УДК 303.732.4

ИНТЕЛЛЕКТУАЛЬНЫЙ АНАЛИЗ МЕДИКО-СОЦИОЛОГИЧЕСКИХ ДАННЫХ С ИСПОЛЬЗОВАНИЕМ МЕТОДА MICROSOFT DECISION TREES

Г. Г. Рапаков, В. А. Горбунов

Вологодский государственный университет

Поступила в редакцию 10.02.2015 г.

Аннотация. В статье рассмотрена реализация интеллектуальной поддержки принятия управленческих решений в рамках областной целевой программы «Профилактика и лечение артериальной гипертензии среди населения Вологодской области». На основе медико-социологических показателей была выполнена оценка значимых факторов социально-экономической эффективности медицинской помощи с использованием алгоритма дерева принятия решений в реализации Microsoft Decision Trees.

Ключевые слова: Data Mining, интеллектуальный анализ данных, Microsoft Business Intelligence, поддержка принятия решений, социологические исследования, профилактика, артериальная гипертензия.

Annotation. To study influence of pharmacotherapy profile on quality of arterial hypertension control in patients with cardiovascular diseases Microsoft Decision Trees algorithm was used. There are some results and the way these results can be used is shown.

Keywords: data mining, intelligent analysis, Microsoft Business Intelligence, decision support, sociological research, prophylaxis, arterial hypertension.

ВВЕДЕНИЕ

В Вологодской области в настоящее время разработана и внедрена технология раннего выявления больных артериальной гипертензией (АГ) врачами территориальных поликлиник, действующая в рамках Областной целевой (в дальнейшем – ведомственная) программы (ОЦП) [1–3]. Отметим, что в структуре общей смертности населения региона $\approx 55\,\%$ занимают болезни системы кровообращения (БСК), а важнейшим фактором риска развития БСК является артериальная гипертензия. Данные российских стандартизованных эпидемиологических исследований показывают, что распространенность АГ составляет около 40 %. В 2010–2011 гг. первичная заболеваемость АГ населения региона изменялась от 622 до 764 (на 100 тыс. чел.). Географическое распределение показателя заболеваемости АГ для муниципальных образований Вологодской области в динамике изменения за 2008–2010 гг. представлено на рис. 1. Показатели смертности от острого инфаркта миокарда в регионе находились на уровне средних по РФ, а от мозгового инсульта – превышали их. Оценка величины упущенной выгоды для Вологодской области за период 2009–2010 гг., вызванной болезнями системы кровообращения, в связи с временной нетрудоспособностью и выплатами по социальному страхованию, пенсий по инвалидности и преждевременной смертностью трудоспособного населения составляет 1,5 млрд. руб. в год [4].

В связи с вышесказанным, актуальной является задача аналитической поддержки принятия управленческих решений в сфере региональной профилактики неинфекционных заболеваний, выявления и своевременного лечения больных АГ, предупреждения инвалидности и смертности населения [5–11]. Важность и практическая значимость работы обусловлена выработкой рекомендаций для

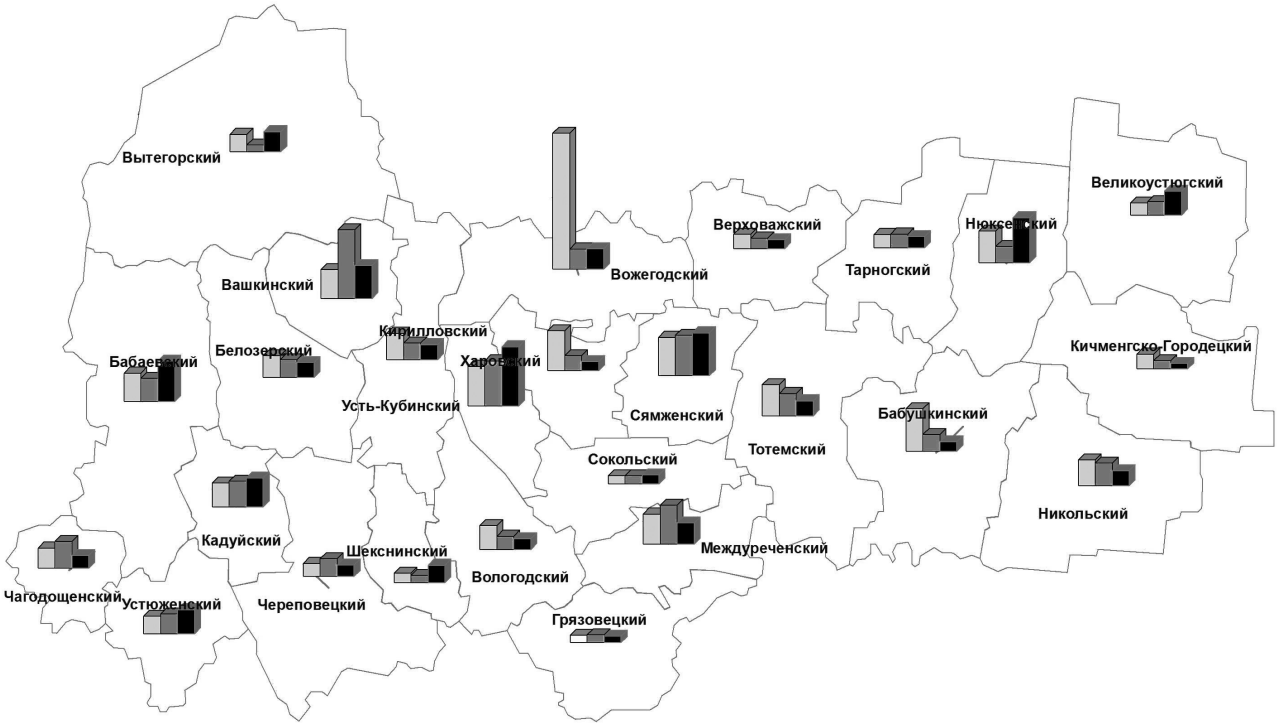

Рис. 1. Заболеваемость артериальной гипертензией населения Вологодской области в 2008–2010 гг. (на 100 тыс. чел.)

лиц, принимающих решения в области организации здравоохранения и медицинской профилактики, которые способствовали бы снижению социально–экономического бремени БСК.

Цель настоящего исследования: анализ значимых факторов социально-экономической эффективности медицинской помощи в ходе реализации Областной целевой программы «Профилактика и лечение артериальной гипертензии среди населения Вологодской области» на основе применения методов интеллектуального анализа данных. Его новизна состоит в оценке возможностей алгоритма дерева решений для формирования набора решающих правил при информационно-аналитической поддержке управленческих решений в региональном здравоохранении.

РЕЗУЛЬТАТЫ ИССЛЕДОВАНИЙ

Департаментом здравоохранения Вологодской области проведен медико-социологический мониторинг на основе данных карт для проведения экспертной оценки эффективности лечения пациента с АГ (данные амбулаторной карты больного, форма №025у) [12]. Объем выборки обеспечивает необходимую точность оценки в пределах 0,05 с доверительной вероятностью $\alpha = 0,95$. Результаты анализа показали, что эффективная антигипертензивная терапия отмечена для 82,84 % пациентов, для 16,27 % эффективность диспансеризации является недостаточной, а для 0,89 % – лечение неэффективно.

Особенностью работы с большими массивами сложно структурированных данных являются существенные трудности при формировании рабочей гипотезы. Распространенным подходом при проведении исследований является извлечение или добыча данных. Как известно, методы Data Mining базируются на статистических методах анализа, методах искусственного интеллекта и используют СУБД, они позволяют выявлять новые и нетривиальные знания, доступные для интерпретации и полезные при принятии управленческих решений [13–15].

Изучение и анализ источников позволил выполнить сопоставление целей и методов исследования с литературными данными. Вопросы заполнения базы знаний логическими правилами в виде продукций в практике использования программного пакета See 5/C 5.0

при формировании базы знаний медицинской информационной системы прогнозирования исхода беременности на основе технологии построения деревьев решений рассматриваются в работе [16]. Решению проблемы выявления скрытых взаимосвязей между диагнозами пациентов, их учетными данными и применяемыми лекарственными препаратами на основе сведений, содержащихся в хранилище данных клиники, реализованном на СУБД Microsoft SQL Server, посвящена статья [17]. Советующая подсистема диагностики заболеваний и оценки здоровья пациентов по результатам биохимических исследований на основе продукционных правил разработана в составе медицинской информационной системы электронной истории болезни [18].

Обработка результатов социометрии выполнялась авторами работы с применением аналитических методов исследований – алгоритма дерева принятия решений в реализации фирмы «Майкрософт» (Microsoft Decision Trees) [19]. Достоинства метода деревьев решений: наглядная модель классификации, доступная для интерпретации, как лицом, принимающим решение, так и экспертом; создание правил в случаях, когда эксперт испытывает трудности с формализацией знаний; представление правил на естественном языке; достаточная точность прогноза даже на малых выборках и др. Использование деревьев решений ограничено их неспособностью находить наилучшие, т. е. наиболее полные и точные правила [14, 15].

Интеллектуальный анализ данных (ИАД) осуществлялся при помощи средств бизнес-аналитики СУБД Microsoft SQL Server 2012 на платформе Microsoft Business Intelligence (рис. 2). Предварительно была выполнена предобработка, очистка и трансформация данных. Модель ИАД содержит собственные метаданные, собранную статистику и выявленные, на основе использования алгоритма, закономерности.

При решении задачи классификации с использованием иерархической древовидной структуры в качестве категориального дискретного выходного атрибута рассматривался показатель оценки эффективности лечения пациентов с АГ. Настройка алгоритма выполнялась с использованием параметров модели деревьев решений.

Результатом работы алгоритма 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. Результат работы алгоритма зачастую представляет собой сложное дерево с большим количеством узлов и ветвей, не пригодное для интерпретации. Дерево с избыточным количеством ветвей может удачно классифицировать обучающие данные, но дает невысокую точность прогноза для новых сведений. Ценность правила становится меньше с уменьшением количества объектов, для которых оно справедливо. С практической точки зрения предпочтительным является такой результат разбиения, при котором малому количеству узлов соответствует большое количество объектов. Для ограничения глубины дерева проводят оценку целесообразности его дальнейшего разбиения. 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 эта операция не приведет к возрастанию ошибки распознавания.

В целях предотвращения эффекта переобучения в исследовании была использована обрезка дерева и подавление роста со значением 0,9 (параметр COMPLEXITY_PENALTY). Вычисление коэффициента разбиения при росте дерева выполнялось по методу Entropy (SCORE_METHOD). При разбиении узла службы Analysis Services самостоятельно определяли наилучшее разбиение из двух вариантов – бинарное или полное (SPLIT_METHOD). Для остальных параметров использовались значения по умолчанию.

Полученная модель корректно отражает особенности области исследований и согласуется с экспертной оценкой. Диаграмма

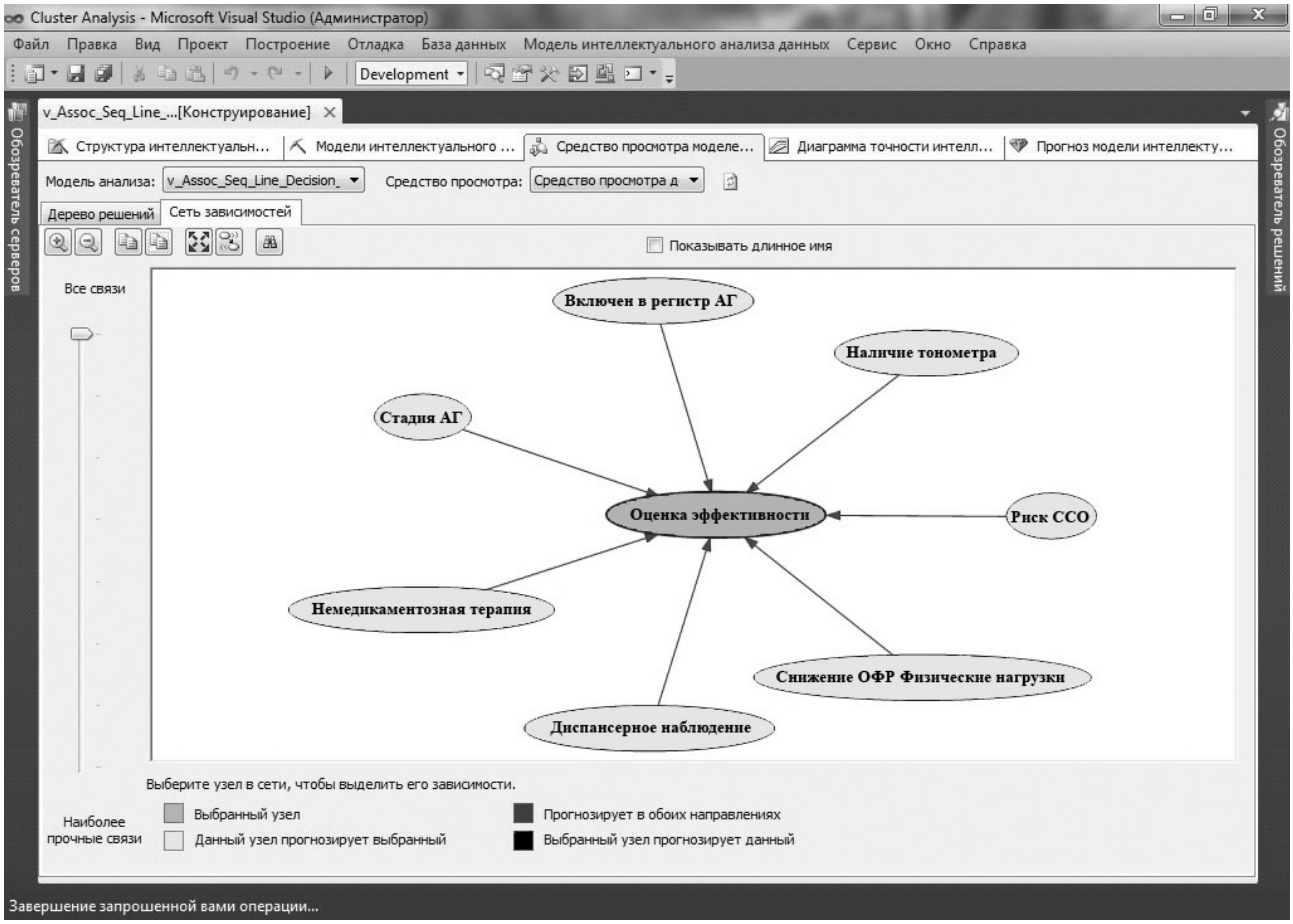

Рис. 2. Диаграмма сети зависимостей

сети зависимостей отображает выявленные взаимосвязи. Для узла целевого показателя представлены все влияющие на него параметры (рис. 2). Варьируя силу связи между входными атрибутами и прогнозируемым показателем модели можно выделить наиболее прочные зависимости. Идентифицировав параметры, не влияющие на результирующую модель ИАД можно уменьшить количество столбцов в исходном наборе данных. Выбор узла на диаграмме дерева решений позволяет ознакомиться с прогнозом алгоритма для соответствующей комбинации параметров модели. Для точки ветвления указывается количество соответствующих случаев в обучающем наборе. Выделив узел, можно получить его описание с гистограммой распределения состояний прогнозируемого атрибута (рис. 3). Интуитивная ясность деревьев решений, которые не содержат неоправданно большого числа ветвей, позволяет наглядно интерпретировать описанные ими зависимости с использованием правил «If – Then».

При помощи средства анализа ключевых факторов влияния был выполнен анализ закономерностей, содержащихся в данных, для целевого показателя – эффективности лечения пациента с АГ. При этом была создана структура ИАД, содержащая ключевые сведения о них; на основе упрощенного алгоритма Байеса (Майкрософт) построена модель ИАД; выполнены прогнозы по столбцам данных и сравнение с результатом; для определения факторов, оказывающих наибольшее влияние на результат, был использован показатель оценки достоверности.

В результате были построены таблицы данных, содержащие факторы, связанные с тремя дискретными значениями прогнозируемого столбца, и графически отображены вероятностные связи между ними.

По результатам проведенного исследования можно сделать следующий вывод. В целях обеспечения роста показателя достаточной эффективности лечения пациента АГ с 82,84 % до 98,67 % с учетом ключевых факто-

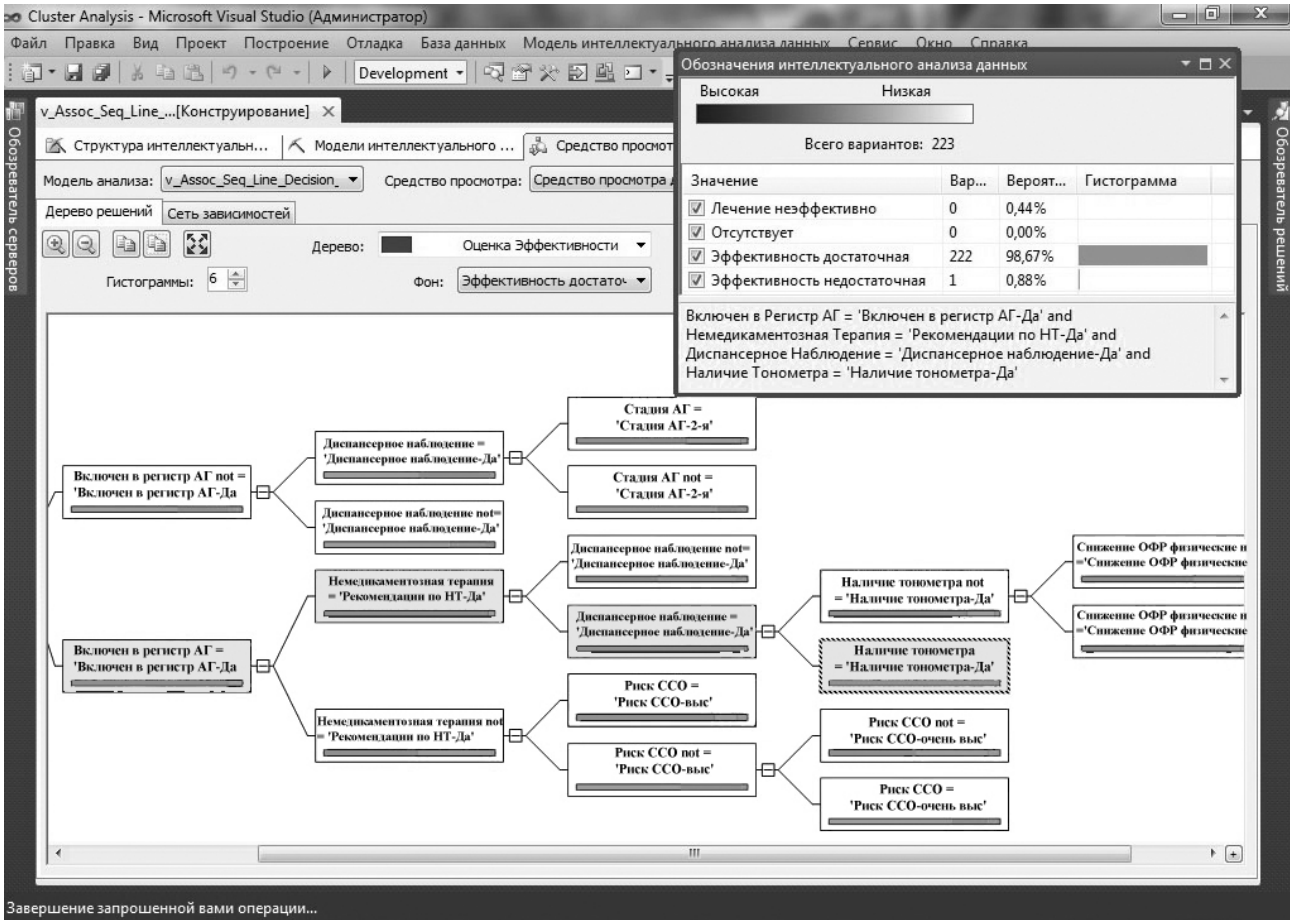

Рис.3 . Диаграмма дерева решений

ров, в наибольшей степени влияющих на достаточную эффективность лечения, необходимо обеспечить:

– включение пациента в регистр АГ (100 %);

– нахождение больного под диспансерным наблюдением (84 %) (сравнительным по степени влияния фактором является отсутствие инвалидности по АГ и ассоциированным клиническим состояниям);

– доведение сведений до пациента и учет им рекомендаций по немедикаментозной терапии (71 %);

– наличие и самоконтроль артериального давления больным при помощи тонометра (53 %).

Для снижения социально-экономического ущерба, значительная часть которого обусловлена непрямыми экономическими потерями в связи с преждевременной смертностью мужчин трудоспособного возраста, необходимо продолжение инвестирования в активную профилактику основных факторов риска.

Верификация модели ИАД позволяет выполнить проверку достоверности добытого знания. Точность классификации прогностической модели была определена при помощи матрицы классификации. Результаты показывают, что 91 % значений достаточной эффективности лечения, предсказанных моделью, соответствуют фактическим значениям. Точность классификации случаев неэффективного лечения максимальна, что обусловлено их редкой встречаемостью. В составе тестовой выборки присутствовал один такой эпизод, который и был успешно опознан. Наибольшую трудность у построенной модели вызывает оценка случаев, когда эффективность является недостаточной. Успех здесь не превышает 53 %. Перспективным является исследование альтернативных подходов анализа данных и использование специальной подготовки обучающей выборки для улучшения точности прогноза. В целом модель ИАД обеспечивает прогностическую точность в 83 %, что вполне достаточно для большинства практических приложений метода.

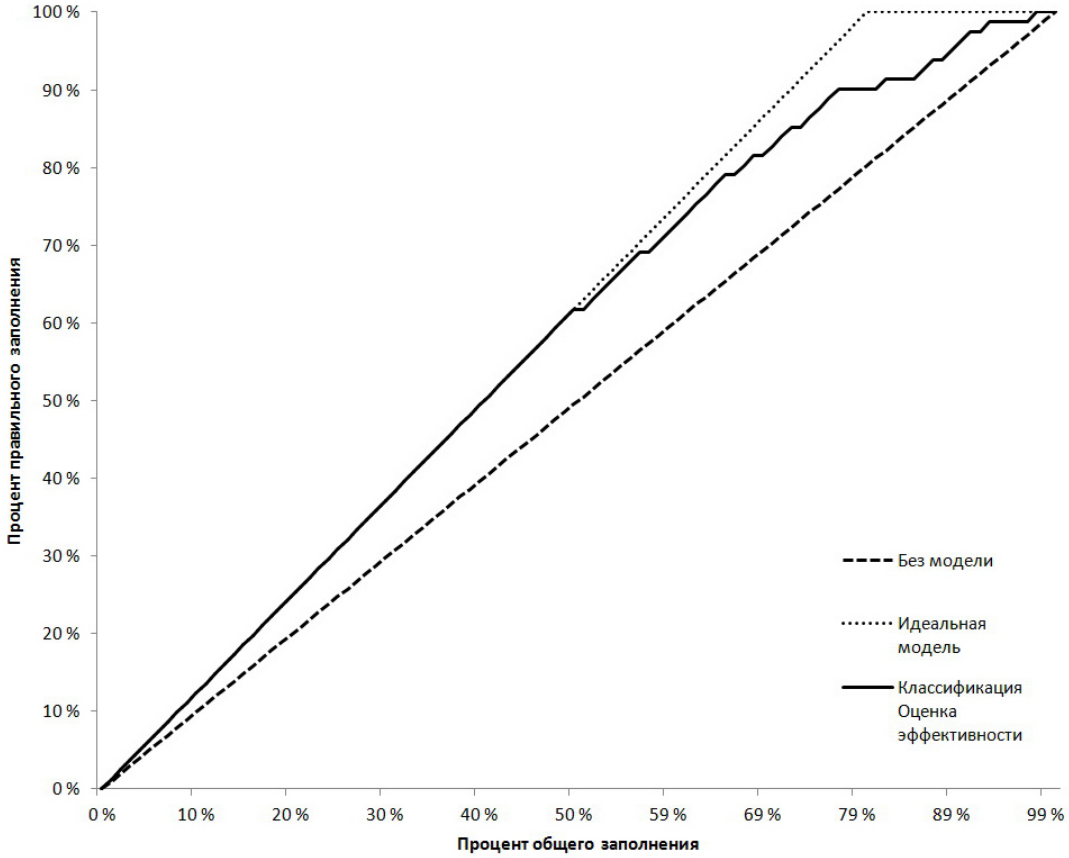

Рис. 4. Диаграмма точности прогностической модели

Наглядное представление о производительности модели ИАД можно получить при помощи диаграммы точности (рис. 4). На графика верхняя ломаная линия соответствует гипотетической идеальной модели, которая выполняет безошибочную классификацию. Нижняя наклонная прямая – биссектриса координатного угла является результатом случайного выбора. Классификация оценки эффективности лечения на основе прогностической модели представлена сплошной кривой линией, которая размещается между ними.

Диаграмма демонстрирует, что количество правильных прогнозов возрастает с увеличением числа случаев, анализируемых моделями. Степень близости кривой оцениваемой модели ИАД к ломаной линии идеальной модели позволяет визуально проверить точность прогноза. Качество классификации начинает уменьшаться от максимального возможного после 50 % от общего объема данных. Снижение продолжается до 80 % от числа всех записей. Разность прогнозов моделей составляет при этом 9,8 %.

ЗАКЛЮЧЕНИЕ

Реализация алгоритма дерева принятия решений MS Decision Trees на базе технологий MS SQL Server 2012 обеспечила аналитическую поддержку при отборе значимых факторов качества медицинской помощи, оказываемой в рамках ОЦП. Полученная модель корректно отображает особенности предметной области.

В ходе работы алгоритма были выявлены системные связи и закономерности в разнородных исходных данных между независимыми показателями и прогнозируемым параметром с приемлемой на практике точностью 83 %. Падение качества прогноза по отношению к идеальной модели не превышает 9,8 %.

Для показателя оценки эффективности лечения пациентов с АГ выделен набор из 4 правил, совместная интерпретация которых позволяет обеспечить рост целевого параметра в 1,19 раза. Закономерности выражены компактно в символьной форме, описываемой системой логических функций и понятны для лица, принимающего решение.

СПИСОК ЛИТЕРАТУРЫ

1. Профилактика и лечение артериальной гипертонии среди населения Вологодской области на 2009–2011 годы [Электронный ресурс]: ведомственная целевая программа: постановление Правительства Вологодской области от 28 июня 2010 г. № 739 // КонсультантПлюс: справ.-правовая система / Компания «Консультант-Плюс»

2. Профилактика и лечение артериальной гипертонии и атеросклероза среди населения Вологодской области на 1998–2002 годы [Электронный ресурс]: областная целевая программа: постановление Законодательного Собрания от 18.03.98. № 97 // КонсультантПлюс: справ.-правовая система / Компания «КонсультантПлюс».

3. Профилактика и лечение артериальной гипертонии в Российской Федерации [Электронный ресурс]: Федеральная целевая программа: постановление Правительства РФ от 17 июля 2001 г. № 540 (128) // КонсультантПлюс: справ.-правовая система / Компания «КонсультантПлюс».

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

5. Реализация программы Профилактика и лечение артериальной гипертонии в Российской Федерации на региональном уровне (опыт г. Вологды) / Г. Т. Банщиков, А. А. Колинько, А. И. Попугаев [и др.] // Кардиоваскулярная терапия и профилактика. – 2004. – № 3. – С. 43–46.

6. Оценка качества контроля артериальной гипертонии среди населения Вологодской области / А. И. Попугаев, Г. Т. Банщиков, Р. А. Касимов и [др.] // Рациональная фармакотерапия в кардиологии. – 2008. – Т. 4, № 5. – С. 6–10.

7. *Попугаев А. И.* Регистр артериальной гипертонии в Вологодской области / А. И. Попугаев, Д. А. Рыбаков, Г. Т. Банщиков // Кардиоваскулярная терапия и профилактика. – 2008. – № 2. – С.18–21.

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

9. *Рапаков Г. Г.* Сравнительная оценка эффективности методов классификационного анализа в социологических исследованиях / Г. Г. Рапаков, В. А. Горбунов // Вестник Воронежского ун-та. Серия: Системный анализ и информационные технологии. – 2014. – № 4. – С. 54–62.

10. *Рапаков Г. Г.* Интеллектуальный анализ данных в здравоохранении региона (на материалах Вологодской области): монография / Г. Г. Рапаков, Г. Т. Банщиков. – Вологда : ВоГУ, 2014. – 79 с.

11. *Рапаков Г. Г.* Методы и алгоритмы машинного обучения при принятии управленческих решений в региональной системе медицинской профилактики (опыт Вологодской области): монография / Г. Г. Рапаков, Р. А. Касимов. – Вологда : ВоГУ, 2014. – 143 с.

12. *Решетников А. В.* Медико-социологический подход к исследованию качества медицинской помощи / А. В. Решетников, М. М. Астафьев // Социология медицины. – 2005. – № 1. – С. 32–36.

13. Технологии анализа данных: Data Mining, Visual Mining, Text Mining, OLAP / А. А. Барсегян, М. С. Куприянов, В. В. Степаненко, И. И. Холод. – СПб. : БХВ-Петербург, 2007. – 384 с.

14. *Дюк В.* Data Mining / В. Дюк, А. Самойленко. – СПб. : Питер, 2001. – 368 с.

15. *Дюк В.* Информационные технологии в медико–биологических исследованиях / В. Дюк, В. Эммануэль. – СПб. : Питер, 2003. – 528 с.

16. Технология формирования баз знаний в медицинских информационных системах / О. Г. Берестнева, К. А. Шаропин, А. В. Старикова, Л. И. Кабанова // Известия Южного федерального университета. Технические науки. – 2010. – № 8. – Том 109. – С. 32–37.

17. *Мутина Е. И.* Применение многомерного интеллектуального анализа данных в социологических исследованиях / Е. И. Мутина // Вестник Московского городского педагогического университета. Серия: Естественные науки. – 2010. – № 1. – С. 66–71.

18. *Петухов А. С.* Модели и методы вывода в многоуровневой компонентной советующей подсистеме в составе электронной истории болезни / А. С. Петухов, Е. А. Оленников, А. А. Захаров // Информационные системы и технологии. – 2008. – № 1–2. – С. 153–158.

19. *MacLenna J.* Data Mining with Microsoft SQL Server 2008 / J. MacLennan, Z. Tang, B. Crivat. – Wiley Publishing, Inc., 2009. – 744 p.

Рапаков Георгий Германович – кандидат технических наук, доцент кафедры информационных систем и технологий, Вологодский государственный университет.
Тел.: +7(8172) 72-95-71
E-mail: grapakov@yandex.ru

Rapakov G. G. – PhD in Technical Science, Associate Professor, Information Systems and Technologies Department, Vologda State University

Горбунов Вячеслав Алексеевич – доктор физико-математических наук, профессор, зав. кафедрой информационных систем и технологий, Вологодский государственный университет.
Тел.: +7(8172)72-95-71
E-mail: gorbunov1945@inbox.ru

Gorbunov V. A. – Doctor of Physics-Mathematical Science, Professor, Information Systems and Technologies Department, Head of Department, Vologda State University