А.В.КУЗНЕЦОВА,
институт биохимической физики РАН им. Н.М.Эмануэля
О.В.СЕНЬКО,
вычислительный центр РАН

ВОЗМОЖНОСТИ ИСПОЛЬЗОВАНИЯ МЕТОДОВ Data Mining ПРИ МЕДИКО-ЛАБОРАТОРНЫХ ИССЛЕДОВАНИЯХ ДЛЯ ВЫЯВЛЕНИЯ ЗАКОНОМЕРНОСТЕЙ В МАССИВАХ ДАННЫХ

До сих пор у врачей привычным подходом при анализе медико-лабораторных данных остаются критерии Стьюдента, Фишера и метод линейных корреляций. Написано много книг по статистике, создано большое число компьютерных программ по статистической обработке медицинской информации. Медикам достаточно сложно разобраться во всем этом многообразии, выбрать подход или критерий, предназначенный именно для их конкретной задачи.

Между тем, математика и вычислительная техника не стоят на месте. Создаются новые алгоритмы, позволяющие не обращать внимание на характер распределения данных (как того требуют традиционные статистические методы), методы для анализа информации даже в условиях нелинейной зависимости признаков. Современная техника легко обрабатывает огромные массивы данных, выбирая из сотни признаков наиболее информативные. Доказательство того, что отобранные признаки достоверны, проводится с помощью алгоритмов перебора многих тысяч возможных вариантов. Это стало реальным благодаря колоссальному быстродействию новых машин.

Возникло целое направление в информатике, называемое **Data Mining**, что в переводе означает «добыча данных», а у нас принято переводить как «интеллектуальный анализ информации (данных)». Это технология выявления скрытых взаимосвязей внутри больших баз данных. В основу Data Mining (discovery-driven data mining) положена концепция шаблонов (паттернов), отражающих ***фрагменты*** многоаспектных взаимоотношений в данных [1, 2].

© А.В.Кузнецова, О.В.Сенько, 2005 г.

Приложения Data Mining успешно применяются в различных областях. В бизнесе, в том числе в розничной торговле и маркетинге, они позволяют компаниям добывать информацию, дающую конкурентные преимущества (иногда до 1000 %). В медицине с их помощью построены экспертные системы для постановки диагнозов на основе правил, описывающих сочетания различных симптомов разных заболеваний. Правила помогают выбирать показания (противопоказания), предсказывать исходы назначенного курса лечения. В молекулярной генетике и генной инженерии – это определение маркеров, под которыми понимаются генетические коды, контролирующие те или иные фенотипические признаки живого организма. Известно несколько крупных фирм, специализирующихся на применении Data Mining для расшифровки генома человека и растений. В прикладной химии эти методы используют для выяснения особенностей химического строения химических соединений.

Data Mining-анализ призван помогать в принятии решений. Для чего нужны не факты сами по себе, а знания – знания о закономерностях в наблюдаемых процессах. Чем специфичнее информация, тем полезнее она для принятия решений. Таким образом, Data Mining (DM) есть процесс обнаружения подобного рода полезных знаний. Причем необходимым требованием является обнаружение в сырых данных: ранее неизвестных, нетривиальных, практически полезных, доступных интерпретации знаний, полезных для принятия решений в различных сферах человеческой деятельности.

Методы Data Mining играют ведущую роль в областях со сложной системной организацией (к такой области несомненно относится исследование различных систем организма человека). Данные, с которыми имеет дело DM-анализ, могут быть неоднородны, гетерогенны, нестационарны и часто отличаются высокой размерностью. Такие данные называют также «сырыми данными» (raw data), и знания, выявляемые из них – «скрытыми знаниями» (hidden knowledge).

Типы закономерностей, которые позволяют выявлять методы Data Mining:

- ассоциация (выявление связи нескольких событий и оценка результативности воздействия на наборы параметров);
- последовательность (выявление временной связи между параметрами);
- классификация (выявление признаков, характеризующих группу, к которой принадлежит тот или иной объект, посредством обучения на уже классифицированных объектах, формулирование набора правил для каждой группы);
- кластеризация (самостоятельно выявляются однородные группы данных);
- прогнозирование (создание шаблонов, адекватно отражающих динамику поведения целевых показателей по временным рядам базы данных).

В основе подходов Data Mining лежат две технологии: машинное обучение и визуализация (визуальное представление информации). Обе технологии дополняют друг друга в процессе осуществления DM-анализа.

Визуализация используется для поиска исключений, общих тенденций и зависимостей. Качество визуализации определяется возможностями графического отображения значений данных путем изменения цветов, форм и других элементов, что упрощает выявление скрытых зависимостей.

Машинное обучение позволяет исследовать бо**о**льшее количество взаимосвязей данных, чем может человек, за счет использования различных методов: деревьев решений; ассоциативных правил; генетических алгоритмов; нейронных сетей.

Деревья решений предназначены для классификации данных, они используют весовые коэффициенты для распределения элементов данных на все более и более мелкие группы. Ме-

тод ассоциативных правил классифицирует данные на основе набора правил, подобных правилам в экспертных системах. Эти правила можно генерировать, используя процесс поиска и проверки комбинаций правил, или извлекать правила из деревьев решений. В нейронных сетях знания представлены в виде связей, соединяющих набор узлов. Сила связей определяет зависимости между факторами данных.

Data Mining является мультидисциплинарной областью, возникшей и развивающейся на базе достижений прикладной статистики, распознавания образов, методов искусственного интеллекта, теории баз данных и др. В программном обеспечении системы Data Mining представлены следующим образом:

Статистические пакеты. Оказались полезными главным образом для проверки заранее сформулированных гипотез (verification-driven data mining) и для «грубого» разведочного анализа, составляющего основу оперативной аналитической обработки данных (online analytical, OLAP). Большинство методов опираются на усредненные характеристики выборки, которые при исследовании реальных сложных жизненных феноменов часто являются фиктивными величинами. Хорошо описаны пакеты STATGRAPHICS, STATISTICA, STADIA [3, 4].

Предметно-ориентированные аналитические системы. Наиболее развиты системы в области исследования финансового рынка, так называемый «технический анализ»: прогноз динамики цен, выбор оптимальной структуры инвестиционного портфеля, основанный на различных эмпирических моделях динамики рынка. Эти методы максимально учитывают специфику приложения (профессиональный язык, индексы и пр.).

Искусственные нейронные сети. Здесь для предсказания значения целевого показателя используются наборы входных переменных, математических функций активации и весовых коэффициентов входных параметров. Выполняется итеративный обучающий цикл, нейронная сеть модифицирует весовые коэффициенты до тех пор, пока предсказываемый выходной параметр соответствует действительному значению. После обучения нейронная сеть становится моделью, которую можно применить к новым данным с целью прогнозирования. Основным недостатком в этом случае является необходимость иметь очень большой объем обучающей выборки. Кроме того, любая нейронная сеть представляет собой «черный ящик» и знания в виде нескольких сотен весовых коэффициентов, полученных с ее помощью, не поддаются анализу и интерпретации. Примеры – BrainMaker, NeuroShell, OWL.

Системы рассуждений на основе аналогичных случаев. Вывод путем сопоставления (Memory-based Reasoning, MBR) или вывод, основанный на прецедентах (Case-based Reasoning, CBR). Эти алгоритмы основаны на обнаружении некоторых аналогий в прошлом, наиболее близких к текущей ситуации, с тем, чтобы оценить неизвестное значение или предсказать возможные результаты (последствия). Эти методы называют еще методом «ближайшего соседа». В выборе решения они основываются на всем массиве доступных исторических данных, поэтому невозможно сказать, на основе каких конкретно факторов строятся ответы. Примеры: KATE tools (Франция), Pattern Recognition Workbench (США), КОРА (Россия).

Деревья решений и Алгоритмы классификации. Создается иерархическая структура классифицирующих правил типа «ЕСЛИ…, ТО…», имеющая вид дерева. Для принятия решения, к какому классу отнести некоторый объект или ситуацию, требуется ответить на вопросы, стоящие в узлах этого дерева, начиная с его корня. Определяют естественные «разбивки» в данных, основанные на целевых переменных. Сначала выполняется разбивка по наиболее важным переменным. Ветвь дерева можно представить как условную часть прави-

ла. Наиболее часто встречающимися примерами являются алгоритмы классификационных и регрессионных деревьев (Classification and regression trees, CART) либо □□-квадрат индукция (Chi-squared Automatic Induction, CHAID). Недостаток: деревья решений принципиально не способны находить «лучшие» (наиболее полные и точные) правила в данных. (IDIS, KnowledgeSEEKER, See5/C5.0).

Эволюционное программирование. Искомая зависимость целевой переменной от других переменных моделируется несколькими вариантами алгоритмов, из которых отбирается тот, который воспроизводит зависимость более точно. Программы, совершенствуясь, конкурируют друг с другом как живые организмы при естественном отборе в борьбе за выживаемость.

Примером такой системы является PolyAnalyst. Найденные зависимости представляются пользователю в виде математической формулы или таблицы. Иногда зависимость ищется в виде функции какого-то определенного вида, например в виде полинома. Так работает метод группового учета аргументов (МГУА).

Генетические алгоритмы. Исходно это было мощное средство решения разнообразных комбинаторных задач и задач оптимизации. Построение алгоритма начинается с кодировки логических закономерностей в базе данных (в виде так называемых, хромосом).

Популяция таких хромосом обрабатывается при последовательных итерациях с проведением отбора, операции изменчивости (мутации), скрещивания, генетической композиции, как это происходит в природе с настоящими генами. Для отбора определенных особей и отклонения других используется «функция приспособленности» (fitness function).

Генетические алгоритмы в первую очередь применяются для оптимизации топологии нейронных сетей и весов. Однако, их можно использовать и самостоятельно, для моделирования. Пример: GeneHunter.

Ассоциативные правила. Алгоритмы ограниченного перебора. Предложены М.М. Бонгардом для поиска логических закономерностей в данных. Выявляют причинно-следственные связи и определяют вероятности или коэффициенты достоверности, позволяя делать соответствующие выводы. Правила представлены в форме «если <условия>, то <вывод>». Их можно использовать для прогнозирования или оценки неизвестных параметров (значений). На основе частоты встречаемости логических закономерностей делается вывод о полезности какой-либо их комбинации (конъюнкции) для установления ассоциации в данных, для классификации, прогнозирования и т.д. (пример, WizWhy). Недостатки: максимальная длина комбинации в if-then-правиле равна 6; поиск простых логических событий в начале работы производится эвристически. Тем не менее данная система постоянно демонстрирует более высокие показатели при решении практических задач, чем все остальные алгоритмы.

Кластерный анализ. Подразделяет гетерогенные данные на гомогенные или полугомогенные группы. Метод позволяет классифицировать наблюдения по ряду общих признаков. Кластеризация расширяет возможности прогнозирования.

Системы для визуализации многомерных данных. Средства графического отображения данных поддерживаются всеми системами Data Mining. Но некоторые предназначены исключительно для этой цели (например, Data Miner 3D). Их главной характеристикой является дружелюбный пользовательский интерфейс с удобными средствами масштабирования и вращения изображений.

Конечно, для того, чтобы разобраться в достоинствах и недостатках приведенных здесь методов Data Mining, не достаточно столь краткого описания. Требуется гораздо больше информации и времени, чтобы сориентироваться в столь разнообразных и не всегда простых ме-

тодах. Необходимы консультации профессионалов в области Data Mining, рекомендующих наилучший подход в той или иной ситуации. Но затраченные усилия не пропадут даром, так как методы Data Mining значительно расширяют возможности специалистов любой области знаний для выявления наиболее информативных показателей при обработке обширных баз данных и решении конкретных задач; позволяют обнаруживать порой принципиально новые факты, радикально меняющие известные взгляды. Благодаря быстрому прогрессу вычислительной техники и появлению программ с дружественным интерфейсом они становятся все более доступными для пользователя. Нужно грамотно использовать разные методы Data Mining при решении разных задач. Вот какие 6 шагов к успеху в интеллектуальном анализе данных выделяют специалисты в этой области (В.Дюк, [3]): четкое представление цели; сбор релевантных данных; выбор методов анализа; выбор программных средств; выполнение анализа; принятие решения об использовании результатов.

ПРИМЕНЕНИЕ НЕКОТОРЫХ МЕТОДОВ Data Mining В МЕДИЦИНЕ

Далее мы расскажем о методах, которые можно смело отнести к области Data Mining: логико-статистических методах, основанных на оптимальных разбиениях и относящихся к методам теории распознавания образов (а также называемых методом статистически взвешенных синдромов – СВС). Эти методы позволяют провести достаточно полный анализ и достоверное сравнение групп больных по имеющемуся набору переменных. Они основаны на поиске в многомерном пространстве этих переменных логических закономерностей, или синдромов. Преимуществом данных методов перед традиционными методами регрессионного анализа или нейронными сетями являются гибкость, позволяющая описывать сложные зависимости, а также высокая наглядность представления результатов анализа и выявленных закономерностей. С помощью данных методов распознавания образов были успешно решены многие задачи в медико-биологических исследованиях

В Институте биохимической физики им. М.Н. Эмануэля в лаборатории математической биофизики с 1993 г. данные методы распознавания образов применялись для создания алгоритмов диагностики и прогнозирования в онкологии (прогноз выживаемости при остеогенной саркоме, при раке желудка), в неврологии (анализ иммунологических показателей при болезни Вильсона, диагностика видов инсульта), в педиатрии (прогноз обострения заболеваний верхних дыхательных путей у часто болеющих дошкольников), в психиатрии (прогноз депрессии при сотрясении головного мозга), в гинекологии (прогноз рецидива миомы матки после операции, прогноз осложнения после аборта по иммунологическим показателям) и других областях теоретической медицины [5–7].

Для медико-биологической информации характерны небольшие выборки, большое число параметров и наличие пропущенных значений в данных. Эти трудности, принципиальные для традиционных статистических методов, для нашего подхода не страшны. Мы любим работать именно с такими сложными данными. Причем данные могут быть и количественными, и качественными, непрерывными или дискретными. Главное, чтобы они имели вид таблицы, в которой один из столбцов является **группирующим** (целевым), то есть содержит номера групп, к которым относятся каждый из объектов (данная строка – есть информация об одном объекте, «запись»). Обучение идет на данных с известным разделением на группы. Имеется в виду, что сравниваемые группы заранее известны. Это могут быть группы больных с различным исходом лечения или заболевания, группы экспериментальных данных и контроля и т.д. После получения **решающего правила** можно любой

предлагаемый новый объект, группа которого не известна, с некоторой вероятностью отнести к одной из групп, то есть сделать для него прогноз или диагностику. Это и есть распознавание образов в действии.

Электронная цифровая запись (будет позже)

Распознавание образов – что это такое? В данном случае, образ – это все, что может называться информацией, то есть что-то, имеющее некоторые характеризующие его признаки: любые базы данных, состоящие из колонок цифр и строчек; любое оцифрованное изображение.

Вы видите на рис. 1, 2 (на диаграммах рассеяния) данные двух групп (се-

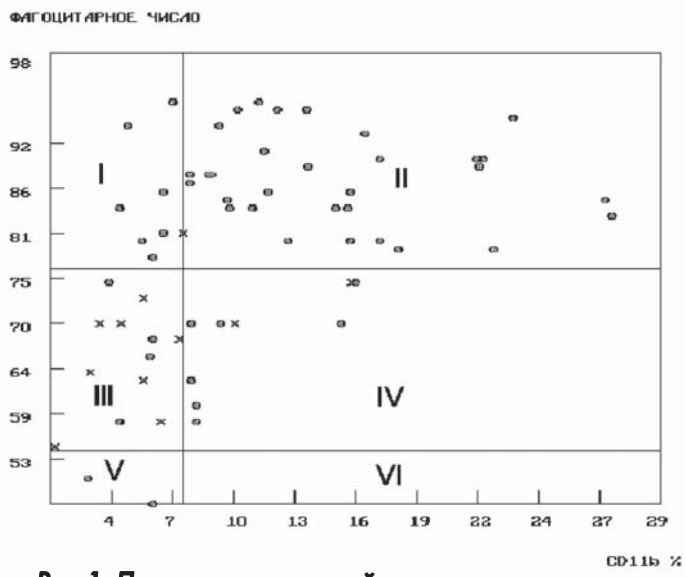

Рис.1. Прогноз осложнений после искусственного прерывания беременности.
По оси X – CD116⁺-лимфоциты (%), по оси Y – фагоцитарное число. Значения пациенток с осложнениями находятся преимущественно в квадранте III

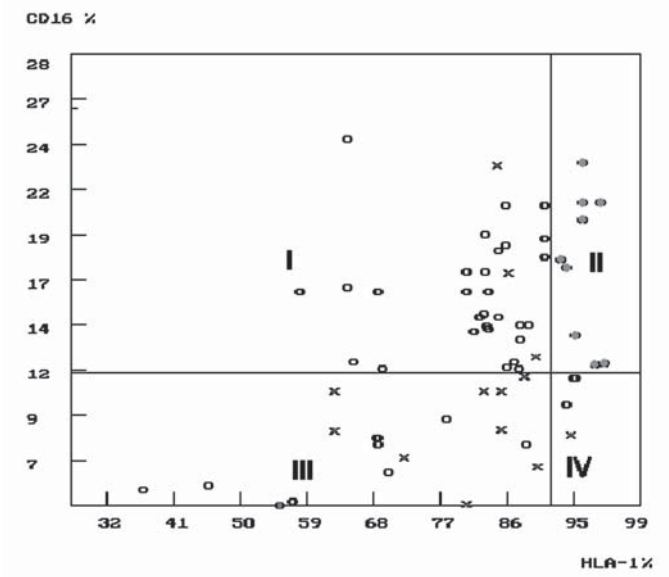

Рис.2. Прогноз осложнений после искусственного прерывания беременности.
По оси X – HLA-1⁺-лимфоциты (%), по оси Y – CD16⁺-лимфоциты (%). Значения пациенток без осложнений находятся преимущественно в квадранте I и II. Значения пациенток с осложнениями преобладают в квадранте III

рые и голубые значки) по двум признакам, отложенным по соответствующим осям. Решалась задача прогноза осложнений после искусственного прерывания беременности по иммунологическим показателям, полученным до операции. Было обследовано 2 группы женщин: 12 человек с осложнениями и 55 без них.

После проведения анализа методом статистически взвешенных синдромов (СВС) был выявлен набор признаков, достоверно характеризующих группу больных с осложнениями после аборта.

С помощью полученного решающего правила доктор в предоперационный период может выявить тех пациентов, у которых высок риск появления осложнений, и использовать в данном случае особую схему оперативного вмешательства с дополнительными превентивными мерами.

Статистические различия между группами по каждому из признаков могут отсутствовать, распределение может быть далеким от нормального. Как же найти то, чем одна группа отличается от другой?

Именно так, как указано на рисунке и происходит распознавание при работе **метода статистически взвешенных синдромов**:

1. Сначала ставятся границы градации (не обязательно одна) по одному из признаков таким образом, чтобы с одной стороны границы преобладали значения одной из групп, а с другой стороны было больше значений второй группы. То же делается для второго признака, и для всех признаков, участвующих в обучении (благодаря современной технике число признаков может быть почти не ограничено).

2. Далее из всех признаков оставляют только те, которые наиболее информативны с точки зрения отделения одной группы от другой.

3. Статистически взвешенное голосование проводится суммарно по всем базовым множествам. Так называют каждый прямоугольник, образованный границами градаций.

4. Создается **решающее правило**, которое включает в себя набор наиболее информативных признаков с их границами градаций. Это и есть найденный синдром, ограниченный набор симптомов, отобранный машиной. По нему новый объект, не участвующий в обучении можно распознать, то есть отнести с некоторой вероятностью к одной из групп. А значит сделать прогноз или диагностику данного объекта.

5. В результате распознавания мы имеем одно число, находящееся между номерами групп 1 и 2. Например: 1,79. К какому номеру группы оно ближе, к той группе и будет относиться распознаваемый объект. Существует зона неопределенности. Если результат попадает в нее, мы называем его отказом, то есть решение неопределенно.

6. Самое главное – доказать, что различия между группами, найденные в результате распознавания, достоверно значимы. Для этого существует перестановочный тест, использующий метод Монте-Карло. Номера группы каждому объекту присваиваются произвольно и на скользящем контроле опять проводят обучение и распознавание. Так делается в автоматическом режиме тысячу раз. Если хорошее распознавание получается в 5 случаях из этой тысячи, то считаем, что достоверность равна 0,005. Если таких случаев достаточно много, то скорее всего различия между исследуемыми группами нет.

Описанные методы были использованы для прогноза динамики депрессивных синдромов в остром периоде сотрясения головного мозга (СГМ) у группы военнослужащих по показателям первичного обследования. Первичное обследование включало клинико-психопатологическую оценку 29 анамнестических, 229 феноменологических признаков, а также исследование функциональной асимметрии, чувства времени, когнитивной деятельности, личностной и ситуационной тревожности у 50 военнослужащих, у которых в остром периоде СГМ имели место депрессивные расстройства. При этом у

24 больных была выражена положительная динамика и у 26 депрессивное состояние оставалось без изменений [8].

Для решения задачи прогноза динамики депрессивных синдромов в остром периоде СГМ была использована следующая процедура. На первом этапе метод оптимальных разбиений был использован для исследования прогностических возможностей отдельных признаков, а также для исследования совместной прогностической силы попарных сочетаний признаков. В результате был выявлен предварительный информативный набор из 29 признаков, который далее был проанализирован экспертом, и из него выделен набор из 14 наиболее интересных в клиническом плане признаков.

Рис. 3 иллюстрирует различия в распределениях значений пары прогностических переменных (показателя ситуационной тревоги и коэффициента правого уха) в группах пациентов с отрицательной и положительной динамикой, выявленных с помощью метода оптимальных разбиений с использованием модели II . Видно, что в область II попали только значения, соответствующие пациентам без изменения состояния, а в область III попали преимущественно значения, соответствующие пациентам с положительной динамикой.

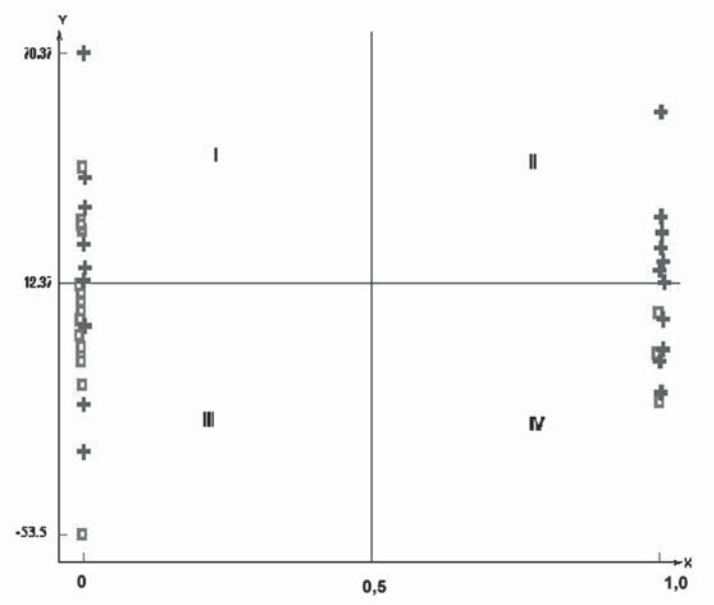

Рис.3. По оси X – ситуационная тревога, по оси Y – коэффициент правого уха.
Крестик (+) соответствует пациентам без видимых изменений их состояния и кружочек (о) соответствует пациентам с положительной динамикой. Во втором базовом множестве преобладают наблюдения группы без динамики – +, в третьем БМ – наблюдения группы с положительной динамикой – о.

Статистическая значимость выявленных различий между группами больных ($p<0,005$) рассчитана с помощью перестановочного теста.

Далее с помощью пошаговой процедуры в методе статистически взвешенных синдромов (СВС) из вышеуказанных признаков был выделен набор четырех наиболее информативных показателей. В него вошли (в порядке убывания информативности):

1. Результаты измерения коэффициента правого уха (Кпу) (признак, характеризующий доминантность одного из двух полушарий головного мозга).
2. Ситуационная тревога
3. Психотравмирующее событие в предшествующий травме период.
4. Травма затылочной области.

На 24 человека в группе с положительной динамикой 19 прогнозов было верными, 3 ошибочных прогноза и 2 отказа. Процент правильных прогнозов с учетом отказов – 87%, без учета отказов – 86%. Общая точность прогноза для двух групп составила 86% с учетом отказов и 85% без учета отказов. Полученный алгоритм краткосрочного прогноза динамики состояния депрессивных синдромов в остром периоде СГМ может быть использован при лечении.

ЗАКЛЮЧЕНИЕ

Различия в группах больных часто имеют сложный характер и их можно достоверно выявить и описать только учитывая взаимодействие между переменными, что не позволяют сделать одномерные статистические тесты.

Подходом, позволяющим провести достаточно полный анализ и достоверное сравнение групп больных по имеющемуся набору признаков, являются методы, основанные на поиске в **многомерном** пространстве признаков подобластей с существенным преобладанием объектов, принадлежащих какой-нибудь одной группе. Такие области (логические закономерности) по аналогии с медициной можно назвать синдромами. С помощью логико-статистических методов распознавания образов, принадлежащих к большому классу алгоритмов Data Mining, можно с успехом решать многие задачи медико-биологических исследований. Более подробно можно ознакомиться с методами на сайте: http://azfor.narod.ru/datmin/datmin.htm.

ЛИТЕРАТУРА

1. *Дюк В., Самойленко А.* «Data Mining: учебный курс». – СПб.: «Питер», 2001.
2. «Что такое Data Mining», сайт: Intersoft Lab http://www.iso.ru/
3. *Дюк В.* Обработка данных на ПК в примерах. – СПб.: «Питер», 1997. – 240 с.
4. *Реброва О.Ю.* Статистический анализ медицинских данных. Применение пакета прикладных программ STATISTICA. – М.: «Медиа Сфера», 2002. – 305 с.
5. *Кузнецова А.В.* Диагностика и прогнозирование опухолевого роста по иммунологическим данным с помощью методов синдромного распознавания//Автореф. дис. канд. биол. наук. – М., 1995. – 23 с.
6. *Кузнецов В.А., Сенько О.В., Кузнецова А.В. и др.* Распознавание нечетких систем по методу статистически взвешенных синдромов и его применение для иммуногематологической нормы и хронической патологии//Химическая физика. – 1996. – Т.15. – №1. – С. 81–100.
6. *Zhirnova I.G., Kuznetsova A.B., Rebrova O.Yu., Labunsky D.A., Komelkova L.V., Poleshchuk V.V., Sen'ko O.V.* Logical and Statistical Approach for the Analysis of Immunological Parameters in Patients with Wilson's Disease//Russian Journal of Immunology. – 1998. – Vol.3. – № 2. – C.174–184.
7. *Kuznetsova A.V., Sen'ko O.V., Matchak G.N., Vakhotsky V.V., Zabotina T.N., Korotkova O.V.* The Prognosis of Survivance in Solid Tumor Patients Based on Optimal Partitions of Immunological Parameters Ranges//Journal Theoretical Medicine. – 2000. – Vol. 2. – C.317–327.
8. *Доровских И.В., Кузнецова А.В., Сенько О.В., Реброва О.Ю.* Прогноз динамики депрессивных синдромов в остром периоде сотрясения головного мозга по показателям первичного обследования (с использованием логико-статистических методов)//Социальная и клиническая психиатрия. – 2003. – №4. – С.18–24.