосвязей между различными факторами: эффективность алгоритмов лечения при определенных диагнозах, воздействие тех или иных лекарственных средств на общее течение болезни и т.д.

Для того чтобы понять каким именно образом можно использовать ИАД в лечебных учреждениях, необходимо ближе рассмотреть понятие ИАД и методы данной дисциплины

## ИНТЕЛЛЕКТУАЛЬНЫЙ АНАЛИЗ ДАННЫХ

### Введение

ИАД стал очередной ступенью после онлайн-аналитической обработки данных (ОАОД) в развитии анализа в реляционных базах данных и информационных хранилищах [6]. После внедрения реляционных баз данных большое количество информации стало храниться в электронном виде. Так согласно исследованию GTE Labs только научные институты генерируют ежедневно до терабайта новых данных. Можно представить насколько значительный объем информации может быть собран всего за один год. Но если эта информация лишь находится на хранении и никак не обрабатывается, вся информационная система превращается в своеобразное кладбище данных – которое редко кто-либо посещает [6].

Однако появившиеся методы ИАД позволяют проводить глубинные исследования больших объемов данных с целью поиска и выведения скрытых и неочевидных закономерностей. В отличие от ОАОД, где задается определенный параметр поиска или проверяется заранее установленное предположение, ИАД позволяет находить новые, еще неизвестные, но существующие, гипотезы и взаимосвязи [1].

### Методы ИАД

В ИАД существуют две основные задачи: *прогнозирование* и *описание*.

Задача *прогнозирования* состоит в создании общей модели всей системы, основываясь лишь на ограниченном наборе данных.

В то время как целью *описания* является обнаружение новой нетривиальной информации, основываясь на существующих данных.

Обе задачи достигаются с помощью использования следующих шести основных методов [7]:

1. Классификация – заключается в поиске группы функций (или моделей), которые смогут классифицировать неизвестный объект на принадлежность к одному из существующих классов.
2. Регрессионный анализ – заключается в поиске функций, с помощью которых, было бы возможно предсказать значение целевой (входной) переменной.
3. Кластеризация – занимается поиском определенных сегментов (кластеров) по которым можно было бы распределить все исходные переменные. В отличие от классификации, мы заранее не имеем определенных классов, нашей задачей является их поиск. Кластеры формируются таким образом, чтобы объекты одного кластера были максимально схожи друг с другом и максимально не похожи на объекты других кластеров.
4. Суммирование – включает в себя методы поиска краткого описания для определенного множества данных.
5. Моделирование зависимостей – заключается в поиске моделей, которые могли бы описать существенные зависимости между значениями переменных или значениями тех или иных свойств в массиве данных.
6. Выявление отклонений и изменений – поиск наиболее значительных изменений во множестве данных.

Итак, мы рассмотрели основные понятия и принципы ИАД. Теперь мы можем рассмотреть, насколько методы и принципы ИАД применимы к медицинской сфере.

## **ИНТЕЛЛЕКТУАЛЬНЫЙ АНАЛИЗ ДАННЫХ В МЕДИЦИНЕ**

### **Область применения**

В практической медицине можно обозначить два основных направления применения ИАД:

- 1) Помощь в установлении диагноза больного на основе полученных данных (жалоб, анамнеза, лабораторных исследований и т.д.)
- 2) Поиск новых взаимосвязей между различными факторами лечения (поиск новых медицинских знаний) с помощью анализа медицинской информации (историй болезней и др.).

Рассмотрим вышеуказанные направления подробнее.

#### **1) Установление диагноза.**

При установлении клинического диагноза, врач основывается на анамнезе, симптомах заболевания и результатах всевозможных обследований [5]. Применение методов ИАД в этом процессе приведет к более точной и оперативной установке диагноза. В отличие от человеческих возможностей, средства ИАД обладают способностью одновременно рассматривать и анализировать более широкий спектр возможных патологий и использовать все полученные данные для установки диагноза.

90-е годы прошлого века характеризуются ростом пропасти между огромными массивами хранимых данных и количеством полезной информации извлекаемой из них. Это и привело к активному развитию различных направлений в ИАД, нацеленных на прогнозирование и поиск неизведанных знаний в базах данных.

#### **2) Поиск новых знаний.**

В этом направлении мы используем средства ИАД для извлечения новых медицинских знаний из собраний историй болезней и другой медицинской информации. При этом, необходимо учитывать, что зачастую исследуемая информация является неполной или ложной [8].

Для медицины, важно чтобы результаты ИАД были точны и понятны. Для улучшения этих показателей, становится важным непосредственное участие экспертов в процессе ИАД и в оценке результатов анализов. Только точные закономерности и связи, интерпретированные на необходимом уровне с помощью экспертов в области медицины, могут оказаться полезными для практического врача в процессе принятия ежедневных решений.

Качество, полученной с помощью ИАД, информации оценивается *исполнением* (точность классификации и прогнозирования), *понятностью* и *значимостью* извлеченных знаний.

Рассмотрев основные направления в использовании ИАД в медицине, стоит взглянуть на методы ИАД, которые могут быть использованы в этих направлениях [8].

## Методы, применимые в медицине

### 1) Вывод правил методом индукции.

Данный метод работает по следующему принципу: системе предоставляется набор классифицированных примеров, проанализировав которые, система генерирует набор if-then правил.

IF УСЛОВИЯ Then ВЫВОДЫ

Сторона условий содержит необходимые обстоятельства, при которых могут быть выведены заключения в стороне выводов. Рассмотрим этот метод подробнее на примере диагностики гипертонической болезни.

```
IF Систолическое_давление > 140
AND Диастолическое_давление > 80
AND Период_наблюдения => 1 месяц
AND Жалобы = головные_боли, сердечные_боли
THEN Гипертония
```

В данном методе, могут возникать трудности в вычислении вероятностей, например, вероятности того что новый случай будет правильно классифицирован тем или иным правилом. Этот показатель может быть повышен, если правило было сформировано, используя большое количество примеров. Однако, в реальной ситуации, количество примеров не всегда бывает достаточным для получения необходимой степени вероятности. В таких случаях, необходимо использовать методы вычисления вероятности, способные выдавать более точные результаты, имея небольшое количество примеров.

### 2) Анализ дерева решений.

Системы, использующие этот метод, генерируют деревья решений, основываясь на предоставленных записях параметров и их значений. Каждому из стволов дерева присваивается название того или иного параметра, в то время как веткам, идущим от стволов, присваиваются значения параметра. Также существуют ответвления, представляющие собой классы.

Процесс построения дерева на каждом этапе эвристически нацелен на поиск *самых информативных* параметров. Это делается для уменьшения количества тестов необходимых для классификации.

Рассмотрим простейший алгоритм построения дерева решений [6]

1. Создать ствол  $N$
2. Если все примеры принадлежат к одному классу  $K$ , тогда
3.     обозначить  $N$  как ответвление, обозначенное классом  $K$ ;
4. Если список параметров пуст, тогда
5.     обозначить  $N$  как ответвление, обозначенное наиболее часто встречающимся классом среди примеров // голос большинства
6. Выбрать *тестовый параметр* – параметр из списка, который является наиболее информативным
7. Обозначить ствол  $N$  названием *тестового параметра*.
8. Для каждого известного значения  $A$  *тестового параметра* // разделение примеров
9.     вырастить ветвь от ствола  $N$  соответствующую условию *тестовый параметр = A*
10.     пусть  $S_i$  будет разделом примеров, для которых значение *тестового параметра = A* // раздел
11.     если  $S_i$  является пустым множеством, тогда

12.           прикрепить ответвление, обозначенное наиболее часто встречающимся классом среди примеров.
13.    в остальных случаях прикрепить новый сгенерированный ствол

Этот метод очень хорошо развит и активно используется сообществом специалистов в ИАД. Одним из наиболее известных и широко используемых средств построения дерева решений является C4.5 (C5.0 – последняя версия). Эта система очень надежна и эффективна, а также обладает подробной документацией. Она с успехом может быть использована и в медицинской сфере.

### 3) Обучение по примеру (ОПП).

Алгоритмы ОПП, в отличие, например, от вывода правил методом индукции используют единичные примеры для задач классификации, а не обобщение всех примеров. Алгоритмы ОПП также называются «Ленивые алгоритмы обучения», так как они просто сохраняют все примеры и откладывают задачу обобщения примеров непосредственно до момента классификации.

Алгоритмы ОПП используют алгоритм *ближайшего соседа* для осуществления классификации. Алгоритм ближайшего соседа рассматривает параметры как измерения Евклидова пространства, а значения параметров – как точки в этом пространстве. В процессе классификации измеряется расстояние между точками пространства и группа точек наиболее приближенных друг к другу образуют новый класс.

## **Необходимые условия для внедрения ИАД**

Интеллектуальный анализ данных является сферой деятельности, которая тесно связана с достижениями в области информатизации и машинного обучения.

Для успешного внедрения ИАД в повседневную деятельность медицинских учреждений необходимо учитывать следующие факторы:

### *1) Программное обеспечение*

ИАД подразумевает под собой поиск знаний в существующих базах данных. Соответственно в лечебном учреждении должно быть установлено программное обеспечение, полностью поддерживающее процесс ведения медицинской документации в электронном виде. Примерами таких систем являются «Медиалог» разработки Пост Модерн Текнолоджи и система «Амулет» фирмы ЦентрИнвестСофт [2]. Эти продукты направлены на полную автоматизацию лечебных учреждений и объединяют в себе несколько взаимосвязанных подсистем.

Помимо вышеуказанного ПО, также необходимо позаботиться о приобретении ПО нацеленного непосредственно на ИАД. Такое программное обеспечение обычно имеет высокую стоимость, но будучи правильно использовано может дать значительный положительный результат. Одним из таких средств является Clementine, производства компании SPSS. При стоимости около трех тысяч долларов США (сентябрь 2007) это ПО предлагает комплексное решение по ИАД во множестве направлений, включая медицинское.

### *2) Оборудование.*

Для успешного внедрения системы ИАД в медицинском учреждении необходимо использовать соответствующее оборудование, способное эффективно работать с необходимым программным обеспечением. Это включает в себя следующие компоненты:

1. Диагностическое оборудование, способное взаимодействовать с программным обеспечением и поддерживающее моментальный обмен информации в электронном виде (рентгенограммы, кардиограммы, УЗИ, компьютерная томография и т.д.).
2. Клиентское оборудование – подразумевает персональные компьютеры (ПК), установленные непосредственно в каждом из отделений лечебного учреждения. Они используются для ввода и извлечения информации из базы данных. Важно обеспечить количество ПК – достаточное для оперативной и комфортной работы медицинского персонала.

3. Серверное оборудование – комплексные серверные решения, необходимые для эффективного хранения и управления огромным количеством медицинской информации.
4. Оборудование для ИАД - это отдельные вычислительные машины, деятельность которых будет ограничена лишь ИАД. Поскольку процесс поиска новых знаний требует значительных вычислительных мощностей, использование клиентского или серверного оборудования в ИАД может негативно сказаться на быстродействии и времени отклика информационной системы в целом.

### *3) Обучение персонала*

Немаловажным этапом во внедрении системы ИАД и сопутствующей медицинской информационной системы (МИС) является также обучение персонала. Переход с бумажной на электронную систему ведения документации должен быть постепенным и последовательным. Помимо работы с МИС, некоторые сотрудники должны будут уметь работать и с системой ИАД (задавать зону поиска, отбирать полезные знания и т.д.)

## **Ограничения ИАД**

Помимо тех преимуществ, которые приносит с собой внедрение ИАД, у этого подхода существует также ряд слабых мест, которые необходимо учитывать [7]. Так Питер Кой, обозреватель журнала Business Week подчеркивает следующие недостатки в существующих средствах ИАД:

1. Существует склонность развивать теории из «странных» связей, найденных в хранящихся данных
2. Порой становится возможным найти подтверждение практически любому убеждению, если позволить системе достаточно долго «копаться» в данных.
3. Чем больше параметров и факторов предоставлены системе для поиска, тем выше вероятность, что будет найдена какая-либо связь (не всегда полезная).

Таким образом, необходимо осознать, что хотя средства ИАД способны найти ранее неизвестные знания и тем самым оптимизировать многие процессы, не стоит рассматривать ИАД как волшебную палочку для легкого достижения той или иной цели.

Очень важную роль играет правильная постановка вопроса и правильный выбор исходных данных. Но помимо трудностей, встречающихся в ИАД в целом, существуют барьеры – специфичные для медицинского ИАД, рассмотрим их в следующем разделе.

## **ПРОБЛЕМЫ ВНЕДРЕНИЯ ИАД В ОБЛАСТИ МЕДИЦИНЫ**

ИАД информации о пациентах является одной из самых необходимых и в то же время сложных из всех видов анализа биологической информации [5]. Этот факт обосновывается рядом проблем, обоснованных следующими особенностями медицинской информации:

1. Неоднородность медицинской информации
2. Этические и социальные вопросы

Рассмотрим каждый из этих факторов в деталях:

### **1. Неоднородность медицинской информации**

Медицинская информация может быть собрана из множества источников. Рассмотрим основные причины неоднородного характера информации:

#### *а) Объем и сложность медицинской информации*

Необработанная медицинская информация очень объемна и неоднородна. Это включает в себя результаты анализов, ежедневные записи и т.д. Особенно большие объемы информации (до

нескольких гигабайт в день) занимают графические изображения рентгенограмм, компьютерная томография, ЭКГ и т.д.

*b) Интерпретация врача*

Обычно записи врачей написаны в свободной несистемной манере. Зачастую даже между специалистами в одной области медицины возникают разногласия по поводу применения того или иного термина в описании состояния больного. Это сильно усложняет процесс ИАД во врачебных записях.

*c) Плохое математическое представление*

По сравнению, например, с физикой или финансовой средой, понятия и обозначения в медицине трудно описать и вычислить математическими формулами и уравнениями.

*d) Каноническая форма*

Каноническая форма - это предпочтительное обозначение определенной концепции, включающее в себя все равнозначные обозначения. Так, например, понятие половины в математике обозначается  $1/2$ , и все равнозначные формы:  $3/6$ ,  $5/10$ ,  $125/250$ , в конце концов, приводятся к значению  $1/2$ . В медицине же, наблюдается отсутствие канонической формы во многих обозначениях. Так одно и то же заболевание может иметь несколько равнозначных названий. Это создает дополнительные трудности в ИАД медицинских записей.

## **2. Этические и социальные вопросы**

Так как медицинская информация собирается с реальных людей, существует вероятность злоупотребления этой информацией. Рассмотрим основные сложности, связанные с этим:

*a) Охрана частной информации*

Использование ИАД подразумевает под собой передачу огромного количества информации, включая частную информацию о больных и их историях болезней. Так как передача электронной информации через интернет небезопасна, должны быть рассмотрены альтернативные пути передачи и защиты большого количества информации.

*b) Ожидаемая выгода*

Любое использование информации о пациентах должно быть оправдано и обосновано. Недозволительно разрешать анализ данных о пациентах и их заболеваниях для неоправданных или вредоносных целей.

## **ЗАКЛЮЧЕНИЕ**

Медицина занимает особое место среди всех наук. От результатов исследований и методов лечения зависят человеческие жизни. Медицина является жизненной необходимостью, а не привилегией или особым удобством. Соответственно, всем решениям и исследованиям в медицинской области должно быть уделено особое внимание и применена тщательная проверка.

По вышенаписанному, можно судить, что уже проведена определенная работа в изучении интеллектуального анализа данных в медицине, хотя все ещё существует огромное поле для исследований. Несмотря на все существующие трудности, интеллектуальный анализ данных в медицине является одним из самых полезных и первоочередных направлений в ИАД. Ведь, правильный вывод, предоставленный в нужный момент, может означать улучшение в состоянии тяжелого пациента или, даже, спасение человеческой жизни.

Литература:

- [1] Буров К., (1999), «Обнаружение знаний в хранилищах данных», *Открытые системы N05-06/1999*.
- [2] Гусев А., Романов Ф., Дуданов И. (2005) «Медицинские информационные системы: анализ рынка», *PCWeek/Russian Edition, №47/2005*
- [3] Дюк Вячеслав, *Data Mining - интеллектуальный анализ данных*, Санкт-Петербургский институт информатики и автоматизации РАН.
- [4] Петровский Б. (1993), *Популярная Медицинская Энциклопедия, 2-е издание*, Советская Энциклопедия. стр. 180
- [5] Cios K., Moore G. (2002), «Uniqueness of Medical Data Mining», *Artificial Intelligence in Medicine, University of Colorado Denver*.
- [6] Han J., Kamber M. (2000), *Data Mining: Concepts and Techniques*, Morgan Kaufmann Publishers. Chapter 1.
- [7] Kantardzic M. (2003), *Data Mining: Concepts, Models, Methods, and Algorithm*”, John Wiley & Sons. Chapter 1, Appendix B.
- [8] Lavrac N (1998), «Data Mining in Medicine: Selected Techniques and Applications», *Second International Conference on the Practical Application of Knowledge Discovery and Data Mining, London*.
- [9] Lavrac N., Keravnou E., Zupan B. (1997), «Intelligent Data Analysis in Medicine», *Intelligent Data Analysis in Medicine and Pharmacology 1997*.