Мельник К.В., Глушко В.Н.

Применение аппарата Байесовых сетей при обработке данных из медицинских карточек

Мельник Карина Владимировна, ассистент,
Глушко Виталий Николаевич, ассистент
Национальный технический университет "Харьковский политехнический институт", г. Харьков, Украина

Аннотация. Медицинские информационные технологии представляют собой незаменимый инструмент для решения различного рода медицинских задач, чему посвящено большое количество публикаций. На сегодняшний день медицинские учреждения стали переходить на электронный документооборот, это обусловлено удобным хранением информации о пациенте, легким и быстрым доступом к информации, а также повышением эффективности работы медицинского учреждения при обработке медицинских данных. Использование статистической информации из электронной базы пациентов помогает выявить разнообразные факторы риска развития различных заболеваний на обслуживаемой территории. Зачастую пациенты обращаются не регулярно в медицинские учреждения, поэтому процесс постановки диагноза представляет собой задачу классификации на неполной информации. Подходящим математическим аппаратом для решения подобных задач является аппарат байесовых сетей доверия. В статье приведен обзор применения байесовых сетей доверия при проведении процесса диагностирования различных заболеваний. Анализ практических реализаций байесовых сетей показал целесообразность данного подхода для решения задач диагностирования. Поэтому в данной работе для определения рисков возможных заболеваний при проведении процесса раннего диагностирования в условиях неопределенной и неполной информации, полученной из карточек пациентов, предлагается применять байесовы сети доверия. В статье предложена сеть, которая помогает обнаружить нарушения работы сердечнососудистой системы. Информация, используемая для построения байесовой сети, была разбита на три группы: первая группа – это факторы риска, влияющие на возникновение и развитие определенных заболеваний; вторая группа – это совокупность рассматриваемых диагнозов; третья группа представляет собой наблюдения врача, различные симптомы и результаты лабораторных или приборно-компьютерных анализов. В статье приведен пример использования разработанной байесовой сети.

***Ключевые слова**: медицинские информационные технологии, байесовы сети доверия, обработка медицинских данных, сердечнососудистая система.*

Введение. На сегодняшний день медицинские информационные технологии (МИТ) представляют собой незаменимый инструмент для решения различного рода медицинских задач. Например, с помощью МИТ можно эффективно выполнять такие задачи, как постановка диагноза, обработка лабораторных анализов, осуществление доступа к специальной медицинской литературе или медицинской карточке пациента. Большое количество публикаций посвящено решению подобных задач. Подробный обзор применения МИТ в различных сферах медицины приведен в работе [1].

На сегодняшний день медицинские учреждения стали переходить на электронный документооборот. Многие больницы, клиники, медпункты стали использовать электронные медицинские карточки, это обусловлено удобным хранением информации о пациенте, легким и быстрым доступом, а также повышением эффективности работы медицинского учреждения при обработке этих данных. Вопросы обработки медицинских данных из электронных карточек пациентов в медицинских учреждениях с применением разных подходов более подробно освещены, например, в работах [2, 3]. Истории болезней, результаты различных медицинских процедур, дата следующего профосмотра - эта и многая другая информация хранится в электронной карточке. Использование статистической информации из электронной базы пациентов помогает выявить разнообразные факторы риска развития различных заболеваний на обслуживаемой территории. Эта информация, в свою очередь, может использоваться при обработке данных каждого пациента в отдельности для задач раннего диагностирования.

Вопросы обработки медицинских данных. В основном обработка информации из электронных карточек пациентов связана с процессом диагностирования состояния пациента. Этот процесс можно рассмотреть для двух постановок задачи. В первом случае рассматривается классическая задача постановки диагноза при наличии полной информации о пациенте (перечень симптомов, результаты лабораторных анализов, результаты приборно-компьютерных процедур), решения для которой предлагает большое количество авторов, что отражено, например, в работе [2]. Вторая постановка рассматриваемой задачи представляет больший интерес. В этом случае рассматривается задача классификации на неполной информации. Подходящим математическим аппаратом для решения подобных задач является аппарат Байесовых сетей доверия (БСД или БС) [4].

Анализ разных источников информации показал, что существуют различные успешные реализации байесовых сетей в медицине, например, Alarm [5] – медицинское диагностическое приложение для наблюдения за пациентом, реализованное на основе БС объемом в 37 узлов. Сеть выдает 8 диагнозов и использует 16 различных симптомов, выдавая при этом пользователю специфические текстовые сообщения о возможных проблемах в организме пациента. БС Asia или Lung Cancer [6], состоящая из 8 узлов, диагностирует наличие рака легких или туберкулеза в зависимости от некоторых факторов риска. БСД Diabetes [7], состоящая из 413 узлов, применяется для коррекции дозы инсулина в зависимости от состояния пациента. БС Pathfinder [8] состоит из 135 узлов, применяется как аппарат для определения заболеваний лимфатической системы пациента. Сеть Munin [9] используется для определения нервно-мышечных заболеваний. Данная сеть существует в нескольких вариациях: в виде одной полной сети (размер – 1041 узлов), а также в виде совокупности 4 подсетей (3 подсети размером около 1000 узлов, четвертая – 186 узлов).

БС Hepar размером в 70 узлов предложена в работе [10] для диагностирования заболеваний печени. В данной сети рассматривается 16 заболеваний, остальные узлы – это различные характеристики, например, симптомы или результаты лабораторных исследований.

Анализ практических реализаций БС показал целесообразность данного подхода для решения задач диагностирования. Поэтому в данной работе для определения рисков возможных заболеваний при проведении процесса раннего диагностирования в условиях неопределенной и неполной информации, полученной из карточек пациентов, предлагается применять БСД.

Использование байесовой сети доверия при проведении раннего диагностирования. БСД представляют собой вероятностный ориентированный граф, в вершинах которого находятся различные понятия, а ребра отображают условную зависимость одной вершины от другой. Применительно к диагностике заболеваний в роли вершин могут выступать вредные привычки, нездоровый образ жизни, неблагоприятная наследственность, заболевания и их симптомы.

Для эффективной оценки состояния пациента необходимо использовать различные байесовы сети, каждая из которых направлена на определенную группу заболеваний. Например, для диагностики нарушений работы сердечнососудистой системы (ССС) необходимо использовать сеть, которая определяет риски заболеваний ССС; для обнаружения возможных заболеваний дыхательной системы необходимо использовать сеть, которая отражает информацию о дыхательной системе.

Рассмотрим в работе БСД, которая помогает обнаружить риски заболеваний ССС. Для оценивания состояния сердца пациента необходимо использовать анамнез, который представлен в карточке. Вся доступная информация разбивается на три группы: первая группа – это факторы риска, влияющие на возникновение и развитие определенных заболеваний; вторая группа – это совокупность рассматриваемых диагнозов; третья группа представляет собой наблюдения врача, различные симптомы и результаты лабораторных или приборно-компьютерных анализов.

В качестве диагнозов были выбраны четыре заболевания ССС, которые имеют между собой весьма сильную связь: атеросклероз, артериальная гипертензия (АГ), стенокардия, ишемическая болезнь сердца (ИБС). Атеросклероз представляет собой болезнь кровеносных сосудов, которая проявляется в том, что на внутренней поверхности сосуда образовываются наросты, так называемые бляшки, которые затрудняют течение крови. Основными симптомами служат повышенное давление, увеличение холестерина, а также нарушения сердечного ритма. В свою очередь, давление, неправильный сердечный ритм и боль в груди являются симптомами для стенокардии; повышенное давление, головная боль и бессонница – это симптомы АГ; а для ИБС являются симптомами все вышеперечисленные. В качестве симптомов были выбраны следующие характерные показатели работы сердечнососудистой системы: боли в области сердца, изменения на электрокардиограмме, повышенное артериальное давление, маркер биохимического анализа крови – уровень холестерина, результаты клинического анализа крови. В роли факторов риска были выбраны разнообразные характеристики жизни и привычек пациентов, которые не являются причиной этих заболеваний, но способствуют их развитию. Для заболеваний ССС существуют следующие факторы риска: курение, злоупотребление алкоголем, ожирение, гиподинамия – отсутствие физической нагрузки, наличие диабета, стресс.

В результате анализа информации о работе сердечнососудистой системы была БСД, которая представлена на рисунке 1. Структура сети имеет следующие характеристики: часть узлов, которые отвечают за наличие определенных факторов риска в жизни пациента, представляют собой независимые вершины; некоторые вершины имеют большое количество родительских вершин – вершины, отвечающие за диагнозы и их симптомы; часть вершин не имеет потомков, но имеет предков – вершины-симптомы.

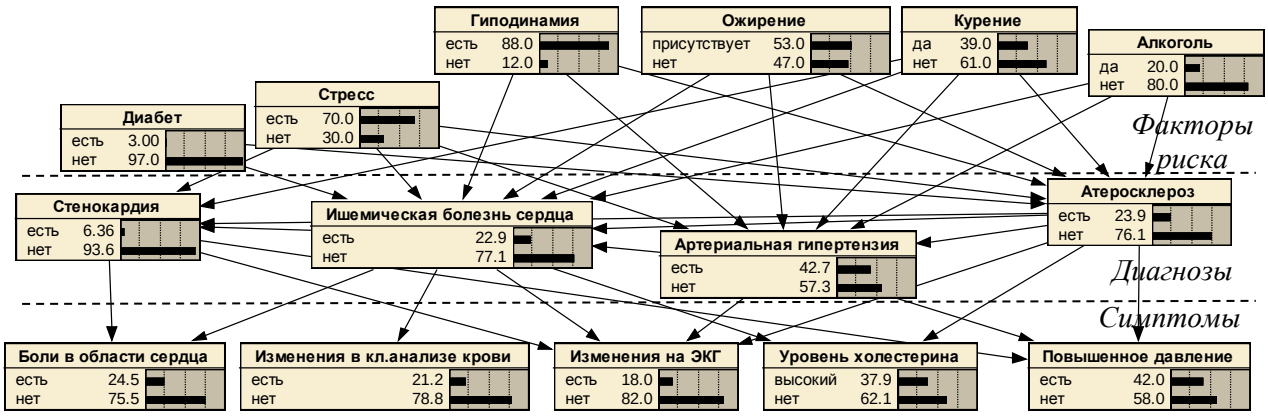

Рисунок 1. Структура БСД

Начальные параметры рассматриваемой сети заданы экспертами с помощью программного средства Netica, поэтому сеть готова к использованию для принятия различных медицинских решений по обработке имеющихся данных. Фрагмент таблицы вероятностей для узла под названием "атеросклероз" представлен на рисунке 2.

Для уменьшения субъективизма и увеличения достоверности принимаемых решений необходимо сеть пополнять новыми медицинскими наблюдениями и

данными, то есть проводить процесс обучения. Существует два различных подхода к обучению БС. Первый подход приводит к изменению топологии сети, так как при построении начальной версии сети эксперт мог не учесть некоторые латентные переменные, которые существенно влияют на результаты исследований. Методы обучения второго подхода позволяют скорректировать значения условных вероятностей в узлах рассматриваемой сети. Если медицинские данные, которые подаются на обучение, не имеют пробелов, то тогда в качестве метода обучения используется классический байесовский подход к вычислению условных вероятностей, основываясь на теореме Байеса; в противном случае рекомендуется применять алгоритм максимизации математического ожидания (expectation maximization) [4].

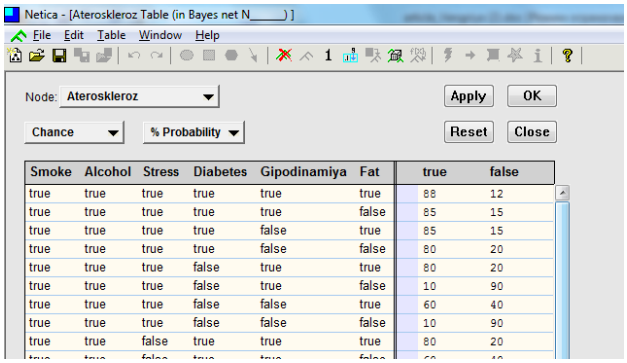

Рисунок 2. Таблица вероятностей для атеросклероза

Алгоритм максимизации математического ожидания основывается на том факте, что наблюдаемое событие заменяется на ожидаемое количество выполнения события. Это происходит следующим образом: если данные, с помощью которых необходимо обучить сеть, имеют пробелы, то пробелы необходимо заполнить возможностями появления того или иного события. Например, причиной повышенного давления могут служить атеросклероз, артериальная гипертензия и стенокардия, а в наборе данных для сети нет информации о стенокардии, тогда одна строчка данных заменяется на два случая: когда при этих исходных условиях есть стенокардия и когда она отсутствует. После рассчитывается ожидаемое количество выполнения события по формуле, $N(x) = \sum_k I(x / D(k))$ где $I(x / D(k)$ - функция индикатор, которая равна единице, если событие x происходит в обучающем процессе $D(k)$, нулю – если не происходит. Итоговая вероятность подчиняется бета-биномиальному распределению. При этом учитываются те данные, которые уже есть в сети.

Рассмотрим использование построенной сети на примере обработки данных из одной конкретной карточки. Исходные данные пациента представлены в таблице 1 в строке под названием "I вариант". Затем пациент прошел ряд обследований. Информация о результатах различных анализов и процедур была записана в карточку. Новые данные из медкарты отражены в таблице 1 в строке "II вариант".

Таблица 1.

Исходные данные пациента

Набор данных о пациенте	Диабет	Стресс	Гиподинамия	Ожирение	Алкоголь	Курение	Боли в сердце	Кл. анализ	Измен. ЭКГ	Холестерин	Давление
I вариант	нет								нет	да	
II вариант	нет		есть	есть	нет	есть	нет	нет	нет	да	есть

Вероятности возникновения заболеваний ССС у конкретного пациента в результате применения разработанной БСД до и после обследований представлены в таблице 2. Новые данные, которыми пополнилась карточка пациента, изменили риски возможных заболеваний: риск заболеть ИБС практически исчез, а вероятности заболеть АГ и атеросклерозом существенно увеличились. Теперь лечащему врачу нужно обратить внимание пациента на устранение факторов риска, присутствующих в жизни пациента, которые могут привести к этим заболеваниям.

Таблица 2.

Результаты применения БСД

Результаты обработки	Стенокардия	ИБС	АГ	Атеросклероз
До исследований	3,82	31,9	59,1	64,9
После исследований	0,69	0,65	86,6	91,5

Выводы. В данной работе были получены следующие результаты:

1. Проведен анализ математических моделей, которые применяются для решения задачи раннего диагностирования, в результате которого был обоснован выбор БСД.

2. Разработана БС объемом в 15 узлов, из которых 6 узлов представляют собой факторы риска, 5 узлов – симптомы заболеваний, 4 узла – заболевания ССС.

3. Приведен пример использования построенной сети для обработки медицинских данных из карточки пациента. Предложенный математический аппарат может использоваться не только для определения рисков, но и для прогнозирования возникновения определенного симптома.

ЛИТЕРАТУРА (REFERENCES TRANSLATED AND TRANSLITERATED)

1. Мельник К.В. Задача создания информационной системы скрининга в медицинских учреждениях // Восточно-европейский журнал передовых технологий. – Харьков. – 2012. – №1 /11(55). – с.55-57.

Mel'nik K.V. Zadacha sozdaniya informatsionnoy sistemy skrininga v meditsinskikh uchrezhdeniyakh [The task of creating a screening information system in healthcare facilities] // Vostochno-yevropeyskiy zhurnal peredovykh tekhnologiy. – Kharkiv. – 2012. – №1 /11(55). – s.55-57.

2. Мельник К.В., Єршова С.І. Проблемы и основные подходы к решению задачи медицинской диагностики // Системи обробки інформації. – Харків. – 2011. – №2 (92). – с.244-248.

Mel'nik K.V., Ershova S.I. Problemy i osnovnyye podkhody k resheniyu zadachi meditsinskoy diagnostiki [Issues and approaches to solving the problem of medical diagnosis] // Sistemi obrobki informatsii. – Kharkiv. – 2011. – №2 (92). – s.244-248.

3. Melnik K., Cherednichenko O., Glushko V. Towards medical screening information technology: the healthgrid-based approach. H.C.Mayr et al.(Eds.): UNISCON 2012, LNBIP 137. – Springer, 2013. – P. 202-204.
4. Richard E.Neapolitan. Learning Bayesian networks. Northeastern Illinois University. Chicago, Illinois: Prentice Hall, 2003. – 674 p.
5. I. A. Beinlich, H. J. Suermondt, R. M. Chavez, and G. F. Cooper. The ALARM Monitoring System: A Case Study with Two Probabilistic Inference Techniques for Belief Networks. In Proceedings of the 2nd European Conference on Artificial Intelligence in Medicine. Springer-Verlag, 1989. – P. 247-256.
6. S. Lauritzen, D. Spiegelhalter. Local Computation with Probabilities on Graphical Structures and their Application to Expert Systems (with discussion). Journal of the Royal Statistical Society: Series B (Statistical Methodology), 50(2):157-224, 1988.
7. S. Andreassen, R. Hovorka, J. Benn, K. G. Olesen, and E. R. Carson. A Model-based Approach to Insulin Adjustment. In Proceedings of the 3rd Conference on Artificial Intelligence in Medicine. Springer-Verlag, 1991. – P. 239-248
8. D. Heckerman, E. Horwitz, and B. Nathwani. Towards Normative Expert Systems: Part I. The Pathfinder Project. Methods of Information in Medicine, 31:90-105, 1992.
9. S. Andreassen, F. V. Jensen, S. K. Andersen, B. Falck, U. Kjærulff, M. Woldbye, A. R. Sørensen, A. Rosenfalck, and F. Jensen. MUNIN - an Expert EMG Assistant. In Computer-Aided Electromyography and Expert Systems, Chapter 21. – Elsevier (Noth-Holland), 1989.
10. A. Onisko. Probabilistic Causal Models in Medicine: Application to Diagnosis of Liver Disorders. Ph.D. Dissertation, Institute of Biocybernetics and Biomedical Engineering, Polish Academy of Science, Warsaw, March 2003.

Melnik K.V., Glushko V.N. Application of Bayesian networks for data processing from medical cards

Abstract. Medical information technologies are the essential tool for solving various health problems. It is reflected in a large number of publications. Nowadays medical facilities have started to use electronic documents. There are many reasons for this: a convenient storage of the patient information, easy and quick access to information, increasing the efficiency of the healthcare facility in medical data processing. Using statistical information from the electronic database of patients helps to identify risk factors for various diseases in the service area. Patients don't visit medical institutions regularly, so the diagnosis is a classification task on incomplete information. Bayesian networks are the appropriate mathematical tool to solve such problems. An overview of the application of Bayesian networks for disease diagnosis is provided in the article. Analysis of practical implementations of Bayesian networks showed the feasibility of this approach for diagnosis. Therefore, Bayesian belief networks are encouraged to apply for early diagnosis in uncertain and incomplete information from patients cards. The paper proposes a network, which helps to detect violations of the cardiovascular system. The information used to construct a Bayesian network was divided into three groups: the first group is risk factors that influence the emergence and development of certain diseases, the second group is a collection of diagnoses under consideration, and the third group is a doctor's supervision, a variety of symptoms and results of laboratory or instrument-computer analyses. The article shows how to use the developed Bayesian network.

Keywords*: medical information technologies, Bayesian networks, medical data processing, cardiovascular system.*