УДК 004.67

ИСПОЛЬЗОВАНИЕ МЕТОДОВ DATA MINING ПРИ ПРИНЯТИИ МЕДИЦИНСКИХ ДИАГНОСТИЧЕСКИХ РЕШЕНИЙ

Мокина Е.Е., Марухина О.В., Шагарова М.Д., Дубинина И.А.

ФГАОУ ВО «Национальный исследовательский Томский политехнический университет», Томск, e-mail: alisandra@tpu.ru

В настоящее время в связи с увеличением темпа жизни в крупных городах встает проблема роста числа больных неврологическими заболеваниями. Настоящая статья посвящена рассмотрению вопросов использования методов Data Mining при исследовании медицинских данных и построения системы поддержки принятия решений на основе результатов исследования, в данном случае выявления наличия неврологических заболеваний по результирующим показателям опросников качества жизни и тревоги и депрессии. С помощью методов Data Mining было построено дерево решений и выявлены правила вывода, которые легли в основу системы поддержки принятия решений. В статье были сформулированы основные требования к такой системе, которая позволит сократить временные затраты персонала медицинского учреждения на обработку данных опросников и предоставить рекомендации по постановке диагноза пациенту.

Ключевые слова: Data Mining, информационные системы, SF-36, HADS_T, системы поддержки принятия решений

USING DATA MINING METHOD IN THE MAKING OF MEDICAL DIAGNOSTIC DECISIONS

Mokina E.E., Marukhina O.V., Shagarova M.D., Dubinina I.A.

National Research Tomsk Polytechnic University, Tomsk, e-mail: alisandra@tpu.ru

At the present time the problem of the rise number of patients with neurological diseases appeared in connection with the increasing pace of big cities life. This article discusses the use of Data Mining methods in the medical data analysis and building making decisions support systems on the basis of research results, in this case detecting the presence of neurological diseases on summary indicators of the questionnaire of quality of life and anxiety and depression. With help of Data Mining methods have been built a decision tree and identified rules of inference for forming the making decisions support systems. In the article authors formulated the main requirements to such system, which will allow to reduce time expenses of health facility staff on data processing of the questionnaires and providing recommendations on the diagnosis of the patient.

Keywords: Data Mining, information system, SF-36, HADS_T, decision support system

В настоящее время информация является одним из важнейших ресурсов в мире, а информационные системы стали необходимым инструментом практически во всех сферах деятельности организаций, в том числе и медицине. Прикладные программы и инструменты медицинских информационных систем позволяют модернизировать рабочий процесс и совершенствовать методы диагностирования и лечения пациентов. Актуальность данной работы определяется, с одной стороны, необходимостью разработки программного продукта, поддерживающего процесс постановки диагноза и анализа показателей здоровья и качества жизни, а с другой стороны, необходимостью обработки и анализа существующих массивов данных, позволяющих рассматривать данные о пациентах в динамике, поскольку в течение периода лечения показатели могут иметь различные значения. Для разработки системы поддержки принятия решения следует установить взаимосвязи диагнозов заболеваний со значениями показателей по методикам оценки, принятыми в медицинской организации.

Анализ качества жизни пациентов

Исследования в данной области являются одним из важных направлений современной медицины, о чем свидетельствуют многочисленные специальные опросники, разработанные с целью оценить качество жизни, уровень физического и психологического самочувствия пациента. Данные опросники содержат варианты ответов на вопросы и составлены для подсчета по методу суммирования рейтингов. Одним из наиболее известных опросников для исследования качества жизни больных является опросник «Краткая форма оценки здоровья – MOS SF-36» (Medical Outcomes Study-Short Form). Методика SF-36 позволяет зарегистрировать и количественно оценить изменения качества жизни у больных с определенным видом заболевания на протяжении конкретного периода стационарного лечения, а также выделить составляющие, которые вносят наиболее весомый вклад в обусловленные лечением изменения качества жизни. Выбор данного опросника исследователями (медицинскими работниками) обусловлен

применением полученных результатов для оценки качества жизни при любых заболеваниях, возможностью сравнения результатов с данными российского популяционного контроля по соответствующим группам, а также возможностью оценки качества жизни больных комплексно (в том числе социальные и психологические нарушения). Для расчета значений результирующих показателей (РН «физический компонент здоровья» и МН «психический компонент здоровья»), согласно методике, применяются значения среднего отклонения популяционных показателей. Установлено, что для разных нозологий результирующие показатели качества жизни отличаются. Если при исследовании врачу не известны средние значения популяции, то рекомендуется использовать полученные ранее данные российского популяционного контроля по соответствующим группам [5]. Интегральная оценка качества жизни зависит и от географического региона, в котором проводится исследование. Поэтому для расчета интегральной оценки качества жизни пациентов с определенным видом нозологии (в нашем случае пациентов с неврологическими заболеваниями) конкретного географического региона (в нашем случае Томской области) целесообразно выявить свои популяционные показатели. При исследовании качества жизни могут быть использованы дополнительные средства, например такие, как опросник «Госпитальная шкала оценки тревоги и депрессии», который предназначен для оценки выраженности тревожных переживаний. С помощью данной методики можно оценить тревогу как преходящее состояние, переживаемое в особых ситуациях (в течение болезни).

Диагноз, в том числе связанный с неврологическими заболеваниями, влияет на показатели качества жизни и уровень тревожности пациента. Выявленные зависимости можно будет применить для постановки диагноза пациенту в соответствии с проводимым исследованиями по изучению его показателей качества жизни и уровня тревоги и депрессии (по методикам SF-36, «Госпитальная шкала оценки тревоги и депрессии» и т.д.).

Поддержка процесса постановки диагноза пациенту с определенной нозологией с помощью системы поддержки принятия решения включает в себя такие задачи, как: анализ данных опроса пациента и сопоставление его с моделью для постановки диагноза; отслеживание изменений показателей качества жизни пациента в динамике, поскольку в ходе лечения дополнительная диагностика позволяет улучшить лечебный процесс и прогнозировать изменения показателей качества жизни.

Использование интеллектуального анализа данных в поиске закономерностей

Основными задачами в ходе анализа данных для разработки интеллектуального компонента системы поддержки принятия решения являются:

- выявление скрытых закономерностей в имеющихся данных с помощью средств интеллектуального анализа данных;
- построение правила логического вывода по результатам применения средств интеллектуального анализа данных;
- построение правила выдачи сообщений о предполагаемом диагнозе и прогнозе изменения показателей;
- разработка логических правил вывода результатов прохождения опросов на языке программирования, формирование базы знаний.

После обработки накопленного массива данных можно выявить зависимости и на их основе построить правила логического вывода и представить знания в виде предложений типа если (условие), то (действие). Одним из наглядных способов представления таких результатов исследования в интеллектуальном анализе данных являются деревья решающих правил, вид которых понятен как специалисту в области информационных технологий, так медицинскому работнику [1, 2].

В качестве инструментария для построения дерева решений и логических правил выбрана среда RapidMiner, представляющая собой комплексную систему, реализующую методы Data Mining (методы интеллектуального анализа, ИАД) и статистического анализа, которая обладает набором алгоритмов для обработки и анализа, в том числе обработку больших массивов данных. Работая с любым набором данных, как в конструкторе, можно добавлять различные операторы ввода/вывода, обработки, визуализации, анализа и т.п. [3]. Весь процесс представлен в древовидной форме (рис. 1).

Пример графического представления дерева решений определения тревоги (HADS_T) в зависимости от значений показателей МН («психологический компонент здоровья») и РН («физический компонент здоровья»), сформированный программой Rapid Miner в случае отсутствия показателей тревоги и депрессии, представлен как в графическом виде, так и в виде текста на рис. 2.

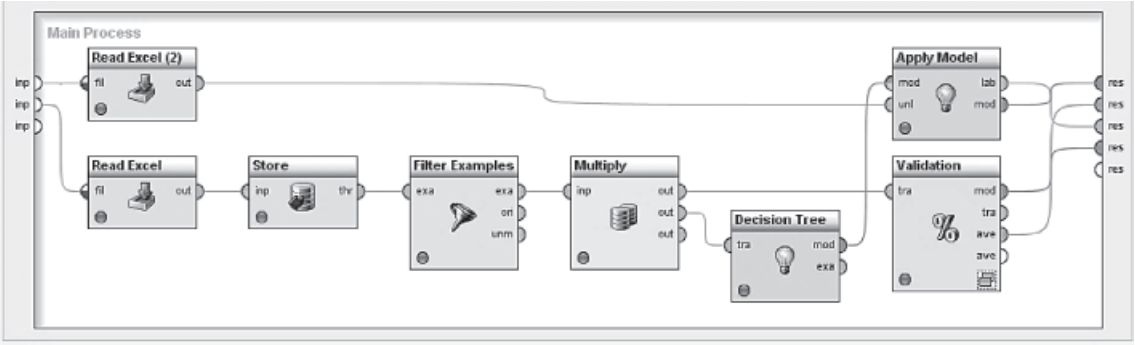

Рис. 1. Этапы процесса анализа в Rapid Miner

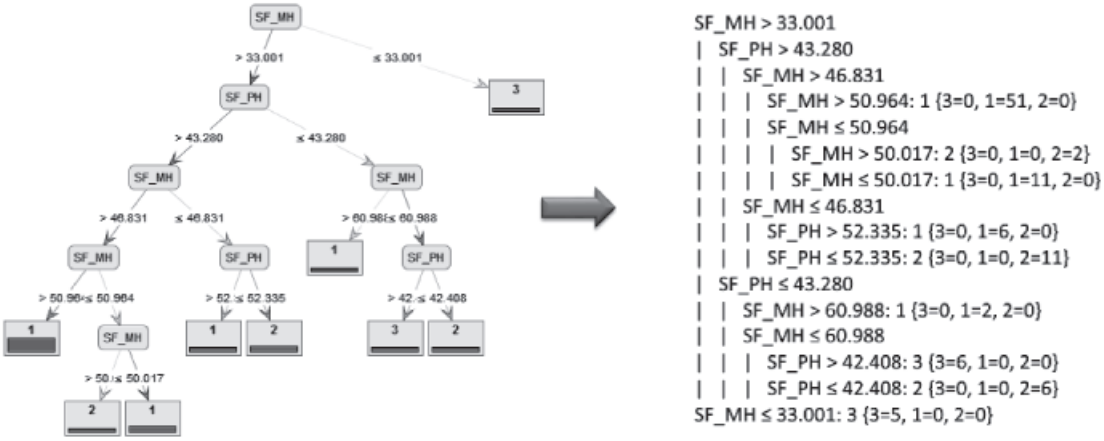

Рис. 2. Дерево решений. Значения 1, 2, 3 показателя HADS_T:
1 – норма; 2 – субклинически выраженная тревога/депрессия;
3 – клинически выраженная тревога/депрессия

При исследовании пациентов с неврологическими заболеваниями и качества жизни пациентов было проведено исследование на предмет зависимости физического компонента здоровья, психологического компонента здоровья, показателей тревоги и депрессии от диагностируемого заболевания.

В качестве примера ниже приведена часть правил определения диагноза, в зависимости от физического компонента здоровья (SF_PH), психологического компонента здоровья (SF_MH), уровня депрессии (HADS_D) и тревоги (HADS_T), диагноза (D со значениями здоров, G20, G24, G35), где:

HADS_T ≤ 14.500
| SF_MH ≤ 60.988
| | SF_PH > 42.408: G20 {G20 = 6, G35 = 0, G24 = 0, здоров = 0}
| | SF_PH ≤ 42.408: G24 {G20 = 0, G35 = 0, G24 = 6, здоров = 0}

Система поддержки принятия решения

В соответствии с выявленными задачами и требованиями при проведении медицинских исследований необходимо разработать систему информационной поддержки, обеспечивающей процесс прохождения тестирования пользователем (пациентом), расчета результирующих показателей, а также постановки диагноза и прогноза изменения показателей [4]. При разработке следует уделить особое внимание таким функциям, как:

- механизм определения диагноза, определения динамики и характеристики изменения показателей по результатам проведенного опроса;
- обработка полученных данных, введенных пользователем (опрашиваемым), и предоставление результата после каждого пройденного опроса;
- хранение полученных результатов по пройденному тесту в систематизированном виде;
- обеспечение доступа (исследователю) к хранящимся результатам для выборки данных по необходимым параметрам.

● экспорт выбранных данных, необходимых исследователю для анализа в других статистических программах.

При разработке системы информационной поддержки следует предусмотреть следующие возможности для обеспечения гибкости системы в случае изменений в процессе работы специалистов:

● изменение/добавление популяционных значений показателей стандартного отклонения и математического ожидания;

● добавление новых методик (опросников);

● добавление новых правил вывода определения диагноза, прогноза изменений показателей для расширения базы знаний;

● хранение номера проводимого исследования, цель проводимого исследования (в частности, для методики SF-36 возможность прохождения теста должна быть представлена в вариантах: с функцией подсчета показателей (PF, RP, BP, GH, VT, SF, RE, MH), без расчета результирующих показателей PH («физический компонент здоровья» и MH («психический компонент здоровья») или с функцией определения диагноза).

Модуль для работы с данными по проведенным опросам (исследованиям) должен позволять производить: расчет результирующих значений показателей по пациенту (возможность применения необходимых средних значений популяций), расчет средних значений популяции (при выборе необходимой совокупности записей).

Исходя из требований, предъявляемых к разрабатываемому программному решению, модель СППР можно представить следующим образом (рис. 1).

После того как пользователь отвечает на вопросы, «сырые» данные (ответы на вопросы) сохраняются в системе, а затем запускается процесс обработки результатов: по определенному алгоритму (в зависимости от конкретных ответов) рассчитываются показатели и сохраняются в базе данных. На следующем шаге в зависимости от того, какое по счету проводится исследование, запускается процесс выборки фактических значений (не прогнозируемых). Эти данные сравниваются с правилами логического вывода (продукционными правилами) и осуществляется вывод предполагаемого диагноза, также вывод прогнозируемых значений показателей.

База знаний может быть представлена набором правил, на основе которых алгоритм логического вывода определяет выходные данные. Подходом к разработке системы поддержки принятия решения выбран подход, основанный на правилах логического вывода. Данный подход выбран исходя из поставленных задач (определение диагноза, анализ данных в динамике), требуемых решения, и наличием имеющихся накопленных данных о диагнозах, показателях здоровья пациентов с заболеваниями. Такие правила позволяют представить знания в виде предложений типа если (условие), то (действие). После обработки накопленного массива данных можно получить выявленные зависимости и на основе этих зависимостей построить правила вывода.

Пример кодирования правила:

IF HADS_T < = 14,500 AND SF_MH < = 60.988 AND SF_PM > 42,408 THEN: D = G20
IF HADS_T < = 14,500 AND SF_MH < = 60.988 AND SF_PM < = 42,408 THEN: D = G24

Рис. 3. Система поддержки принятия решения

Таким образом, можно сформировать все полученные правила, совокупность которых станет базой знаний для системы поддержки принятия решений, а программирование модуля «Сравнение показателей с логическими правилами» позволит определить диагноз, прогноз изменений показателей, обеспечит вывод новых заключений.

При проектировании структуры базы данных учтены положения, приведённые ниже.

● Респондент (опрашиваемый) может несколько раз проходить тестирования по опросникам. По одному опроснику для респондента может проводиться тестирование несколько раз.

● Опросник состоит из вопросов. Некоторые вопросы относятся к определенной категории вопросов. Вопросы в опросниках не повторяются.

● Один вопрос состоит из нескольких вариантов ответов (пользователь должен выбрать только один, подходящий ему). Если несколько вопросов относятся к одной категории, то эти вопросы могут иметь одинаковые варианты ответов.

● После прохождения тестирования на основе полученных ответов от опрашиваемого рассчитывается показатель.

● Вопрос относится к определенному показателю. Показатель может состоять из нескольких вопросов. Пример показателя: эмоциональное состояние, моральное состояние и т.п.

● Рассчитанное значение для респондента находится в определенном числовом диапазоне.

● Для каждого показателя есть соответствующая граница, которая определяет нахождение показателя в норме, выше или ниже границы нормального значения. По одному опроснику вычисляется несколько показателей.

Проектируемая система поддержки принятия решений включает три основные роли пользователей, функции которых представлены на рис. 4, 5.

С помощью компьютерная обработки тестов есть возможность достигнуть таких положительных эффектов, как повышение скорости работы благодаря быстроте обработки данных и получению результатов тестирования; исключение ошибок обработки исходных данных, которые могут иметь место при ручных методах расчета, а также возможность накапливать результаты тестирования и анализировать их в динамике.

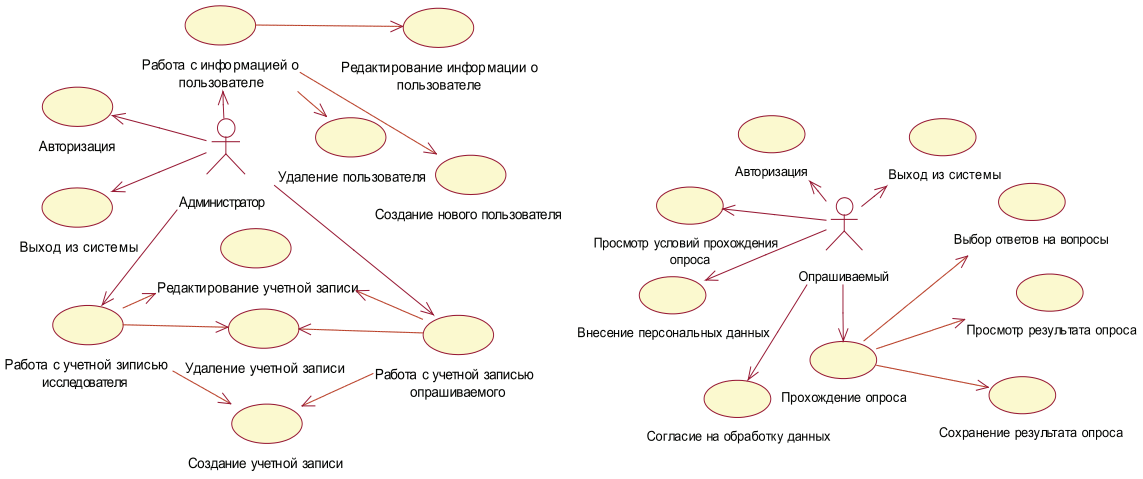

Рис. 4. Варианты использования системы пользователем и опрашиваемым

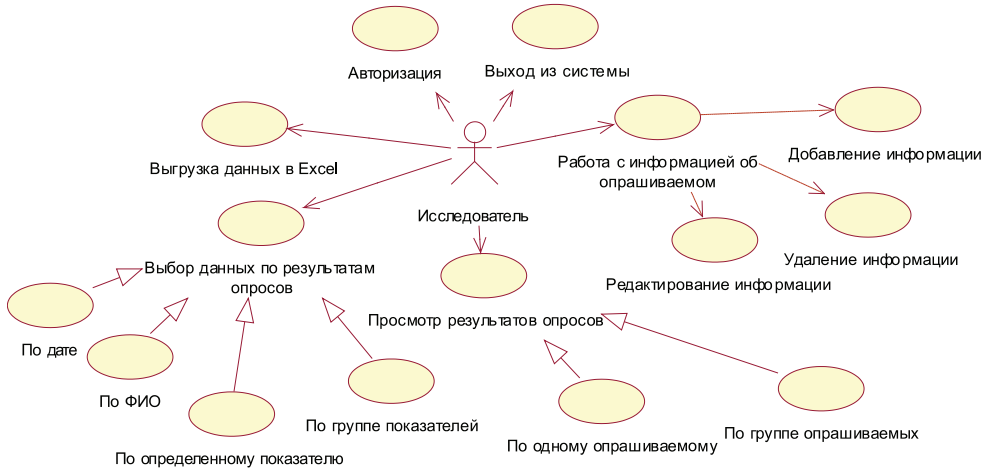

Рис. 5. Варианты использования системы исследователем

Заключение

Предложен подход к проектированию системы поддержки принятия решений с целью определения диагноза, прогнозирования состояния здоровья и прогноза показателей в динамике. В соответствии с рассмотренным подходом можно сделать вывод о том, что главным процессом является установление зависимостей в накопленных данных и формирование правил логического вывода, применяемых далее в программном решении, и в этом процессе важную роль играют методы Data Mining. Так как именно от правильности построения модели и выбора критериев для анализа зависит адекватность и ценность вывода рекомендаций для определения диагноза, отслеживание показателей здоровья в динамике и работа системы в целом. База знаний, представленная продукционными правилами, полученными с помощью средств интеллектуального анализа данных (в нашем случае построения деревьев решений), позволяет решить данные задачи.

Работа выполнена при финансовой поддержке РФФИ, проект № 14-06-00026, проект № 14-07-00675.

Список литературы

1. Берестнева О.Г., Осадчая И.А., Немеров Е.В. Методы исследования структуры медицинских данных // Вестник науки Сибири. – 2012. – № 1 (2). – С. 333–338.

2. Берестнева О.Г., Пеккер Я.С. Выявление скрытых закономерностей в сложных системах // Известия Томского политехнического университета. – 2009. – Т. 315. – № 5. – С. 138–143.

3. Марухина О.В., Мокина Е.Е., Берестнева Е.В. Применение методов Data Mining для выявления скрытых закономерностей в задачах анализа медицинских данных // Фундаментальные исследования. – 2015. – № 4–0. – С. 107–113.

4. Мещеряков Р.В., Балацкая Л.Н., Чойнзонов Е.Л. Специализированная информационная система поддержки деятельности медицинского учреждения // Информационно-управляющие системы. – 2012. – № 5. – С. 51–56.

5. Новик А.А., Ионова Т.И. Исследование качества жизни в медицине: учебное пособие для вузов / под ред. Ю.Л. Шевченко. – М.: Изд-во ГЭОТАР-МЕД., 2004. – 304 с.

References

1. Berestneva O.G., Osadchaja I.A., Nemerov E.V. Metody issledovanija struktury medicinskih dannyh // Vestnik nauki Sibiri. 2012. no. 1 (2). pp. 333–338.

2. Berestneva O.G., Pekker Ja.S. Vyjavlenie skrytyh zakonomernostej v slozhnyh sistemah // Izvestija Tomskogo politehnicheskogo universiteta. 2009. T. 315. no. 5. pp. 138–143.

3. Marukhina O.V., Mokina E.E., Berestneva E.V. Primenenie metodov Data Mining dlja vyjavlenija skrytyh zakonomernostej v zadachah analiza medicinskih dannyh // Fundamentalnye issledovanija. 2015. no. 4–0. pp. 107–113.

4. Meshcheryakov R.V., Balatskaya L.N., Choinzonov E.L. Specializirovannaja informacionnaja sistema podderzhki dejatelnosti medicinskogo uchrezhdenija // Informacionno-upravljajushhie sistemy. 2012. no. 5. pp. 51–56.

5. Novik A.A., Ionova T.I. Issledovanie kachestva zhizni v medicine. Uchebnoe posobie dlja vuzov. [pod red. Ju.L. Shevchenko] Moskva: Izd-vo GJeOTAR-MED., 2004. 304 p.