УДК 004.65:519.23 **DOI: 10.34670/AR.2022.69.56.081**

Изучение возможностей методов Data Mining для проведения анализа медицинских данных

Зелинская Снежана Александровна

Доктор педагогических наук,
доцент кафедры медицинской химии,
Луганский государственный медицинский университет им. Святителя Луки,
91045, Российская Федерация, Луганск,
квартал 50-летия Обороны Луганска, 1Г;
e-mail: snejana.zelinskaya@mail.ru

Зелинский Сергей Сергеевич

Кандидат педагогических наук,
доцент кафедры социальной медицины и экономики здравоохранения,
Луганский государственный медицинский университет им. Святителя Луки,
91045, Российская Федерация, Луганск,
квартал 50-летия Обороны Луганска, 1Г,
e-mail: snejana.zelinskaya@mail.ru

Аннотация

Сейчас обработка колоссальных объемов разрозненных медицинских данных немыслима без использования средств современных информационно-коммуникационных технологий. В то же время не все технологии одинаково эффективны и могут обеспечить информационные потребности специалиста. Интеллектуальный анализ Data Mining является уникальным представителем комплекса современных методов для проведения статистического анализа данных, используя которые можно получить наилучший результат. Использование алгоритма Apriori позволяет избавиться от полного перебора всех возможных наборов в процессе построения ассоциативных правил в процессе статистического анализа данных. Предложенный алгоритм и программный инструментарий могут эффективно применяться при решении различных задач диагностики как в медицине, так и в других прикладных областях.

Для цитирования в научных исследованиях

Зелинская С.А., Зелинский С.С. Изучение возможностей методов Data Mining для проведения анализа медицинских данных // Педагогический журнал. 2022. Т. 12. № 5А. С. 621-634. DOI: 10.34670/AR.2022.69.56.081

Ключевые слова

Интеллектуальная система, метод, медицинские данные, Data Mining, Apriori, множество кандидатов.

Введение

Актуальность данной статьи обусловлена тем, что стремительные темпы развития информационно-коммуникационных технологий, в частности происходящие прогрессы в методах сбора, хранения и обработки данных, позволяют многим организациям собирать огромные массивы данных, которые необходимо эффективно анализировать. Объемы этих данных настолько велики, что возможностей экспертов уже не хватает и необходимо использовать специализированные средства интеллектуального анализа данных.

На сегодняшний день интенсивно развивается направление, которое связано с интеллектуализацией методов обработки и анализа данных. Интеллектуальные системы анализа данных, основанные на методах Data Mining, призваны минимизировать усилия лиц, принимающих решения, в процессе анализа данных, а также оптимизировать усилия в тонкой настройке алгоритмов анализа. Большинство интеллектуальных систем анализа данных позволяют не только быстро решать классические задачи принятия решения, но и способны выявлять причинно-следственные связи, скрытые закономерности в исследуемой системе.

В связи с тем, что интеллектуальный анализ данных на базе использования методов Data Mining предоставляет огромное количество инструментальных средств, каждый из исследователей выбирает наиболее подходящий и использует его в своей сфере профессиональной деятельности. Так, например, в работах А.А. Берсегян описаны высокопродуктивные технологии анализа данных: Data Mining, Visual Mining, Text Mining, OLAP. Исследователи О.В. Мурахина, А.А. Непотребная занимались вопросами практического применения методов Data Mining для анализа медицинских данных. В работе А.С. Сеидовой, Е.В. Берестневой, И.А. Осадчей представлены результаты использования методов Data Mining для создания медицинских баз знаний. Изучением вопросов практического использования BI-технологий для многомерной обработки данных занималась Б.М. Саданова. Уделили большое внимание вопросу заполнения базы знаний логическими правилами в виде продукций следующая группа авторов: О.Г. Берестнева, К.А. Шаропин, А.В. Старикова. Н.В. Пивоварова, С.И. Видунова рассматривали алгоритм Apriori для получения ассоциативных правил из множества накопленных данных аптек фармацевтического предприятия.

Однако следует отметить, что в рамках изучаемой темы существует множество разрозненных данных, каждый из исследователей акцентирует свое внимание на интересующих именно его характеристиках анализируемой задачи, что не позволяет получить целостное представление о практическом использовании технологий интеллектуального анализа Data Mining в рамках предметной сферы. В связи с этим необходимы дополнительные исследования методов Data Mining в сфере проведения статистического анализа медицинских данных.

Цель исследования заключается в изучении возможностей методов Data Mining для проведения анализа медицинских данных.

В соответствии с целью была определена необходимость постановки и решения следующих задач: изучить технологии Data Mining и сферу их практического использования; проанализировать современное состояние изученности методов Data Mining; описать практическое применение алгоритмов интеллектуального анализа Data Mining, в частности алгоритм Apriori.

Основная часть

В биологических и медицинских исследованиях, равно как и в практической медицине, спектр решаемых задач настолько широк, что позволяет использовать любые методы Data Mining. Примером может служить построение диагностических систем или исследование эффективности хирургических вмешательств.

Известно множество экспертных систем для постановки медицинских диагнозов. Главным образом, они построены на базе использования правил, которые описывают сочетания различных симптомов отдельных заболеваний. При помощи использования таких правил можно узнать не только, чем болен пациент, но и какое необходимо назначить ему лечение. Сформулированные правила интеллектуальных систем помогают выбирать необходимые средства медикаментозного воздействия, определять показания или противопоказания, ориентироваться в назначаемых лечебных процедурах, создавать условия наиболее эффективного лечения, предсказывать последующий исход назначенного курса лечения и т.п. Технологии Data Mining позволяют в медицинских данных обнаруживать шаблоны, составляющие основу указанных правил.

Одним из наиболее передовых направлений современной медицины является биоинформатика, которая представляет собой область науки, разрабатывающую и применяющую разнообразные вычислительные алгоритмы для оперативного анализа и последующей систематизации генетической информации с целью выявления структуры и функции макромолекул, последующего использования этих знаний для объяснения различных биологических явлений и создания новых лекарственных препаратов.

Data Mining представляет собой сочетание многообразного математического инструментария (начиная от классического статистического анализа до передовых кибернетических методов обработки информации) и самых последних достижений в сфере информационно-коммуникационных технологий. В Data Mining гармонично объединены строго формализованные методы и методы неформального анализа, т.е. реализуется качественный и количественный анализ данных.

Data Mining (добыча данных, интеллектуальный анализ данных, глубинный анализ данных) является собирательным названием, которое используется для обозначения совокупности разных методов обнаружения в данных ранее не известных, нетривиальных, практически полезных и доступных интерпретации знаний, которые необходимы для оперативного принятия эффективных решений в различных сферах человеческой жизнедеятельности [Берсегян и др., 2009].

Одним из наиболее значимых назначений методов Data Mining является наглядное представление полученных результатов в процессе выполненных вычислений, что позволяет использовать предоставляемый инструментарий Data Mining людьми, у которых практически нет специализированной математической подготовки для их практического применения в профессиональной деятельности. В то же время практическое использование статистических методов анализа данных требует достаточно высокого владения математической статистикой и теорией вероятностей.

Основой Data Mining является широкий спектр методов прогнозирования, моделирования, классификации, кластеризации, ассоциации, которые основаны на использовании деревьев решений, искусственных нейронных сетей, генетических алгоритмов, элементов эволюционного программирования, ассоциативной памяти, нечёткой логики. Однако

необходимо отметить то, что использование такого рода методов предполагает наличие некоторых априорных представлений об обрабатываемых данных, что может существенно расходиться с целями технологии Data Mining (обнаружение ранее неизвестных, нетривиальных и практически полезных знаний).

Статистические методы Data Mining представлены четырьмя взаимосвязанными разделами:

– предварительный анализ статистических данных (проверка гипотез нормальности, стационарности, однородности, независимости, оценка вида функции распределения, ее параметров и т.п.);
– выявление закономерностей и связей (корреляционный анализ, линейный и нелинейный регрессионный анализ и др.);
– многомерный статистический анализ (кластерный анализ, линейный и нелинейный дискриминантный анализ, компонентный анализ, факторный анализ и др.);
– прогноз и динамические модели на базе использования временных рядов [Берсегян, Куприянов, Степаненко, 2007].

Посредством использования перечисленных методов Data Mining в медицине можно:

– изучать состояние общественного здоровья населения в целом и его основных групп посредством сбора и последующего анализа полученных статистических данных о численности и составе населения, его физическом развитии, воспроизводстве, распространенности и длительности различных заболеваний и т.д.;
– выявлять и устанавливать связи общего уровня заболеваемости и смертности от каких-либо отдельных болезней с учетом различных факторов окружающей среды;
– собрать и изучить числовые данных о сети медицинских учреждений, их практической деятельности и кадрах для последующего планирования медико-санитарных мероприятий, контролировать выполнение установленных планов развития вычислительной сети и непосредственной деятельности учреждений здравоохранения и оценки качества работы отдельных медицинских учреждений;
– оценивать эффективность проводимых мероприятий по предупреждению и последующему лечению заболеваний;
– определять статистическую значимость полученных результатов исследований в клинике и экспериментах.

Для более целостного понимания проблематики использования статистических методов Data Mining были изучены актуальные работы исследователей в этой области.

В работе А.С. Сеидовой, Е.В. Берестневой, И.А. Осадчей описано применение высокопроизводительных методов Data Mining для создания медицинских баз знаний. Полученные результаты могут быть использованы для расширения уже существующих баз знаний системы поддержки научных исследований бронхиальной астмы. Также при непосредственном создании прототипа виртуального центра, предназначенного для оценки и последующего мониторинга состояния детей с наиболее распространенными неинфекционными заболеваниями [Сеидова, Берестнева, Осадчая, 2016].

В продолжение затронутой темы предыдущими авторами Б.М. Саданова описала возможности использования BI-технологий для многомерной обработки данных. Центральным инструментальным средством непосредственного создания хранилищ и витрин данных является интегрированная CASE-среда Oracle Warehouse Builder, которая построена на базе использования современной архитектуры Common Warehouse Metadata. Она предназначена для

комплексного описания структуры хранилищ и витрин данных, проектирования и создания процедур извлечения, согласования и последующей загрузки данных, а также генерации метаданных для создания специализированных средств доступа к ним, например таких, как Discoverer, Oracle Workflow and Oracle Enterprise Manager [Саданова, 2014].

Отметим, что Warehouse Builder представляет собой средство генерации программного кода, основанное на использовании репозитория метаданных. Затем полученный программный код может быть использован для создания хранилищ данных и для последующего сопровождения, преобразования данных. Средство Warehouse Builder сильно интегрировано с другими прикладными программными продуктами фирмы Oracle, такими как Oracle Enterprise Manager, которые предназначены для эффективного выполнения процедур загрузки анализируемых данных по расписанию и отслеживания состояния выполнения задач в базе данных [Дорожкин, Климанов, 2004]. А также с Oracle Workflow можно выполнять задания нелинейных последовательностей, выполнять разные этапы процесса загрузки данных и планировать оповещения в случаях возникновения программных ошибок.

Среди преимуществ практического использования средства генерации программного кода Warehouse Builder для создания, последующей поддержки хранилищ данных и разработки аналитических прикладных программных решений можно выделить:

1) Уменьшение времени стадий разработки. За счет наличия обширного набора различных мастеров, удобного графического пользовательского интерфейса и наличия большой библиотеки, заранее разработанных преобразований существенно уменьшается время создания объектов хранилища данных и программных процедур его последующего заполнения.
2) Централизованное проектирование. Вся информация о вычислительной системе хранится в центральном репозитории, что позволяет избежать несогласованности обрабатываемых данных, а также позволяет существенно упростить вопросы эффективного управления этим данными.
3) Защита от ошибок. За счет наличия мощного механизма проверки данных, а также централизованного репозитория сводится к минимуму не только вероятность возникновения программных ошибок, но и становится более легким повторное создание, обновление и последующая поддержка хранилища данных.
4) Повышение эффективности технологий. Warehouse Builder предназначен для эффективного использования новых функций, которые появились в последних версиях базы данных Oracle.

Другая рассматриваемая Б.М. Садановой полезная OLAP-технология является BI-решением от корпорации Microsoft, которое построено на базе использования специализированных средств платформы SQL Server и включает специализированные компоненты анализа данных Analysis Services и Integration Services. Microsoft SQL Server Analysis Services (SSAS) построены на базе использования унифицированной многомерной модели (Unified Dimensional Model, UDM), позволяющей различным типам клиентских прикладных программных приложений получать доступ к обрабатываемым данным как из реляционных, так и из многомерных баз данных без использования отдельных моделей, для каждого типа баз данных. UDM предоставляет широкие возможности для практического использования множества источников данных (data sources) для создания многомерной модели обработки данных.

Наличие возможностей по работе с многомерным массивом данных позволяет получить достаточный объем информации для нахождения максимально точного результата, на

основании которого могут быть приняты соответствующие решения.

Также можно отметить одну из ключевых проблем современной медицины – диагностирование бронхиальной астмы. Бронхиальная астма является причиной значительных ограничений жизнедеятельности, снижения социальной активности больных, т.е. снижения их качества жизни. Ограничение физической и социальной активности отрицательно сказывается на развитии человека, вызывает значительные трудности у больного. На развитие болезни влияют не только такие факторы, как наследственность, профессиональные факторы, экологические факторы, нервная и иммунная системы, но и, возможно, ряд других факторов. Для выявления скрытых закономерностей у больных бронхиальной астмой группой авторов (О.В. Мурахина, А.А. Непотребная) были использованы преимущественно продукционные модели [Марухина, Непотребная, 2014].

Продукционная модель знания представляет собой модель, которая основана на правилах, позволяет представить знания в виде предложений следующего типа – «Если (условие), то (действие)». Данный метод реализован в пакете See 5/C 5.0, основная задача которого состоит в предсказании диагностического класса некоторого объекта по значениям его уникальных признаков. При этом пакет See 5/C 5.0 позволяет сконструировать классификатор в виде дерева решений, которому может быть поставлено в соответствие определенное множество логических правил. В большинстве случаев полученное дерево решений может оказаться слишком сложным для восприятия, но в то же время позволит эффективно обработать медицинскую информацию.

Например, при построении задач высокой размерности для неоднородных данных часто дерево получается кустистым и довольно большим. Поэтому для упрощения логического вывода можно воспользоваться логической связкой «И». Если по смыслу существует логическая связка «ИЛИ», то формируется второе аналогичное правило, которое содержит только связки «И».

Необходимо обратить внимание на то, что продукционная модель может являться фрагментом семантической сети, которая основана на временных отношениях между состояниями объектов. В сравнении с другими формами представления знаний продукции имеют такие преимущества:

– модульность;
– единообразие структуры (основные компоненты продукционной модели могут использоваться для построения интеллектуальных систем с различной проблемной ориентацией);
– естественность (в продукционной модели вывод заключения во многом аналогичен процессу рассуждений эксперта);
– гибкость родовидовой иерархии понятий, которая поддерживается только как связь между правилами (изменение правила влечет за собой изменение в иерархии) [Берестнева, Шаропин, 2010].

В то же время нужно сказать и о недостатках практического использования продукционной модели: при накоплении достаточно большого числа продукций они начинают, вследствие необратимости дизъюнкций, противоречить друг другу. В этом случае разработчикам потребуется усложнять систему, включая в нее модули нечеткого вывода или иные специализированные средства разрешения возникающих конфликтов, – правила по глубине, правила по приоритету, эвристические механизмы исключений, возврата и т.п.

Продолжая затронутую тему, нельзя не упомянуть о получении новых знаний на базе

использования ассоциативных правил.

Так, в статье Н.В. Пивоваровой, С.И. Видуновой рассмотрена методика практического использования данных для повышения их непосредственной эффективности и действенности. Показано, как можно эффективно использовать алгоритм Apriori для нахождения ассоциативных правил в данных, которые были получены из аптек фармацевтической компании (в процессе чего может быть выявлена корреляция между санитарно-гигиеническими изделиями и обезболивающими средствами) [Пивоварова, Видунова, 2016]. Авторами подробно рассмотрена часть набора всех медикаментов, которые реализуются в аптеках фармацевтической компании, и часть ее аптечных транзакций. Частота появления группы предметов или отдельного предмета, которая выражается в процентах, называется распространенностью. Низкий уровень распространенности (менее одной тысячной процента) говорит о том, что такая ассоциация является несущественной.

В статье В.А. Биллига рассматривается алгоритм построения ассоциативных правил AprioriScale. Алгоритм используется для непосредственного решения конкретных задач диагностики в медицине. Автором построена программная реализация алгоритма AprioriScale при помощи использования средств языка программирования C#. Создан специализированный прикладной программный инструмент, который позволяет медикам проводить важные исследования в процессе решения задач диагностики различных заболеваний.

Алгоритм AprioriScale является модификацией классического алгоритма Apriori, который позволяет извлекать ассоциативные правила из некоторой базы данных. Основной особенностью алгоритма AprioriScale являются способы представления данных и последующего построения достоверных ассоциативных правил [Биллиг, Иванова, Царегородцев, 2016].

Алгоритм AprioriScale и разработанный упомянутым автором инструментарий могут быть использованы как для решения широкого спектра задач медицинской диагностики, так и для решения задач других прикладных областей.

Существует достаточно большое количество реализаций алгоритма Apriori. Описание некоторых из них можно найти в работах [Olson, Delen, 2008; Vercellis, 2009]. Большая часть из них ориентированы на оперативный анализ потребительских корзин.

Алгоритм Apriori является одним из стандартных алгоритмов нахождения правил ассоциаций среди набора данных [Ahmad et al., 2015]. Ассоциация позволяет выделить устойчивые группы объектов, между которыми существуют неявно заданные связи.

Цель алгоритма Apriori заключается в поиске всех ассоциативных правил, у которых частота и качество будут выше заданных пользователем минимальных значений.

$$Rules = \{Rule \mid Support(Rule) > \\ > \text{support\_min} \,\&\, \text{Confidence(Rule)} > \\ > \text{quality\_min}\}$$

Шаги следования вычислений алгоритма Apriori заключаются в реализации следующих задач:

1) обнаружение обычных наборов обрабатываемой записи базы данных;

2) построение ассоциациативных правил на основе использования найденных наборов.

Основная идея алгоритма Apriori, предложенная в работе [Agrawal, Srikant, 1994], заключается в том, что использование алгоритма позволяет избавиться от полного перебора всех возможных наборов в процессе построения ассоциативных правил.

Детализация алгоритма Apriori может быть представлена следующим образом. Назовем

набор свойств X частым, если $\text{Support(X)}>>\text{support\_min}$. В таком случае будет справедливо такое утверждение: если X будет частым набором, то и все его полученные подмножества будут частыми наборами.

Более важным будет утверждение, которое следует из отрицания такой импликации: если X не будет частым набором, то и все его надмножества не будут являться частыми наборами. Это свойство наборов, которое называется антимонотонностью, при построении правил позволяет выполнить исключение из рассмотрения большое число наборов.

Так, например, при обнаружении, что определенное свойство p редко появляется в наборах обрабатываемой базы данных, можно не рассматривать все наборы базы данных, содержащие свойство p.

Отсюда следует и общая схема практической реализации алгоритма Apriori. Вначале строится множество частых наборов F, которое содержит наборы длины 1, и множество правил R, где посылка и заключение содержат наборы из F. Затем в цикле на базе использования уже построенных множеств частых наборов F и R длины k строятся множества F и R длины $k+1$. Цикл завершится, когда вновь создаваемое множество F длины $k+1$ будет пустым, то есть не будет существовать частых наборов длины $k+1$.

Таким образом, использование алгоритма будет оправданным при обнаружении обычных наборов записи и построении ассоциативных правил на основе использовании найденных наборов.

В качестве примера рассмотрим ситуацию, когда диагностируется заболевание, которое похоже на астму («подражатели астмы»). Если у пациента есть характерные симптомы, которые проявляются через кашель, свистящее, затрудненное дыхание, это еще не может означать, что у пациента астма. Множество других заболеваний могут иметь похожие симптомы. Так как одни и те же симптомы могут относиться и к астме, и к широкому спектру других заболеваний, то лечащему врачу необходимо провести тщательное обследование, чтобы убедиться, что данные симптомы относятся именно к астме. Вариантом решения такого рода задач является применение методов Data Mining для проведения статистического анализа медицинских данных, например, алгоритм Apriori.

Одной из наиболее распространенных задач анализа данных является определение часто встречающихся наборов объектов в большом множестве наборов. В общем виде эту задачу можно описать следующим образом. Для этого необходимо обозначить объекты, которые составляют изучаемые наборы, следующим множеством:

$$I=\{i_1,i_2,...,i_j,...,i_n\}.$$

В представленном множестве i_j представляет собой объекты, которые входят в изучаемые наборы; n представляет собой общее число объектов. Базовое множество кандидатов представлено в таблице 1.

Из представленной таблицы можно получить следующее множество объектов:

$I=\{$воспаление носовых пазух, нарушение кровообращения в мышечных тканях сердца, сгустки крови в легочных артериях, боль в груди, повреждение стенок дыхательных путей, голосовые связки перестают выполнять свои функции, грибковая инфекция легочных тканей$\}$

Таблица 1 - Базовое множество кандидатов

Идентификатор	Наименование симптома	Заболевание
0	Воспаление носовых пазух	Синусит
1	Нарушение кровообращения в мышечных тканях сердца	Ишемия миокарда
2	Сгустки крови в легочных артериях	Эмболия легких
3	Боль в груди	Стенокардия
4	Повреждение стенок дыхательных путей и легких	Бронхоэктаз
5	Голосовые связки перестают выполнять свои функции	Паралич голосовых связок
6	Грибковая инфекция легочных тканей	Аспергиллез легких

Алгоритм Apriori представляет собой алгоритм поиска ассоциативных правил и используется для выявления частных наборов объектов, реализуется через большое количество вычислений и машинного времени на их непосредственную обработку.

В своей работе алгоритм Apriori использует свойство, которое заключается в поддержке любого набора объектов, который не может превышать минимальной поддержки из его подмножеств, что может быть описано следующим образом:

$Supp_F \le Supp_E$ при $E \subset F$.

В качестве примера можно привести то, что поддержка трехобъектного набора (боль в груди, воспаление носовых пазух, повреждение стенок дыхательных путей и легких) будет всегда меньше или равна поддержке двухобъектных наборов (боль в груди, воспаление носовых пазух), (воспаление носовых пазух, повреждение стенок дыхательных путей и легких(.

Это можно объяснить тем, что любая выполненная транзакция, которая будет содержать следующие наборы: боль в груди, воспаление носовых пазух, повреждение стенок дыхательных путей и легких, также содержит и наборы (боль в груди, воспаление носовых пазух), (воспаление носовых пазух, повреждение стенок дыхательных путей и легких), (боль в груди, повреждение стенок дыхательных путей и легких), причем обратное отношение будет неверным.

За несколько этапов при помощи использования алгоритма Apriori можно определить часто встречающиеся наборы. На i-м этапе выполняется определение всех часто встречающихся i-элементных наборов. Каждый отдельный этап включает несколько шагов, к которым относятся формирование кандидата; подсчет поддержки кандидата.

Рассмотрим более подробно i-й этап формирования кандидатов, в процессе которого будет создано множество кандидатов из i-элементных наборов. Подсчет кандидата предполагает сканирование множества проведенных транзакций в процессе вычисления поддержки набора кандидатов. После выполнения операций сканирования происходит отбрасывание кандидатов, поддержка которых будет ниже установленного минимального значения, и будут сохранены только частные i-элементные наборы.

Представим алгоритм Apriori обработки кандидатов в виде следующего псевдокода:

Apriori(*T,e*)

L_1={1-элементные наборы, которые часто встречаются}

while k=2;

L_{k-1} <> f; k++

for transactions $t \in$ D

C_t = subset(C_k, t) // удаление правил, которые избыточны

for candidates c $\in$ C_t

c.count++

Lk = { c $\in$ C_k | c.count >= $Supp_{\min}$ } // выполнение отбора кандидатов

return $\bigcup_k L_k$

Опишем основные обозначения, используемые в представленном алгоритме.

L_k – представляет собой определенное множество k -элементных частных наборов, поддержка которых будет не ниже установленного минимального значения. Все члены множества имеют наборы упорядоченных ($i_j < i_p$, если $j < p$) элементов F и следующее значение поддержки такого набора $Supp_F > Supp_{\min}$:

$$L_k = \{(F_1, Supp_1), (F_2, Supp_2), ..., (F_q, Supp_q)\},$$

где $F_j = \{i_1, i_2, ..., i_k\}$;

C_k – представляет собой множество *k*-элементных наборов потенциально частых кандидатов. Каждый отдельный член множества имеет определенный набор элементов F, которые упорядочены ($i_j < i_p$, если $i < p$), и специальное значение поддержки набора $Supp$.

Рассмотрим работу алгоритма Apriori на основании данных, приведенных в таблице 1, с учетом того, что $Supp_{\min} = 0{,}5$. На первом шаге получим следующее множество кандидатов C_1 (указаны соответствующие идентификаторы симптомов) (табл. 2).

Таблица 2 - Множество кандидатов C_1

№	Набор	Значение ($Supp$)
1	{0}	0
2	{1}	0,5
3	{2}	0,75
4	{3}	0,25
5	{4}	0,75
6	{5}	0,75
7	{6}	0,5

Заданной минимальной поддержке удовлетворяют только кандидаты под следующими номерами: 2, 3, 5, 6, 7.

$L_1 = \{\{1\}, \{2\}, \{4\}, \{5\}, \{6\}\}$.

На следующем шаге можно увеличить до двух значение k. В связи с тем, что можно получить двухэлементные наборы, будет получено множество C_2, которое представлено в таблице 3.

Таблица 3 - Множество кандидатов C_2

№	Набор	Значение ($Supp$)

1	{1,2}	0,25
2	{1,4}	0,5
3	{1,5}	0,25
4	{1,6}	0,5
5	{2,3}	0,75
6	{2,5}	0,5
7	{3,6}	0,5

На базе построенных кандидатов в рамках заданной минимальной поддержки удовлетворяют кандидаты под следующими номерами: 2, 4, 5, 6, 7.

$L_2 = \{\{1,4\},\{1,6\},\{2,3\},\{2,5\},\{3,6\}\}$.

На последнем шаге следует перейти к созданию трехэлементных кандидатов и выполнению подсчета их поддержки. В результате будет получено такое множество C_3, которое представлено в таблице 4.

Таблица 4 - Множество кандидатов C_3

№	Набор	Значение ($Supp$)
1	{2,4,5}	0,5

Представленный набор удовлетворяет самой низкой поддержке, следовательно:

$L_3 = \{\{2,4,5\}\}$.

В связи с тем, что четырехэлементные наборы получить нет возможности, тогда в результате работы алгоритма будет получено следующее множество:

$L = L_1 \cup L_2 \cup L_3 = \{\{1\},\{2\},\{4\},\{5\},\{6\},\{1,4\},\{1,6\},\{2,3\},\{2,5\},\{3,6\},\{2,4,5\}\}$.

Для подсчета поддержки кандидатов нужно сравнить все транзакции с каждым отдельным кандидатом. Очевидным является то, что количество кандидатов может быть достаточно большим и нужен высокоэффективный способ выполнения такого рода интеллектуального подсчета. Гораздо эффективнее и быстрее будет использовать подход, который основан на хранении кандидатов в хэш-дереве. Внутренние узлы дерева будут содержать хэш-таблицы с информационными указателями на потомков, а листья – на кандидатов. Этим деревом можно воспользоваться для быстрого подсчета поддержки кандидатов.

Заключение

Методика анализа данных с использованием механизмов Data Mining базируется на использовании различных алгоритмов извлечения закономерностей из исходных данных, результатом работы которых являются специальные модели. Таких алгоритмов довольно много, но, несмотря на их обилие, использование машинного обучения и т.п., они не способны в полной мере гарантировать качественное решение поставленной задачи. Никакой самый сложный метод сам по себе не даст хороший результат, так как критически важным становится вопрос качества исходных данных. Чаще всего именно качество данных является причиной неудачи.

Для того чтобы найти новое знание на основе данных большого хранилища, недостаточно просто использовать алгоритмы Data Mining, запускать их и ждать появления интересных результатов. Нахождение нового знания – это процесс, который включает в себя несколько

шагов, каждый из которых необходим для уверенности в эффективном применении средств Data Mining.

Также был использован алгоритм Apriori для решения конкретной задачи медицинской диагностики. Построенные ассоциативные правила позволили сделать важный вывод о возможности выявления на ранних стадиях параметров, свидетельствующих о риске возникновения «подражателей астмы». Предложенный алгоритм и программный инструментарий могут эффективно применяться при решении различных задач диагностики как в медицине, так и в других прикладных областях.

Библиография

1. Берестнева О.Г., Шаропин К.А., Старикова А.В., Кабанова Л.И. Технология формирования баз знаний в медицинских информационных системах // Известия Южного федерального университета. Технические науки. 2010. № 8. С. 32-37.
2. Берсегян А.А. и др. Анализ данных и процессов. 3-е изд., перераб. и доп. СПб.: БХВ-Петербург, 2009. 512 с.
3. Берсегян А.А., Куприянов М.С., Степаненко В.В. Технологии анализа данных: Data Mining, Visual Mining, Text Mining, OLAP. 2-е изд., перераб. и доп. СПб.: БХВ-Петербург, 2007. 384 с.
4. Биллиг В.А., Иванова О.В., Царегородцев Н.А. Построение ассоциативных правил в задаче медицинской диагностики // Программные продукты и системы. 2016. № 2 (114). С. 146-157.
5. Дорожкин А.К., Климанов В.А. Использование Oracle Warehouse Builder для организации хранилищ данных // Научно-технический вестник информационных технологий, механики и оптики. 2004. № 14. С. 105-110.
6. Марухина О.В., Непотребная А.А. Применение методов Data Mining для анализа медицинских данных // Труды Всероссийской молодежной научно-практической конференции «Перспективы развития информационных технологий». Кемерово, 2014. С. 247-248.
7. Пивоварова Н.В., Видунова С.И. Интеллектуальный анализ данных в фармацевтическом бизнесе // Интернет-журнал «Науковедение». 2016. № 6. (37). С. 166.
8. Саданова Б.М. Использование BI-технологий для многомерной обработки данных // Труды Всероссийской молодежной научно-практической конференции «Перспективы развития информационных технологий». Кемерово, 2014. С. 127-128.
9. Сеидова А.С., Берестнева Е.В., Осадчая И.А. Применение методов Data Mining для создания медицинских баз знаний // Труды Международной конференции студентов, аспирантов и молодых ученых «Перспективы развития фундаментальных наук». Томск, 2016. С. 123-125.
10. Agrawal R., Srikant R. Fast algorithms for mining asso-ciation rules in large databases // Proc. 20th Intern. Conf. on Very Large Data Bases, VLDB. 1994. P. 487-499.
11. Ahmad Y. et al. Identifying Association Rules among Drugs in Prescription of a Single Drugstore Using Apriori Method // Intelligent Information Management. 2015. P. 253-259.
12. Olson D.L., Delen D. Advanced Data MiningTechniques. Springer, 2008. 181 p.
13. Vercellis C. Business Intelligence: Data Mining and Opti-mization for Decision Making. Willey, 2009. 436 p.

Exploring the opportunities of Data Mining methods for analyzing medical data

Snezhana A. Zelinskaya

Doctor of Pedagogy,
Associate Professor of the Department of medical chemistry,
Lugansk State Medical University named after Saint Luke,
91045, 1G kvartal 50-letiya Oborony Luganska, Lugansk, Russian Federation;
e-mail: snejana.zelinskaya@mail.ru

Sergei S. Zelinskii

PhD inPedagogy,
Associate Professor of the Department of social medicine and health economics,
Lugansk State Medical University named after Saint Luke,
91045, 1G kvartal 50-letiya Oborony Luganska, Lugansk, Russian Federation;
e-mail: snejana.zelinskaya@mail.ru

Abstract

Today the processing of colossal volumes of disparate medical data is unthinkable without the use of modern information and communication technologies. At the same time, not all technologies are equally effective and can meet the information needs of a specialist. Data Mining is a unique representative of a set of modern methods for statistical data analysis, using which one can get the best result. The use of the Apriori algorithm allows us to get rid of a complete enumeration of all possible sets in the process of constructing association rules in the process of statistical data analysis. The proposed algorithm and software tools can be effectively used in solving various diagnostic problems both in medicine and in other applied areas.

For citation

Zelinskaya S.A., Zelinskii S.S. (2022) Izuchenie vozmozhnostei metodov Data Mining dlya provedeniya analiza meditsinskikh dannykh [Exploring the opportunities of Data Mining methods for analyzing medical data]. *Pedagogicheskii zhurnal* [Pedagogical Journal], 12 (5A), pp. 621-634. DOI: 10.34670/AR.2022.69.56.081

Keywords

Intelligent system, method, medical data, Data Mining, Apriori, plenty of candidates.

References

1. Agrawal R., Srikant R. (1994) Fast algorithms for mining asso-ciation rules in large databases. *Proc. 20th Intern. Conf. on Very Large Data Bases, VLDB*, pp. 487-499.
2. Ahmad Y. et al. (2015 Identifying Association Rules among Drugs in Prescription of a Single Drugstore Using Apriori Method. *Intelligent Information Management*, pp. 253-259.
3. Berestneva O.G., Sharopin K.A., Starikova A.V., Kabanova L.I. (2010 Tekhnologiya formirovaniya baz znanii v meditsinskikh informatsionnykh sistemakh [Technology of formation of knowledge bases in medical information systems]. *Izvestiya Yuzhnogo federal'nogo universiteta. Tekhnicheskie nauki* [News of of the Southern Federal University. Technical science], 8, pp. 32-37.
4. Bersegyan A.A. et al. (2009) Analiz dannykh i protsessov [Analysis of data and processes], 3th ed. Saint Petersburg.: BKhV-Peterburg Publ.
5. Bersegyan A.A., Kupriyanov M.S., Stepanenko V.V. (2007) *Tekhnologii analiza dannykh: Data Mining, Visual Mining, Text Mining, OLAP* [Data analysis technologies: Data Mining, Visual Mining, Text Mining, OLAP], 2nd ed. Saint Petersburg: BKhV-Peterburg Publ.
6. Billig V.A., Ivanova O.V., Tsaregorodtsev N.A. (2016) Postroenie assotsiativnykh pravil v zadache meditsinskoi diagnostiki [Construction of associative rules in the problem of medical diagnostics]. *Programmnye produkty i sistemy* [Software products and systems], 2 (114), pp. 146-157.
7. Dorozhkin A.K., Klimanov V.A. (2004) Ispol'zovanie Oracle Warehouse Builder dlya organizatsii khranilishch dannykh [Using Oracle Warehouse Builder to Organize Data Warehouses]. *Nauchno-tekhnicheskii vestnik informatsionnykh tekhnologii, mekhaniki i optiki* [Scientific and Technical Bulletin of Information Technologies, Mechanics and Optics], 14, pp. 105-110.
8. Marukhina O.V., Nepotrebnaya A.A. (2014) Primenenie metodov Data Mining dlya analiza meditsinskikh dannykh [Application of Data Mining methods for the analysis of medical data]. In: *Trudy Vserossiiskoi molodezhnoi nauchno-*

prakticheskoi konferentsii "Perspektivy razvitiya informatsionnykh tekhnologii" [Proc. All-Russian Conf. "Prospects for the Development of Information Technologies"]. Kemerovo, pp. 247-248.
9. Olson D.L., Delen D. (2008) *Advanced Data MiningTechniques*. Springer.
10. Pivovarova N.V., Vidunova S.I. (2016) Intellektual'nyi analiz dannykh v farmatsevticheskom biznese [Data Mining in the Pharmaceutical Business]. *Internet-zhurnal "Naukovedenie"* [Journal of Science Studies], 6(37), pp. 166.
11. Sadanova B.M. (2014) Ispol'zovanie BI-tekhnologii dlya mnogomernoi obrabotki dannykh [Using BI-technologies for multidimensional data processing]. In: *Trudy Vserossiiskoi molodezhnoi nauchno-prakticheskoi konferentsii "Perspektivy razvitiya informatsionnykh tekhnologii"* [Proc. All-Russian Conf. "Prospects for the Development of Information Technologies"]. Kemerovo, pp. 127-128.
12. Seidova A.S., Berestneva E.V., Osadchaya I.A. (2016) Primenenie metodov Data Mining dlya sozdaniya meditsinskikh baz znanii [Application of Data Mining methods to create medical knowledge bases]. In: *Trudy Mezhdunarodnoi konferentsii studentov, aspirantov i molodykh uchenykh "Perspektivy razvitiya fundamental'nykh nauk"* [Proc. Int. Conf. "Prospects for the Development of Fundamental Sciences"]. Tomsk, pp. 123-125.
13. Vercellis C. (2009) *Business Intelligence: Data Mining and Opti-mization for Decision Making*. Willey.