DOI 10.23859/1994-0637-2019-1-88-3
УДК 303.732.4

Рапаков Георгий Германович
Кандидат технических наук, доцент, Вологодский государственный университет (Вологда, Россия)
E-mail: grapakov@yandex.ru

Rapakov Georgij Germanovich
PhD in Technical Sciences, Associate Professor, Vologda State University (Vologda, Russia)
E-mail: grapakov@yandex.ru

Банщиков Геннадий Трофимович
Доктор медицинских наук, главный внештатный специалист-терапевт Департамента здравоохранения области, врач-кардиолог, БУЗ ВО «Вологодская областная клиническая больница» (Вологда, Россия)
E-mail: vologdauzo@inbox.ru, vol_obl_bol1@mail.ru

Banshchikov Gennadii Trofimovich
Doctor of Medicine, M.D. Cardiologist, Vologda region clinical hospital (Vologda, Russia)
E-mail: vologdauzo@inbox.ru, vol_obl_bol1@mail.ru

Горбунов Вячеслав Алексеевич
Доктор физико-математических наук, профессор, Вологодский государственный университет (Вологда, Россия)
E-mail: gorbunov1945@inbox.ru

Gorbunov Vyacheslav Alekseevich
Doctor of Physico-Mathematical Sciences, Professor, Vologda State University (Vologda, Russia)
E-mail: gorbunov1945@inbox.ru

Малыгин Леонид Леонидович
Доктор технических наук, доцент, президент ООО «Малленом Системс» (Череповец, Россия)
E-mail: malygin@mallenom.ru

Malygin Leonid Leonidovich
Doctor of Technical Sciences, Associate Professor, President of company "Mallenom Systems" (Cherepovets, Russia).
E-mail: malygin@mallenom.ru

Ревелев Игорь Михайлович
Врач (сердечно-сосудистый хирург), БУЗ ВО «Вологодская городская больница №2» (Вологда, Россия)
E-mail: postservice2016@yandex.ru

Revelev Igor' Mikhailovich
Physician, Cardiovascular Surgeon, Vologda municipal hospital no. 2 (Vologda, Russia)
E-mail: postservice2016@yandex.ru

АНАЛИЗ ДАННЫХ МЕДИКО-СОЦИОЛОГИЧЕСКОГО МОНИТОРИНГА НА ОСНОВЕ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ

DATA ANALYSIS OF MEDICAL-SOCIOLOGICAL MONITORING WITH MACHINE LEARNING METHODS

Аннотация. Исследованы методы машинного обучения при обработке данных медико-

Abstract. This article studies the methods of the machine learning while processing the data of

социологического опроса. При помощи компьютерной имитации построена модель правил связывания на основе алгоритма априорных значений. Выделены пять решающих правил, представляющих интерес для анализа, с наибольшим значением лифта. Применение метода деревьев решений позволило выявить целевую группу влияния. При оценке модели вычислены важности предикторов. Верификация результатов классификации позволила выполнить проверку достоверности добытого знания. Результаты анализа использованы при принятии управленческих решений в региональной системе медицинской профилактики.

Ключевые слова: компьютерное моделирование, машинное обучение, ассоциативные правила, деревья решений, поддержка принятия решений

the medical-sociological monitoring. The binding rules model on the basis of the priori value algorithm was designed with help of the computer imitation. Five main rules, meaningful for the analysis, with the biggest hoist meaning were identified. The application of the decision tree method allowed to obtain the influence target group. While the model's assessing the meanings of the predictors were calculated. The verification of the classification results helped to check the reliability of the obtained knowledge. The analysis results were used for making the management decisions in the regional system of the medical prevention.

Keywords: computer modeling, machine learning, association rules, decision trees, support of decision making

Введение

Для количественного прогнозирования поведения зависимой переменной при изменении факторов влияния, оценки значимости ковариат в социологических исследованиях широко применяется регрессионный и дискриминантный анализ. К числу современных универсальных подходов, позволяющих выполнить обнаружение связанных событий, образующих социокультурный поведенческий паттерн, относят ассоциативные правила. На основе применения методов машинного обучения с использованием деревьев решений может быть построена модель связывания; соотнесены признаки целевой аудитории; выявлены зависимости между переменными опроса для различных типов данных [14], [15]. Практическая значимость работы определяется результатами исследования методов машинного обучения в целях снижения социально-экономического бремени смертности населения, обусловленной сердечно-сосудистыми заболеваниями (ССЗ) [8]. В связи с этим особую актуальность приобретает выявление проблем региональной медицинской профилактики и их решение при помощи методов интеллектуального анализа данных. Целью настоящей работы является исследование методов машинного обучения на основе компьютерного моделирования данных медико-социологического опроса. Совместная интерпретация наглядных элементарных высказываний позволяет выделить социокультурные шаблоны и выполнить целенаправленный анализ поведенческих стереотипов, что обусловливает новизну работы. Использование компьютерного моделирования позволяет повысить эффективность формирования управленческих стратегий медицинской профилактики в системе медико-социальной поддержки населения [7], [9], [10].

Основная часть

С точки зрения машинного обучения (Machine Learning) поиск ассоциативных правил (Association Rule Induction), или правил связывания, относится к отдельному клас-

су обучения без учителя (Unsupervised Learning). Задача состоит в том, чтобы выявить правила – связанные между собой группы ответов на вопросы анкеты с учетом заданных ограничений.

На множестве объектов X задано n бинарных признаков $F = \{f_1, ..., f_n\}$, $f_j: X \to \{0, 1\}$. Имеется выборка $X^l = \{x_1, ..., x_l\} \subset X$. В нашей задаче выборка объектов соответствует набору анкет. В части факторов риска рассматриваются бинарные признаки – ответы на вопросы анкеты. Единичное значение признака $f_j(x_i) = 1$ свидетельствует о положительном ответе на j-вопрос в i-й анкете. Каждому набору признаков $\varphi \subseteq F$ ставится в соответствие предикат $\varphi(x)$, равный конъюнкции всех признаков из φ:

$$\varphi(x) = \bigwedge_{f \in \varphi} f(x), \; x \in X.$$

Для $\varphi(x) = 1$ говорят, что признаки набора φ совместно встречаются у объекта x. Для количественной оценки связи используются два показателя: поддержка (support) и достоверность (confidence). Поддержку набора φ в X^l принято описывать функцией

$$\nu(\varphi) = \frac{1}{l} \sum_{i=1}^{l} \varphi(x_i).$$

Для ограничения количества правил вводят параметр минимальной поддержки (minsupport) δ. Набор $\varphi \subseteq F$ называется часто встречающимся, если $\nu(\varphi) \geq \delta$. В нашем случае поддержка supp – это число анкет, содержащих как условие, так и следствие относительно их общего количества.

Пара непересекающихся наборов φ, $y \subseteq F$ называется ассоциативным правилом $\varphi \to y$, если выполнены следующие условия [17]:

$$\nu(\varphi \cup y) \geq \delta; \qquad \nu(y \mid \varphi) \equiv \frac{\nu(\varphi \cup y)}{\nu(\varphi)} \geq \theta.$$

Левая часть первого неравенства называется достоверностью правила. Параметр минимальной достоверности (minconfidence) θ позволяет ограничить количество правил. Достоверность ассоциативного правила conf используется для оценки его точности и в нашем случае представляет собой отношение числа анкет, содержащих как условие, так и следствие, к количеству анкет, содержащих только условие. Таким образом, $\nu(y|\varphi)$ можно рассматривать как оценку условной вероятности. Таким образом, для ассоциативного правила $\varphi \to y$ справедливо: наборы φ и y совместно часто встречаются – не реже, чем в доле случаев δ; если встречается набор φ, то с частотой не менее θ встречается и набор y. Дополнительным показателем, оценивающим значимость правила, является лифт (Lift). Лифт – это отношение частоты появления условия среди объектов, содержащих также и следствие, к частоте появления собственно следствия. Лифт рассматривают как обобщенную меру связи признаков и используют для оценки значимости правила. Алгоритм ассоциативных пра-

вил также может быть использован при сегментации опрашиваемых по поведению и при анализе предпочтений.

Традиционно реализация извлечения набора правил из данных осуществляется с использованием алгоритма APriory (Agrawal and Srikant, 1994). Ответы на вопросы анкеты рассматриваются как связанные причинно-следственными отношениями «из А следует С»: условие (Antecedent) → следствие (Consequent). Символ «→» (стрелка) в записи используется для отображения правила. Ассоциативные правила позволяют количественно описать связи между вопросами анкеты, которые соответствуют условиям и следствиям.

При этом выделяются правила с наибольшим информационным содержимым. При решении задач ассоциаций существуют различные рекомендации по выбору пределов поддержки и достоверности. Строгий алгоритм отсутствует. Обычно для исключения тривиальных правил рекомендуется ограничить верхний порог support несколькими десятками процентов. Низкие значения minsupport приводят к генерации статистически необоснованных правил, поэтому для них используется дополнительная проверка на значение lift. Занижение minconfidence до нескольких процентов неоправданно увеличивает количество правил и снижает их ценность.

Модели деревьев решений позволяют классифицировать будущие наблюдения на основе набора решающих правил. Использование деревьев решений ограничено их неспособностью находить наилучшие (наиболее полные и точные) правила. В исследовании был использован алгоритм C&RT – Classification and Regression Trees (Breiman, Friedman, Olshen and Stone,1984). Алгоритм C&RT является одним из наиболее популярных и используется в задачах классификации и регрессии автоматического анализа данных. Результатом его работы является бинарное дерево решений – иерархическая структура правил. Правило – это логическая конструкция вида «если... то...» (If-Then), которая представляет собой путь от вершины до листа (конечного узла) дерева. Для бинарного дерева решений каждый узел имеет двух потомков [16].

В ходе рекурсивной дихотомии метод делит исходное множество на два подмножества так, что записи в каждом из них являются гомогенными. На очередном шаге разбиение проводится по той переменной, которая делает его наилучшим. В качестве правила разбиения выступает статистический критерий – индекс Gini $g(t)$, при помощи которого можно дать оценку «расстояния» между распределениями классов. Индекс $g(t)$ будет равен нулю, если все записи в узле будут относиться к одной и той же категории. Для $p(j)$-вероятности класса j в текущем узле t индекс $g(t)$ определяют как

$$g(t) = 1 - \sum_j p^2(j).$$

Результат работы алгоритма зачастую представляет собой сложное дерево с большим количеством узлов и ветвей, непригодное для интерпретации. Ценность правила становится меньше с уменьшением количества объектов, для которых оно справедливо. Необходимо избегать переобучения модели. С практической точки зрения предпочтительным является такой результат разбиения, при котором малому количеству узлов соответствует большое количество объектов. Для ограничения глубины дерева используют оценку целесообразности дальнейшего разбиения. Так

называемая «ранняя остановка» приводит к ухудшению классификации. Поэтому вместо остановки используют отсечение.

Точность, обеспечиваемая деревом решений, определяется отношением правильно классифицированных объектов к их общему количеству. Для большинства практических задач отсечению или замене поддеревом подлежат те ветви, по отношению к которым это действие не приведет к возрастанию ошибки распознавания. Особенностью алгоритма C&RT является то, что при отсечении достигается компромисс между оптимальным размером дерева и точной оценкой вероятности ошибочной классификации. После получения последовательности деревьев из нее выбирается лучшее. Используются тестирование на соответствующей выборке и механизм перекрестной проверки.

Межведомственная работа по формированию регионального здоровьесберегающего пространства регулируется рядом нормативно-правовых актов [2], [5]. Болезни системы кровообращения (БСК) находятся на первом месте по причинам смерти в регионе. За период 2009–2014 гг. значения показателя смертности от БСК (на 100 тыс. человек населения) в Вологодской области в 1,18–1,28 раза выше, чем в Российской Федерации, и в 1,09–1,17 раза выше, чем в Северо-Западном федеральном округе. Ожидаемые результаты реализации Государственной программы предполагают снижение смертности от БСК к 2020 г. до 649,4 случая на 100 тыс. населения, что соответствует значению аналогичного показателя в РФ в 2014 г. Значимым показателем здоровья населения является инвалидность. БСК занимают первое место в региональной структуре первичной инвалидности. Географическое распределение показателя первичного выхода на инвалидность в связи с БСК (на 10 тыс. взрослого населения) в Вологодской области в 2014 г. представлено на рис. 1 [6].

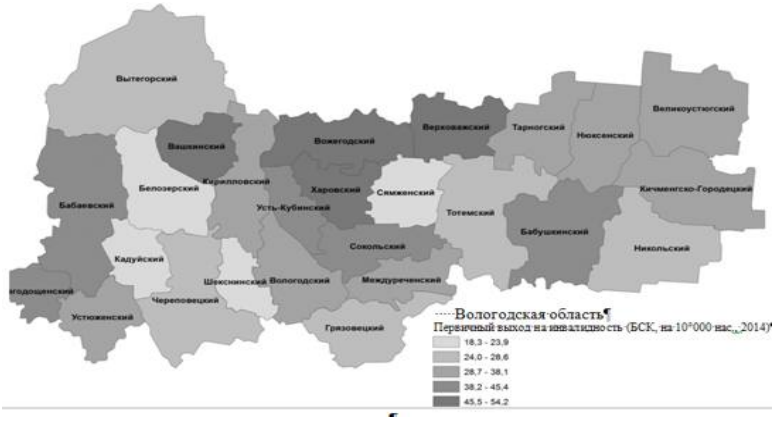

Рис. 1. Географическое распределение показателя первичного выхода на инвалидность в связи с БСК в разрезе муниципальных образований области в 2014 г.

Ознакомление с научными источниками позволило сопоставить цели и методы исследования с данными аналогичной литературы. Применение технологии поиска ассоциативных правил, основанной на модифицированном аппарате линейной алгебры с использованием процедуры самоорганизации данных и эффекта информа-

ционного структурного резонанса, позволяет находить высокоточные связи элементов исходного множества транзакций с заданным элементом [1]. В работе [3] обсуждается поиск ассоциативных правил применительно к анализу смертности. Выявлены заболевания, которые распространены преимущественно в пределах одного района г. Новокузнецка: 92,84 % случаев закупорки и стеноза передней мозговой артерии зарегистрированы в Заводском районе. В публикации [11] содержатся результаты использования ассоциативных правил для выделения социокультурных шаблонов для последующей коррекции здоровьесберегающих активностей населения и модификации факторов риска БСК. В монографии [12] обосновано применение методов и алгоритмов интеллектуальной поддержки принятия управленческих решений в задаче формирования регионального здоровьесберегающего образовательного пространства. На основе ассоциативных правил построена модель связывания. При помощи метода деревьев решений выявлена целевая аудитории влияния. Обнаружены связанные события и выделен поведенческий паттерн. Сформированы решающие правила для принятия управленческих решений в сфере региональной профилактики заболеваемости. Решению задач фармакоэкономического моделирования содействуют смешанные подходы, основанные на привлечении традиционной статистики и аналитических моделей. В области фармакоэкономики наиболее часто встречаются модели Маркова и деревья решений [4]. Модель на основе метода деревьев решений (random forest) позволила повысить качество прогноза отказов в условиях малого количества поломок [13]. Алгоритм строит большое количество деревьев решений по набору на основе исходной обучающей выборки с возвращением. При настройке используются данные об отказах. Обучение модели выполнялось на сведениях, соответствующих нормальному режиму работы оборудования. Для выявления отказов и аномалий используется разность показаний фактического и прогнозного нормального сигналов в последующий промежуток времени.

Информационной базой исследования являются результаты пилотного опроса. Объем выборки обеспечивает точность оценки не ниже 7% ($\alpha = 0,95$). Контингенты представлены тремя группами респондентов: врачи и пациенты регионального центра, пациенты муниципальных образований региона (область).

Сетевой граф иллюстрирует силу взаимосвязей между заданными полями набора данных. Направленная сеть используется для отображения силы взаимосвязи между несколькими полями и значениями одного целевого поля. В качестве стандартного способа показа связей между полями используется подход – сильные связи тяжелее. Результаты визуализации данных медико-социологического опроса при помощи сетевых графов для целевой группы «ССЗ_АГ = да» представлены на рис. 2. На основе данных мониторингового медико-социологического исследования построена модель правил связывания при помощи алгоритма априорных значений. Из всего множества условий, отобранных моделью связывания APriory при варьировании параметров support (от 10 до 100 %) и confidence (от 80 до 100 %), были выделены пять, имеющих отношение к следствию (ССЗ_АГ= да) с достоверностью 100 % и представляющих интерес для анализа (максимальный лифт). Правила связывания объединяют следствие (консеквент), в качестве которого выступает наличие у респондента артериальной гипертонии (ССЗ_АГ = да), с набором условий (антецедентов) (см. таблицу).

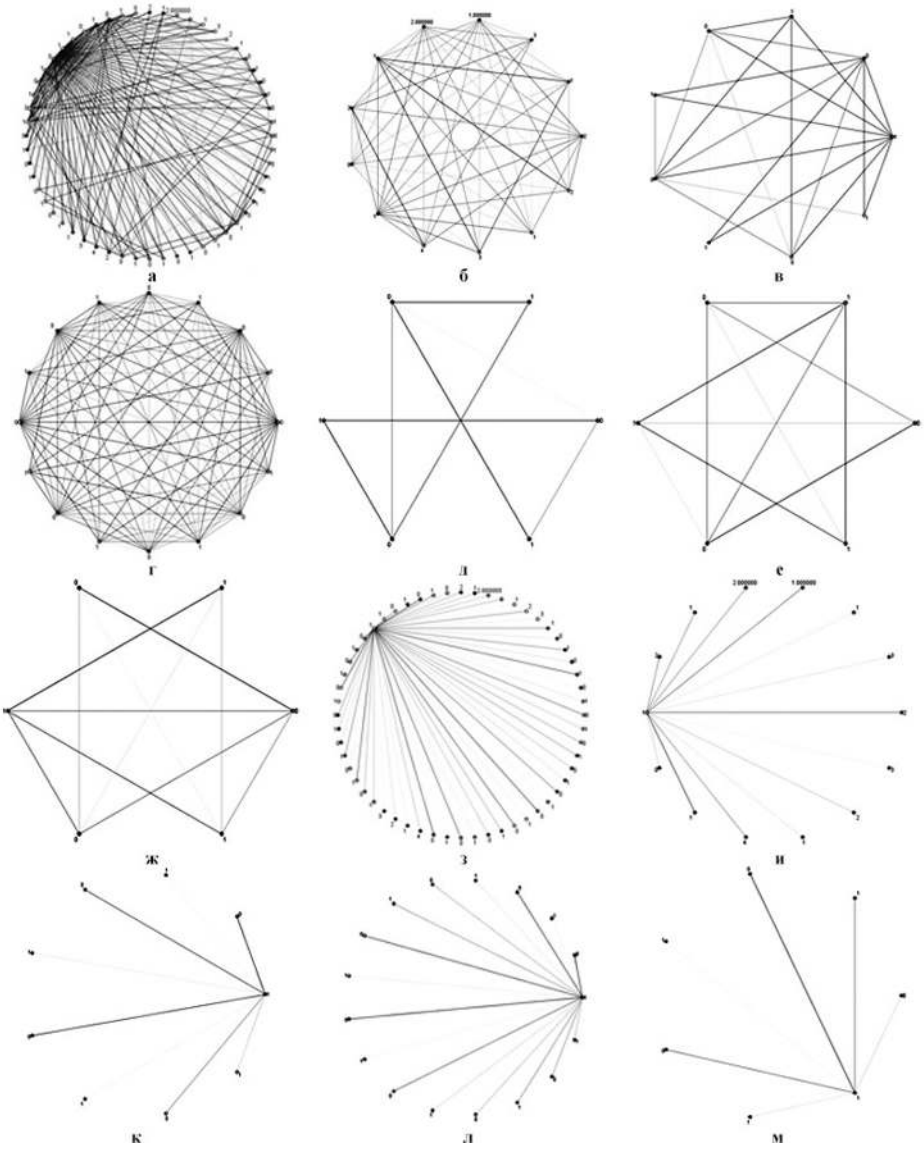

Рис. 2. Визуализация результатов медико-социологического опроса при помощи сетевых графов

Итогом обнаружения полезных зависимостей являются закономерности, присущие анализируемым данным, которые не ограничены традиционными для статистических подходов априорными предположениями о распределении показателей и структуре выборки. В ходе работы алгоритма C&RT были выявлены шаблоны, отражающие системные связи и закономерности в разнородных исходных данных. Многоаспектные взаимоотношения представлены моделью решающих правил, сформированных деревом решений. Результаты применения алгоритма показаны на

рис. 3. При этом целевая группа влияния (Target Group) представлена пациентами из муниципальных образований региона (область) в возрасте более 41,5 года, врачами и пациентами регионального центра в возрасте более 53,5 лет. Объем выделенной целевой аудитории составляет 62,8 %.

Получена оценка влияния переменных набора данных на результаты классификации (Variable Importance). Определяющий вклад в наличие у респондента артериальной гипертонии (ССЗ_АГ = да) вносит возраст (60 %). Оценка значимости принадлежности к контингенту составляет 25 %. Оставшиеся семь предикторов определяют равные доли влияния. Верификация модели позволяет выполнить проверку достоверности добытого знания. Точность классификации прогностической модели была определена при помощи процедуры анализа на основе таблицы сопряженности дерева решений. Полученные результаты показывают, что 81,4 % значений, предсказанных моделью, соответствуют фактическим, что вполне достаточно для большинства практических приложений метода.

Таблица

Модель правил связывания при помощи алгоритма априорных значений

Консеквент	Антецеденты	Поддержка, %	Достоверность, %	Рост, количество раз
ССЗ_АГ	НФР – холестерин; НФР – индекс массы тела; обладаю тонометром	15,349	100	1,667
ССЗ_АГ	НФР – холестерин; НФР – индекс массы тела; знаю артериальное давление; обладаю тонометром	14,884	100	1,667
ССЗ_АГ	НФР – холестерин; НФР – индекс массы тела; знаю холестерин; обладаю тонометром	13,023	100	1,667
ССЗ_АГ	НФР – холестерин; НФР – индекс массы тела; знаю холестерин; знаю артериальное давление	13,023	100	1,667
ССЗ_АГ	НФР – холестерин; НФР – индекс массы тела; семейное положение; обладаю тонометром	11,628	100	1,667

Примечание. НФР – наличие фактора риска.

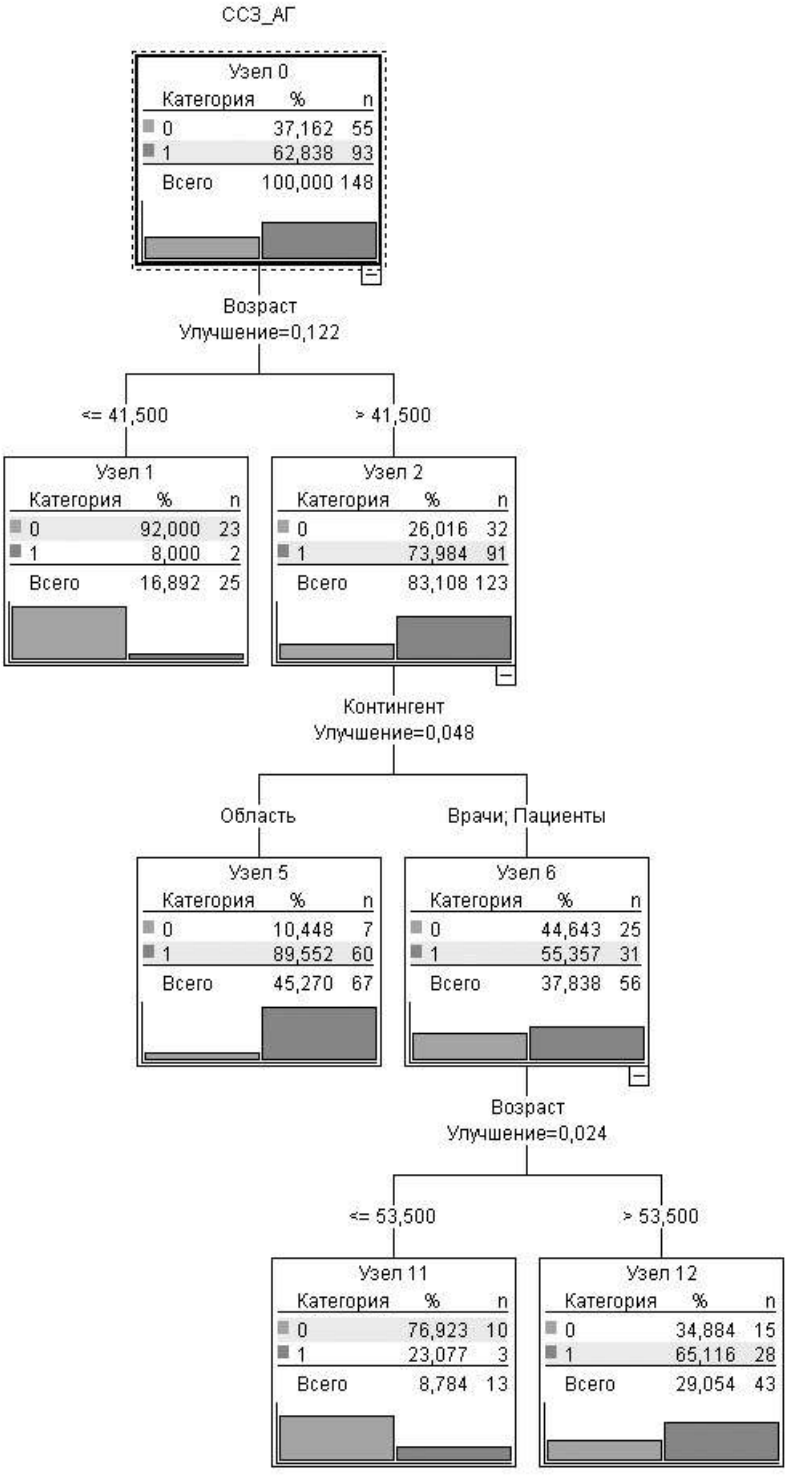

Рис. 3. Модель дерева решений

Выводы

На основе мониторингового медико-социологического исследования построена модель правил связывания при помощи алгоритма априорных значений. Из всего множества условий, отобранных моделью связывания APriory при варьировании параметров support (от 10 до 100 %) и confidence (от 80 до 100 %), были выделены пять, имеющих отношение к следствию (ССЗ_АГ = да) с достоверностью 100 % и представляющих интерес для анализа (максимальный лифт). Визуализация связей набора ассоциативных правил для признаков социологического опроса выполнена при помощи сетевого графа. Модели деревьев решений позволяют классифицировать будущие наблюдения на основе набора решающих правил. В результате применения алгоритма C&RT (Classification and Regression Trees) выявлена целевая группа влияния, представленная пациентами в возрасте более 41,5 года из муниципальных образований региона (область), врачами и пациентами в возрасте более 53,5 года из регионального центра. Объем выделенной целевой аудитории составляет 62,8 %. Оценка важности предикторов свидетельствует о том, что большое влияние на наличие у респондента артериальной гипертонии оказывает возраст (60 %). Оценка значимости принадлежности к контингенту составляет 25 %. Верификация модели была определена при помощи процедуры анализа на основе таблицы сопряженности дерева решений. Полученные результаты показывают, что 81,4 % значений, предсказанных моделью, соответствуют фактическим, что вполне достаточно для большинства практических приложений метода. Дальнейшие перспективы работы связаны с интеллектуальным анализом данных для когорты больных, имеющих онкопатологию, и с увеличением числа предикторов.

Литература

1. Асеев М. Г., Дюк В. А. Применение системы Deep Data Diver для решения задачи анализа рыночных корзин // Труды СПИИРАН. 2004. Т. 1. № 2. С. 127–134.

2. Вологда – город долгожителей: концепция активного долголетия на территории муниципального образования «Город Вологда» на период до 2035 года: решение Вологодской городской Думы от 29 декабря 2014 г. № 129. URL: http://vologda-portal.ru/oficialnaya_-vologda/index.php?SECTION_ID=8401

3. Жилина Н. М., Фадеева А. Е., Чеченин Г. И. Анализ смертности населения г. Новокузнецка на основе электронной базы данных за период 1999–2007 гг. // Социальные аспекты здоровья населения. 2009. № 3. С. 1–11.

4. Крысанов И. С. Введение в фармакоэкономическое моделирование // Фармакоэкономика. Современная фармакоэкономика и фармакоэпидемиология. 2008. № 1. С. 7–9.

5. Развитие здравоохранения Вологодской области на 2014–2020 годы: Государственная программа: постановление Правительства области от 28.10.2013 № 1112. URL: https://vologda-oblast.ru/dokumenty/programmy/23816/

6. Рапаков Г. Г., Банщиков Г. Т. Визуализация показателей в задачах управления здравоохранением Вологодской области // Вузовская наука – региону: материалы XII Всероссийской научно-технической конференции. Вологда: Вологодский государственный технический университет, 2014. С. 61–63.

7. Рапаков Г. Г., Банщиков Г. Т. Интеллектуальный анализ данных в здравоохранении региона (на материалах Вологодской области). Вологда: Вологодский государственный университет, 2014. 79 с.

8. Рапаков Г. Г., Горбунов В. А. Интеллектуальный анализ медико-социологических данных с использованием метода Microsoft Decision Trees // Вестник Воронежского государст-

венного университета. Серия: Системный анализ и информационные технологии. 2015. № 2. С. 130–137.

9. Рапаков Г. Г., Банщиков Г. Т. Организация системы раннего выявления больных артериальной гипертензией и доступность антигипертензивных средств в Вологодской области: опыт использования кластерного анализа // Архивъ внутренней медицины. 2013. № 4. С. 16–23.

10. Рапаков Г. Г., Банщиков Г. Т. Эффективность реализации областной целевой программы лечения пациентов с артериальной гипертензией на региональном уровне (опыт Вологодской области) // Экономические и социальные перемены: факты, тенденции, прогноз. 2014. № 5. С. 206–221.

11. Рапаков Г. Г., Касимов Р. А., Банщиков Г. Т., Горбунов В. А. Распознавание и анализ социокультурных поведенческих паттернов на основе метода ассоциативных правил // Физико-математическое моделирование систем: материалы XII Международного семинара. Воронеж: Воронежский государственный технический университет, 2014. Ч. 2. С. 155–160.

12. Рапаков Г. Г., Касимов Р. А. Методы и алгоритмы машинного обучения при принятии управленческих решений в региональной системе медицинской профилактики (опыт Вологодской области). Вологда: Вологодский государственный университет, 2014. 143 с.

13. Шаханов Н. И., Варфоломеев И. А., Ершов Е. В., Юдина О. В. Прогнозирование отказов оборудования в условиях малого количества поломок // Вестник Череповецкого государственного университета. 2016. № 6 (75). С. 36–41.

14. Harrington P. Machine Learning in Action. Shelter Island: Manning Publications, 2012. 384 p.

15. Witten I. H., Frank E. Data Mining: Practical Machine Learning Tools and Techniques. Burlington: Elsevier Inc., 2005. 525 p.

16. Wu X. The Top Ten Algorithms in Data Mining. Boca Raton: Chapman & Hall, 2009. 201 p.

17. Zhang C., Zhang S. Association Rule Mining: Models and Algorithms. Berlin; Heidelberg: Springer: Verlag, 2002. 248 p.

References

1. Aseev M. G., Diuk V. A. Primenenie sistemy Deep Data Diver dlia resheniia zadachi analiza rynochnykh korzin [Deep Data Diver Application in Market Baskets Analysis]. *Trudy SPIIRAN* [SPIIRAS Proceedings], 2004, vol. 1, no. 2, pp. 127–134.

2. *Vologda – gorod dolgozhitelei*: kontseptsiia aktivnogo dolgoletiia na territorii munitsipal'nogo obrazovaniia “Gorod Vologda” *na period do 2035 goda: reshenie Vologodskoi gorodskoi Dumy ot 29 dekabria 2014 g. № 129.* [Vologda as the city of long-livers: the conception of the active longevity on the territory of the municipal education “Gorod Vologda” for the period up to 2035: the decision of the Vologda’s City Duma dated December the 29th 2014 #129]. Available at: http://vologda-portal.ru/oficialnaya_vologda/index.php?SECTION_ID=8401

3. Zhilina N. M., Fadeeva A. E., Chechenin G. I. Analiz smertnosti naseleniia g. Novokuznetska na osnove elektronnoi bazy dannykh za period 1999–2007 gg. [The death-rates analysis of Novokuznetsk population on the basis of an electronic database for the period 1999–2007]. *Sotsial'nye aspekty zdorov'ia naseleniia* [Social aspects of population health], 2009, no. 3, pp. 1–11.

4. Krysanov I. S. Vvedenie v farmakoekonomicheskoe modelirovanie [Introduction into pharmacoeconomic modelling]. *Farmakoekonomika. Sovremennaia farmakoekonomika i farmakoepidemiologiia* [Pharmacoeconomics. Modern pharmacoeconomics and pharmacoepidemiology], 2008, no. 1, pp. 7–9.

5. *Razvitie zdravookhraneniia Vologodskoi oblasti na 2014–2020 gody*: *Gosudarstvennaia programma: postanovlenie Pravitel'stva oblasti ot 28.10.2013 № 1112*. [The development of Vologda region health care system for 2014-2020: The State Program: the act of the regional government dated 28.10.2013 # 1112]. Available at: https://vologda-oblast.ru/dokumenty/programmy/23816/

6. Rapakov G. G., Banshchikov G. T. Vizualizatsiia pokazatelei v zadachakh upravleniia zdravookhraneniem Vologodskoi oblasti [Visualization of indicators in health management tasks in the Vologda Region]. *Vuzovskaia nauka – regionu: materialy dvenadtsatoi vserossiiskoi nauchno –*

tekhnicheskoi konferentsii [University science – region: proceedings of the 12th all-Russian scientific and technical conference VNR-2014]. Vologda: Vologda Technical State University, 2014, pp. 61–63.

7. Rapakov G. G., Banshchikov G. T. *Intellektual'nyi analiz dannykh v zdravookhranenii regiona (na materialakh Vologodskoi oblasti)* [Data mining in region public health (the material of the Vologda region)]. Vologda: Vologda State University, 2014. 79 p.

8. Rapakov G. G., Gorbunov V. A. Intellektual'nyi analiz mediko-sotsiologicheskikh dannykh s ispol'zovaniem metoda Microsoft Decision Trees [Intelligent analysis of medical-sociological data using microsoft decision trees algorithm]. *Vestnik Voronezhskogo gosudarstvennogo universiteta. Seriia: Sistemnyi analiz i informatsionnye tekhnologii* [Proceedings of Voronezh State University. Series: Systems analysis and information technologies], 2015, no. 2, pp. 130–137.

9. Rapakov G. G., Banshchikov G. T. Organizatsiia sistemy rannego vyiavleniia bol'nykh arterial'noi gipertenziei i dostupnost' antigipertenzivnykh sredstv v Vologodskoi oblasti: opyt ispol'zovaniia klasternogo analiza [Organization of early detection system for patients with arterial hypertension and availability of antihypertensive drugs in the Vologda region: experience of cluster analysis]. *Arkhiv" vnutrennei meditsiny* [The Russian Archives of Internal Medicine], 2013, no. 4, pp. 16–23.

10. Rapakov G. G., Banshchikov G. T. Effektivnost' realizatsii oblastnoi tselevoi programmy lecheniia patsientov s arterial'noi gipertenziei na regional'nom urovne (opyt Vologodskoi oblasti) [Efficiency of implementation of the regional target program for the treatment of patients with arterial hypertension at the regional level (experience of the Vologda Oblast)]. *Ekonomicheskie i sotsial'nye peremeny: fakty, tendentsii, prognoz* [Economic and Social Changes: Facts, Trends, Forecast], 2014, no. 5, pp. 206–221.

11. Rapakov G. G., Kasimov R. A., Banshchikov G. T., Gorbunov V. A. Raspoznavanie i analiz sotsiokul'turnykh povedencheskikh patternov na osnove metoda assotsiativnykh pravil [Recognition and analysis of behavioral sociocultural patterns by association rules technique]. *Fiziko-matematicheskoe modelirovanie sistem: materialy XII mezhdunarodnogo seminara* [Physico- mathematical system modeling: XII international seminar materials]. Voronezh: Voronezh State Technical University, 2014, part 2, pp. 155–160.

12. Rapakov G. G., Kasimov R. A. *Metody i algoritmy mashinnogo obucheniia pri priniatii upravlencheskikh reshenii v regional'noi sisteme meditsinskoi profilaktiki (opyt Vologodskoi oblasti*) [Methods and algorithms of machine learning for decision-making in the regional system of medical prevention (the experience of the Vologda region)]. Vologda: Vologda State University, 2014. 143 p.

13. Shakhanov N. I., Varfolomeev I. A., Ershov E. V., Iudina O. V. Prognozirovanie otkazov oborudovaniia v usloviiakh malogo kolichestva polomok [Forecasting equipment failures at the conditions of a small number of breakdowns]. *Vestnik Cherepovetskovo gosudarstvennovo universiteta* [Cherepoves State University Bulletin], 2016, no. 6 (75), pp. 36–41.

14. Harrington P. *Machine Learning in Action*. Shelter Island: Manning Publications, 2012. 384 p.

15. Witten I. H., Frank E. *Data Mining: Practical Machine Learning Tools and Techniques*. Burlington: Elsevier Inc, 2005. 525 p.

16. Wu X. *The Top Ten Algorithms in Data Mining*. Boca Raton: Chapman & Hall, 2009. 201 p.

17. Zhang C., Zhang S. *Association Rule Mining: Models and Algorithms*. Berlin; Heidelberg: Springer: Verlag, 2002. 248 p.

Для цитирования: *Рапаков Г. Г., Банщиков Г. Т., Горбунов В. А., Малыгин Л. Л., Ревелев И. М.* Анализ данных медико-социологического мониторинга на основе методов машинного обучения // Вестник Череповецкого государственного университета. 2019. № 1 (88). С. 27–38. DOI: 10.23859/1994-0637-2019-1-88-3

For citation*:* Rapakov G. G., Banshchikov G. T., Gorbunov V. A., Malygin L. L.., Revelev I. M. Data analysis of medical-sociological monitoring with machine learning methods. *Bulletin of the Cherepovets State University*, 2019, no. 1 (88), pp. 27–38. DOI: 10.23859/1994-0637-2019-1-88-3