ОБЗОР ИСПОЛЬЗОВАНИЯ ТЕХНОЛОГИЙ БОЛЬШИХ ДАННЫХ В МЕДИЦИНСКОМ ПРОГНОЗИРОВАНИИ

Колтунов Игорь Ильич

д-р техн. наук, профессор, Московский Политехнический Университет, РФ, г. Москва

Панфилов Антон Владимирович

канд. экон. наук, генеральный директор НПГ «Традиция», РФ, г. Москва

Посельский Иван Александрович

руководитель НТЦ «Автоматизированные технические системы» Московского политехнического университета, РФ, г. Москва

Чубуков Николай Николаевич

руководитель проектов, НПГ «Традиция», РФ, г. Москва

Матьков Станислав Сергеевич

инженер, НПГ «Традиция», РФ, г. Москва

Данное исследование выполняется при финансовой поддержке Министерства Образования и Науки России по соглашению №14.577.21.0232 от «29» сентября 2016 года (уникальный номер RFMEFI57716X0232), прикладные научные исследования проводятся по теме «Исследование научно-технических решений и разработка интеллектуальной нательной биосенсорной платформы превентивного мониторинга и оценки показателей организма человека «Сенсорная сеть тела» с возможностью коррелирования данных, полученных от различных датчиков в зашумленной среде».

Аннотация. Целью работы является исследование вариантов аппаратно-программных и конструктивно-компоновочных решений для создания неинвазивной нательной биосенсорной платформы мониторинга физиологических показателей состояния здоровья человека в режиме повседневного ношения с учетом особенностей применения: зашумленности среды, двигательной активности, сложностей учета априорной биомедицинской информации. Актуальность работы обусловлена ожидаемым существенным повышением качества предоставления населению России медицинских услуг, а также снижением их стоимости за счет активного внедрения в оснащение лечебных заведений самых передовых IT-решений, основанных на современных открытиях фундаментальной науки. До настоящего времени не сформировалась стройная идеология разработки биосенсорных сетей для текущей неинвазивной диагностики, что определяет новизну проводимых исследований. Предлагаемый подход является развитием идеи телемедицины для решения задач функциональной диагностики с высокой степенью автоматизации. В статье показаны основные аспекты и проблематика разработки эффективных моделей текущей диагностики и диагностического прогнозирования состояния здоровья пациента – объекта неинвазивного мониторинга, на основе текущего анализа характеристических сочетаний его жизненных показателей по нозологиям и результатам долгосрочного сбора, обработки и семантической классификации биомедицинских данных.

Колтунов И.И. Обзор использования технологий больших данных в медицинском прогнозировании / И.И. Колтунов, А.В. Панфилов, И.А. Посельский, Н.Н. Чубуков, С.С. Матьков // Научный форум: Медицина, биология и химия: сб. ст. по материалам XVII междунар. науч.-практ. конф. — № 9(17). — М., Изд. «МЦНО», 2018. — С. 38-45.

Ключевые слова: биосенсор; биосенсорная платформа; диагностическая информативность; диагностический прогноз; неинвазивный мониторинг; нозология; облачные технологии; телемедицина.

Введение. В настоящее время технологии больших данных нашли широкое применение в области медицины и здравоохранения. Раздел, связанный с большими данными включает в себя методы хранения, аналитики и обработки большого количества разнородной, структурированной и неструктурированной информации. С момента внедрения информационных технологий в сектор здравоохранения, медицинская индустрия хранит и накапливает большие объемы данных, такие как записи наблюдений, история болезней, рентгеновские и томографические изображения, данные страхования и другие. Использование этих данных обещает принести множество преимуществ в таких областях, как поддержка принятия клинических решений, мониторинг заболеваний, выявление критических трендов и прогнозирование тенденций, влияющих на здоровье населения.

Настоящая работа посвящена анализу успешных применений технологии больших данных в медицине, определению перспектив ближайшего развития таких технологий.

Описание проблемы. Разнообразие информации предоставляет возможность поиска закономерностей, которые не видны при рассмотрении отдельных сегментов данных. Однако, традиционные средства обработки данных не позволяют эффективно извлекать значимую информацию из таких источников. Именно поэтому было сформировано само понятие больших данных и разработаны системы, а также методики работы с ними. Применение этих технологий привело ко многим замечательным результатам в социологии и маркетинге.

По большому счету, методология больших данных очень робко проникает в медицину. Данные связанного с медициной характера могут поступать из разных источников: записи наблюдений, клинические системы поддержки принятия решений, государственные источники, данные лабораторий и др. Важным источником данных могут служить «носимые устройства» [4], [5], [6], которые обретают значимое место в медицинских технологиях [3].

Врач же может ориентироваться только не небольшой объем информации о пациенте, которая собрана в медицинских записях или получена в результате целенаправленного обследования. Без специальных средств обработки и анализа накопленных данных врач, при принятии решения способен охватить лишь небольшую часть даже этой информации.

В этом отношении перспектива применения методологии больших данных в медицине весьма обширна, а сами применения требуют исследований [1].

Несмотря на то, что методология больших данных находится лишь на начальной стадии проникновения в медицину, за последние годы опираясь на близкие к ней методы получен ряд практически значимых результатов.

В этом разделе рассмотрены некоторые из опубликованных результатов, полученных на основе методов больших данных. По каждому из них определена направленность и ключевые аспекты проведенного исследования. Результаты представлены в виде 2-х условных групп: диагностика заболеваний и определение состояния.

Методика. Распознавание проблем с коронарной артерией. Исследовательская группа Университета Бойнора, Иран, предложила метод распознавания коронарной недостаточности основанный на применении нейронной сети в связке с генетическим алгоритмом для оптимизации весов. Также, для оценки качества предлагаемого метода было проведено распознавание без использования генетического алгоритма. Для выделения признаков был использован метод опорных векторов. Классификация осуществлялась посредством трехслойной нейронной сети [13].

В результате, нейронная сеть в связке с генетическим алгоритмом дала на 9 % (Accuracy) лучший результат – 93.85 %.

Распознавание инфаркта миокарда с помощью сверточной нейронной сети. В данном исследовании была применена 11-ти слойная сверточная нейронная сеть, распознающая нормальное сердцебиение и отклоняющееся от нормы [7]. Данный способ примечателен тем, что он может работать в условиях зашумленности данных, так как он работает сразу со всеми признаками, не отбирая их и не проводя никакой предобработки.

Количество итераций для тестирования равно 60. Выборка была поделена на обучающую (90 %) и тестовую (10 %). Для разбиения была применена кросс-валидация по 10 блокам.

В результате, в условиях зашумленности, нейронная сеть дала 93.53 % точности (Accuracy). В условиях без шума результат 95.22 % (Accuracy).

Распознавание и локализация инфаркта миокарда. Для распознавания и локализации MI был применен метод «ближайшего соседа» [12]. Этот метод предполагает распознавание MI без обучения. Идея состоит в том, что для каждого измерения вычисляется расстояние до ближайшего измерения.

Данный алгоритм является ресурсоемким и может обрабатывать данные очень много времени. Для уменьшения времени распознавания был применен прунинг (Arif-Fayyaz pruning) входных данных.

В целом, результат распознавания превышает 96 %(Sensitivity) и 97 %(Specificity), исключением является тип Inferio-posterior-lateral, распознавание которого дало 93.37 %(Specificity).

Распознавание гипертонии. Для предсказания гипертонии китайские исследователи применили множество подходов, из которых был выделен алгоритм «случайного леса», давший наилучшие показатели [8].

Для удаления аномалий был использован Межквартильный размах. Этот метод отлично работает для «симметричных» данных, в которых медиана равна среднему значению размаха. Пропуски в значениях диастолического артериального давления были заменены на средние значения по выборке. Также была произведена дискретизация [9].

Для классификации болезни были рассмотрены шесть подходов: нейронная сеть с обратным распространением ошибки, LogitBoost, локально взвешенный наивный Байес, байесовская сеть, метод опорных векторов и «случайный лес».

В качестве алгоритма для отбора признаков был выбран алгоритм ранжирования по показателю прироста информации (information gain [10]), а затем по их индивидуальной оценке [11].

Наилучший результат показал алгоритм «случайного леса» (AUC = 0,93).

В этой работе характерным является использование большого количества медицинских параметров, а также применение характерных для больших данных методов исследования. В результате получена сравнительная характеристика ряда различных алгоритмов классификации с целью диагностики.

Динамика потребления кислорода. Чтобы выявить динамику потребления кислорода человеком, был использован метод предсказания, основанный на алгоритме «случайного леса» [14]. Данная методика позволит спрогнозировать отклонение здоровья человека от нормы на ранней стадии.

Для оценки результатов была использована метрика MNG (mean normalized gain amplitude). Она же используется в качестве показателя динамики потребления.

В этой работе данные получены в результате медицинских измерений с нательных датчиков. Не смотря на относительно не большой объем анализируемых данных, благодаря примененной методики получен практически значимый результат.

Выявление падения. Люди преклонного возраста сталкиваются с проблемой падения, дома или на улице. В таких ситуациях встает вопрос об адекватном реагировании на экстренные ситуации такого рода для систем мониторинга жизненных показателей. Для этого был разработан метод, позволяющий не только определять состояния падения, но и классифицировать их по характеру падения. Более того, эти данные помогут терапевтам понять причины падения для лучшего ухода [15].

Алгоритм состоит из 2 основных шагов: на первом этапе технология smart textile позволяет собрать все необходимые данные о человеке (координаты с акселерометра, параметры дыхания и сердцебиения). На втором этапе, если произошло предполагаемое падение, метод опорных векторов классифицирует падение и причисляет его к одному из классов (во время подъема, во время спуска, во время прогулки, во время пробежки, стоя, падение вперед, падение назад, падение вправо, падение влево, лежа, сидя).

Метрики результатов получились следующие: accuracy = 98 %, sensitivity = 97.6 %, specificity = 98.5 %.

В работе данные сбирались в процессе исследования. Номенклатура измеряемых показателей не велика. Ключевым в получении результата является применение носимой электроники.

Распознавание состояний человека. Методы обработки больших данных могут использоваться не только для предсказания появления или не появления болезни пациента, но и также предсказывать будущие когнитивные (психологические) состояния.

Выявление когнитивных состояний. Для выявления когнитивных состояний были применены методы «ближайшего соседа» и «случайного леса» [16]. В качестве метрики расстояния для метода ближайшего соседа было выбрано расстояние Минковского.

Для сбора данных был использован следующий подход: на испытуемых вешали сенсоры, считывающие жизненные показатели (сердечный ритм, частота дыхания, частота шагов и др.). После получения данных, испытуемые фиксировали свое состояние через специальное мобильное приложение. Таким образом была получена выборка для обучения.

Данные для обучения и для тестирования были поделены с помощью кросс-валидации по 10 блокам. В среднем, результаты показывают 0.56(AUC), 57,2 %(Accuracy) для метода ближайшего соседа и 0,65(AUC), 60.2 %(Accuracy) для «случайного леса».

Работа так же опирается на применение носимой электроники и характерные для больших данных методы обработки. Выявлению состояния пациента посвящены и некоторые другие работы. В частности, электрокардиограмма может быть использована для выявления эмоционального состояния [2]. Ориентированы такие работы на получение результатов в экстренном режиме.

Результаты. В данной статье рассмотрены различные сферы применения технологий больших данных в области медицины и здравоохранения, связанные с системой мониторинга и анализа медицинских данных пациента. Выявлены типичные методы обработки, такие как нейронные сети и машинное обучение.

Заключение. Результаты, полученные в данной статье показывают, что тема исследованию в области систем сбора и обработки данных носимых биосенсоров интеллектуальными методами является перспективной.

Список литературы:

1. W. Raghupathi, V. Raghupathi. (2014). Big data analytics in healthcare: promise and potential. Health Inf Sci Syst. 2014; 2: 3.

2. Chen M., Ma Y., Song J. (2016). Smart Clothing: Connecting Human with Clouds and Big Data for Sustainable Health Monitoring. Mobile Netw Appl (2016) 21: 825. https://doi.org/10.1007/s11036-016-0745-1.

3. Poon CC, Lo BP, Yuce MR, Alomainy A, Hao Y. (2015). Body sensor networks: In the era of big data and beyond. IEEE Rev Biomed Eng 8:4–16.

4. Song Z, Liu CH, Wu J, Ma J, Wang W. (2014). Qoi-aware multi-task-oriented dynamic participant selection with budget constraints. IEEE Trans Veh Technol 63(9):4618–4632.

5. H. Banaee, M. Uddin Ahmed and A. Loutfi. (2013). Data Mining for Wearable Sensors in Health Monitoring Systems: A Review of Recent Trends and Challenges. Sensors 2013,13, 17472-17500; doi:10.3390/s131217472.

6. Data Mining, Soft Computing, Machine Learning and BioInspired Computing for Heart Disease Classification/ Prediction – A Review

7. U. Rajendra Acharya, Hamido Fujita, Shu Lih Oh, Yuki Hagiwara, Jen Hong Tan, Muhammad Adam. (2017). Application of Deep Convolutional Neural Network for Automated Detection of Myocardial Infarction Using ECG Signals. Information Science 201;
190–198.https://doi.org/10.1016/j.ins.2017.06.027.

8. Mouaz H., Al-Mallah, Radwa Elshawi, Amjad M. Ahmed. (2017). Using machine learning on cardiorespiratory fitness data for predicting hypertension: The Henry Ford ExercIse Testing (FIT) Project. The American Journal of Cardiology 120(11) August 2017; DOI:10.1016/j.amjcard.2017.08.029.

9. Kurgan L, Cios KJ. (2001). Discretization algorithm that uses class-attribute interdependence maximization. In: Proceedings of the 2001 International Conference on Artificial Intelligence (IC-AI 2001); 2001. p. 980– 987.

10. Kent JT. (1983). Information gain and a general measure of correlation. Biometrika. 1983; 70(1):163–173. https://doi.org/10.1093/biomet/70.1.163.

11. Guyon I, Elisseeff A. (2003). An introduction to variable and feature selection. Journal of machine learning research. 2003; 3(Mar):1157–1182.

12. Azadeh Noorian, Nader Jafarnia Dabanloo, Saman Parvaneh. (2014). Detection and Localization of Myocardial Infarction using K-nearest Neighbor Classifier. Conference: Computing in Cardiology 2014; January 2014.

13. Zeinab Arabasadi, Roohallah Alizadehsani, Mohamad Roshanzamir, Hossein Moosaei. (2017). Computer aided decision making for heart disease detection using hybrid neural network-Genetic algorithm. Computer methods and programs in biomedicine 141; January 2017; DOI: 10.1016/j.cmpb.2017.01.004.

14. Beltrame T. et al. (2017). Prediction of Oxygen Uptake Dynamics by Machine Learning Analysis of Wearable Sensors during Activities of Daily Living. Scientific Reports. 7, p. 45738, 2017.

15. Webster E. et al. (2017). Predicting Cognitive States from Wearable Recordings of Autonomic Function.IBM Journal of Research and Development, 61, (2/3), p. 2:1-2:11, 2017.

16. «Apache Hadoop», http://hadoop.apache.org.

Принятые обозначения:

Чувствительность (**sensitivity**). Данная метрика показывает отношение правильно распознанных «больных» пациентов ко всем «больным» пациентам в выборке;

Специфичность (**specificity**). Данная метрика показывает отношение правильно распознанных «здоровых» пациентов ко всем «здоровым» пациентам в выборке;

Точность (**accuracy**). Данная метрика показывает отношение всех распознанных пациентов ко всем пациентам в выборке;

True positive (TP). Количество правильно распознанных «больных» пациентов;

True negative (TN). Количество правильно распознанных «здоровых» пациентов;

False positive (FP). Количество неправильно распознанных «больных» пациентов;

False negative (FN). Количество неправильно распознанных «больных» пациентов;

Positive predictive value (PPV). Вероятность того, что пациент, распознанный как «больной», действительно имеет болезнь;

Площадь под ROC-кривой (AUC). Зависимость доли TP от доли FP.