

УДК 004.046

АНСАМБЛЬ КЛАССИФИКАТОРОВ: РЕАЛИЗАЦИЯ, ОЦЕНКА ЭФФЕКТИВНОСТИ И ИНТЕГРАЦИЯ В ОБЛАЧНУЮ ПЛАТФОРМУ ХРАНЕНИЯ, СИСТЕМАТИЗАЦИИ И ОБРАБОТКИ МЕДИЦИНСКИХ ДАННЫХ

¹Алексеев Д.М., ¹Минюк А.Н., ²Понимаш З.А., ¹Шумилин А.С.¹ФГАОУ ВО «Южный федеральный университет», Институт компьютерных технологий и информационной безопасности, Таганрог, e-mail: minyuk@neurotech.ru;²ФГАОУ ВО «Южный федеральный университет», Институт радиотехнических систем и управления, Таганрог, e-mail: ponimash@neurotech.ru

В рамках работы цель исследований заключается в разработке и оценке эффективности общей схемы облачной платформы хранения, систематизации и обработки медицинских данных, а также в разработке ансамбля классификаторов для решения задачи выделения зон эпилептической активности. В рамках данной работы был разработан новый метод быстрого поиска и классификации эпилептических паттернов в электроэнцефалограмме на основе бесконечного анализа независимых компонент, который позволяет выделить характерные признаки судорожной активности. Для решения проблемы обработки данных создается ансамбль подпространств, с помощью сочетания метода опорных векторов, многослойного перцептрона и расширенного метода k-ближайших соседей. Для оценки эффективности предложенного подхода были обработаны контрольные записи ЭЭГ восьми пациентов с височной и экстракратемпоральной эпилепсией. В рамках оценки использовалась распределенная вычислительная система, основанная на многоуровневой облачной архитектуре. Используя методику перекрестного исключения по одному, была проведена оценка точности, чувствительности, конкретности, а также ложноположительных и ложноотрицательных результатов: 0,97; 0,97; 0,95; 0,05 и 0,03 соответственно. В рамках разработанной облачной платформы хранения, систематизации и обработки медицинских данных достигнута эффективная работа классификатора, основанного на комбинации классификаторов различных типов. Точность (97%) и высокая скорость работы данного классификатора являются оптимальными для разработанной облачной платформы.

Ключевые слова: облачные вычисления, классификатор, распределенная информационная система, машинное обучение, машина опорных векторов, нейронные сети, обработка данных, систематизация данных

CLASSIFIERS ENSEMBLE: IMPLEMENTATION, ESTIMATION OF EFFICIENCY AND INTEGRATION IN A CLOUDED PLATFORM FOR STORAGE, SYSTEMATIZATION AND PROCESSING OF MEDICAL DATA

¹Alekseev D.M., ¹Minyuk A.N., ²Ponimash Z.A., ¹Shumilin A.S.¹Federal Autonomous Educational Institution of Higher Education Southern Federal University, Institute of Computer Technologies and Information Security, Taganrog, e-mail: minyuk@neurotech.ru;²Federal Autonomous Educational Institution of Higher Education Southern Federal University, Institute of Radio Systems and Control, Taganrog, e-mail: ponimash@neurotech.ru

In this article the purpose of research is to develop and evaluate the effectiveness of the general scheme of the cloud platform for storing, systematizing and processing medical data, as well as to develop an ensemble of classifiers for solving the problem of identifying zones of epileptic activity. In the framework of this work, a new method was developed for the rapid search and classification of epileptic patterns in an electroencephalogram based on an endless analysis of independent components, which allows us to highlight the characteristic signs of convulsive activity. To solve the data processing problem, an ensemble of subspaces is created using a combination of the support vector method, a multilayer perceptron, and an extended method of k-nearest neighbors. To evaluate the effectiveness of the proposed approach, the EEG control records of eight patients with temporal and extra temporal epilepsy were processed. As part of the assessment, a distributed computing system based on a multi-level cloud architecture was used. Using the method of cross exclusion one at a time, the accuracy, sensitivity, specificity, as well as false positive and false negative results were evaluated: 0.97, 0.97, 0.95, 0.05, and 0.03, respectively. It was used the cloud platform for storing, systematizing and processing medical data, an efficient operation of the classifier based on a combination of classifiers of various types has been achieved. The accuracy (97%) and high speed of this classifier are optimal for the developed cloud platform.

Keywords: cloud computing, classifier, distributed information system, machine learning, support vector machine, neural networks, data processing, data systematization

В век всеобщей информатизации и активного развития информационных технологий медицинские учреждения в ходе выполнения диагностических исследований обрабатывают и систематизируют значительные объемы данных для последующей

реабилитации и лечения пациентов в целом. Эффективность оказываемой медицинской помощи прямо пропорциональна оперативности и удобству использования данной информации специалистами медицинских организаций. Возможность оперирования

данными в электронном виде обеспечивает оперативность получения врачом необходимой информации о пациенте, что увеличивает скорость принятия решения о постановке диагноза и методах лечения [1].

Информационные технологии, модели обработки данных и методы работы со знаниями играют решающую роль как в автоматизации рутинных процессов оказания медицинских услуг, так и в развитии научно-технического прогресса в медицине. Так с использованием накопленной информации удалось обучить искусственный интеллект распознавать онкологические заболевания, создать системы прогнозирования и поддержки принятия клинических решений.

В настоящий момент далеко не вся информация, которая создается и накапливается в процессе оказания медицинских услуг, может быть обработана с целью получения новых знаний. Одной из таких областей слабоструктурированных данных является электроэнцефалография. Данные, созданные во время обследования, как правило, сохраняются только локально, на рабочих станциях специалистов, проводящих исследование, что ограничивает возможность накопления больших данных и их обработку с целью получения новых знаний.

В настоящий момент основной клинической задачей проведения обследования методом регистрации электроэнцефалограмм является диагностика хронического неврологического заболевания – эпилепсии. Регистрация ЭЭГ является основным (но не единственным) методом исследования эпилепсии, на который во многом опирается специалист, делая выводы и назначая терапию.

Обнаружение эпилептической активности в сигналах ЭЭГ представляет собой достаточно сложную задачу в связи с тем, что до настоящего времени неизвестен точный механизм возникновения эпилептических припадков. Для автоматического обнаружения эпилептической активности в основном используются линейные (частотно-временные, математические и статистические) методы, где не учитывается нелинейность анализируемого сигнала [2].

Таким образом, исследование существующих информационных структур создания и хранения данных ЭЭГ и разработка моделей информационных процессов и структур для агрегации и унификации больших данных со всех известных источников являются актуальными задачами.

Для того чтобы созданная информационная структура хранения больших данных позволила приобрести новые знания о предметной области, необходимо разработать модель инфраструктуры, позволяющей

impleментировать различные алгоритмы искусственного интеллекта. В условиях возрастающего объема информации и развития новых методов машинного обучения разрабатываемая модель должна быть легко масштабируемой, гибкой, как с точки зрения увеличения вычислительного ресурса, так и с точки зрения аддитивности новых алгоритмов обучения и классификации данных. Такой подход позволит решать как научные задачи предметной области, так и прагматичные задачи интеллектуализации бизнес-процессов.

В настоящее время широкое распространение получили модели облачных вычислений, позволяющие достичь высокой вычислительной производительности и масштабируемости. Облачные сервисы доступны из любой точки сети Интернет и являются экономичным и универсальным инструментом для реализации различных конфигураций моделей сбора, накопления и обработки информации.

Цель исследования: разработка информационной структуры облачной системы хранения и алгоритма обработки накопленных данных с помощью ансамбля методов машинного обучения.

Для достижения указанной цели в рамках работы необходимо решить следующие задачи: провести анализ существующих моделей информационных процессов и структур в предметной области, изучены средства создания, накопления и обработки информации; разработать архитектуру облачной платформы распределенного хранения данных, позволяющую взаимодействовать с различными информационными системами и комплексами предметной области с целью формирования больших данных; разработать алгоритм обработки больших данных на основе ансамбля методов машины опорных векторов многослойного перцептрона и расширенного метода k-ближайших соседей; создать интегрируемую облачную платформу распределенного хранения электроэнцефалографических данных и систему выделения и классификации паттернов эпилептической активности с использованием разработанного метода машинного обучения.

Материалы и методы исследования

В качестве материалов исследования для решения задачи поиска зон ЭПИ-активности используется ансамбль классификаторов, состоящий из следующих элементов:

1. Классификатор, основанный на машине опорных векторов (SVM).
2. Классификатор, базирующийся на обучении многослойного перцептрона (MLP).

3. Классификатор, построенный на базе метода ближайшего соседа.

Для решения задачи хранения, систематизации и обработки медицинских данных разработана облачная платформа, общая схема которой представлена на рис. 1.

Разработанная облачная система включает четыре основных уровня:

1. *Уровень хранения данных*: глобальное хранилище обезличенных данных, которое включает в себя базу данных для хранения сырых данных обследований и обезличенных отчетов, а также антропометрическая, диагностическая, демографическая информация о пациентах без персональных данных. Хранилище содержит полный объем информации для исследований и обучения машинных алгоритмов, но идентификация пациента возможна только по защищенному идентификатору.

2. *Уровень потребителей данных* – слой, включающий системы, которые принимают и обрабатывают данные из Глобального хранилища или передают в него новые данные. Этот уровень связан с уровнем хранения данных через стандартизированный программный интерфейс (Storage API). Потребителями данных могут быть: сторонние медицинские информационные системы; исследовательские системы; информационная система обработки данных содержит базу персональных данных пациентов, соответствует требованиям безопасности и защиты персональных данных и медицинских данных (Федеральный закон РФ от 27 июля 2006 г. № 152-ФЗ «О персональных данных»; Федеральный закон от 21.11.2011 № 323-ФЗ «Об основах охраны здоровья

граждан в Российской Федерации»; Health Insurance Portability and Accountability Act of 1996, HIPAA) [3]. Данный модуль обеспечивает взаимодействие с конечными клиентскими приложениями посредством распределенного интерфейса (REST API).

3. *Уровень прикладного ПО* – уровень, содержащий программные средства конечных клиентов, где формируются и/или отображаются медицинские данные (обследования в виде сигналов, отчетные и персональные данные пациента): *Windows клиенты* – программное обеспечение для ОС семейства Windows; *Веб-сервер* – предоставляет пользователю возможность доступа через web browser, в соответствии с назначенными этому пользователю ролями; *Мобильный клиент* – предоставляет доступ в информационную систему обработки данных, используя мобильные устройства (Android, iOS).

4. *Уровень аппаратных средств* – физические устройства для проведения обследований. В общем случае могут быть различных видов: электроэнцефалографы, кардиографы, системы биологической обратной связи, носимые фитнес-трекеры и т.д.

Результаты исследования и их обсуждение

Система автоматического поиска участков эпилептической активности представляет собой облачный сервис хранения и обработки электроэнцефалографических сигналов, а также программное обеспечение для рабочей станции невролога-специалиста.

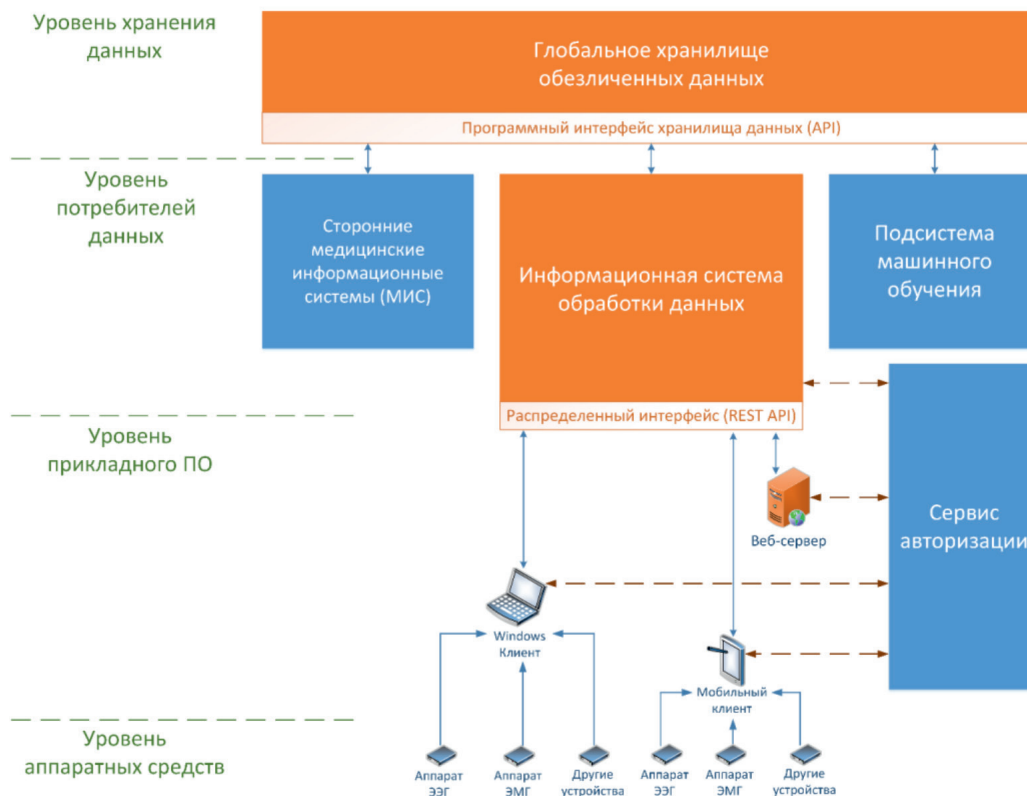


Рис. 1. Общая схема облачной платформы хранения, систематизации и обработки медицинских данных

Задачей системы является упрощение процесса анализа сигналов ЭЭГ специалистом с помощью алгоритма автоматизированного поиска эпилептической активности.

В результате работы алгоритма классификации создается файл разметки, который содержит в себе временные метки найденных участков в соответствующих каналах электроэнцефалограммы (рис. 2). С помощью пользовательского интерфейса врач верифицирует результаты автоматического анализа. В результате обработки данных врачом-экспертом формируется второй файл разметки. Результаты анализа вместе с сигналами загружаются на сервера облачного хранилища. Специализированная подсистема обучения производит сравнение разметки, полученной алгоритмом, и разметки, полученной в результате верификации специалистом, и выделяет ошибки первого и второго рода. Используя исходные сигналы, подсистема обучения переобучает алгоритм и формирует новую матрицу признаков, которая загружается в конфигурационный файл, позволяющий задавать настройки каскада признаков класса. Постоянное переобучение существующего классификатора позволит повышать качество автоматического анализа в ходе эксплуатации системы.

Рассмотрим предложенный ансамбль классификаторов, используемый для решения задачи автоматизированного поиска зон ЭПИ-активности. Схема предложенного подхода на базе облачных виртуальных кластеров представлена на рис. 3.

После предварительной фильтрации данных и удаления артефактов производится извлечение большого количества времен-

ных и частотных признаков сигнала. Так как пространство полученных признаков имеет большую размерность, процесс обучения и оптимизации классификаторов имеет высокую вычислительную сложность. Предлагается сократить пространство признаков без снижения их информативности.

Допустим, нам дано пространство признаков $u \in R^N$, методы выбора признаков находят отображение $x = f(y): R^N \rightarrow R^M$ ($M < N$) таким образом, чтобы вектор $x \in R^M$ сохранил большую часть информации о u . Для выбора наиболее информативных признаков предлагается использовать расширенный вариант метода анализа независимых компонентов, который получил название бесконечного ICA (Infinite Independent Components Analysis) [4].

Ансамблевая классификация рассматривается как альтернатива отдельным классификаторам, чтобы избежать необходимости переобучения, вызванного высокой размерностью данных и малой обучающей выборкой. Эти методы позволяют принять совокупное взвешенное решение, а также улучшают общую эффективность обучения, комбинируя слабые классификаторы и объединяя их с помощью таких алгоритмов, как бэггинг [6]. Для повышения качества обучения системы и точности обнаружения, предлагается использовать комбинацию дополнительных методов классификации в виде ансамбля подпространств, который использует сочетание машины опорных векторов (SVM), многослойный перцептрон (MLP) и расширенный метод k-ближайших соседей (k-NN), который получил название метода расширенного ближайшего соседа (ENN) [5].

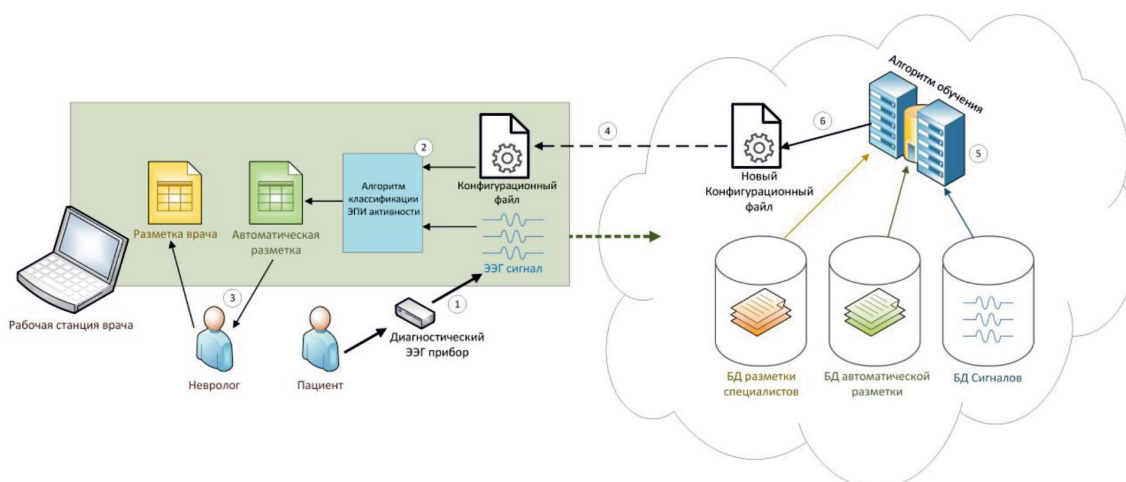


Рис. 2. Обучение алгоритма

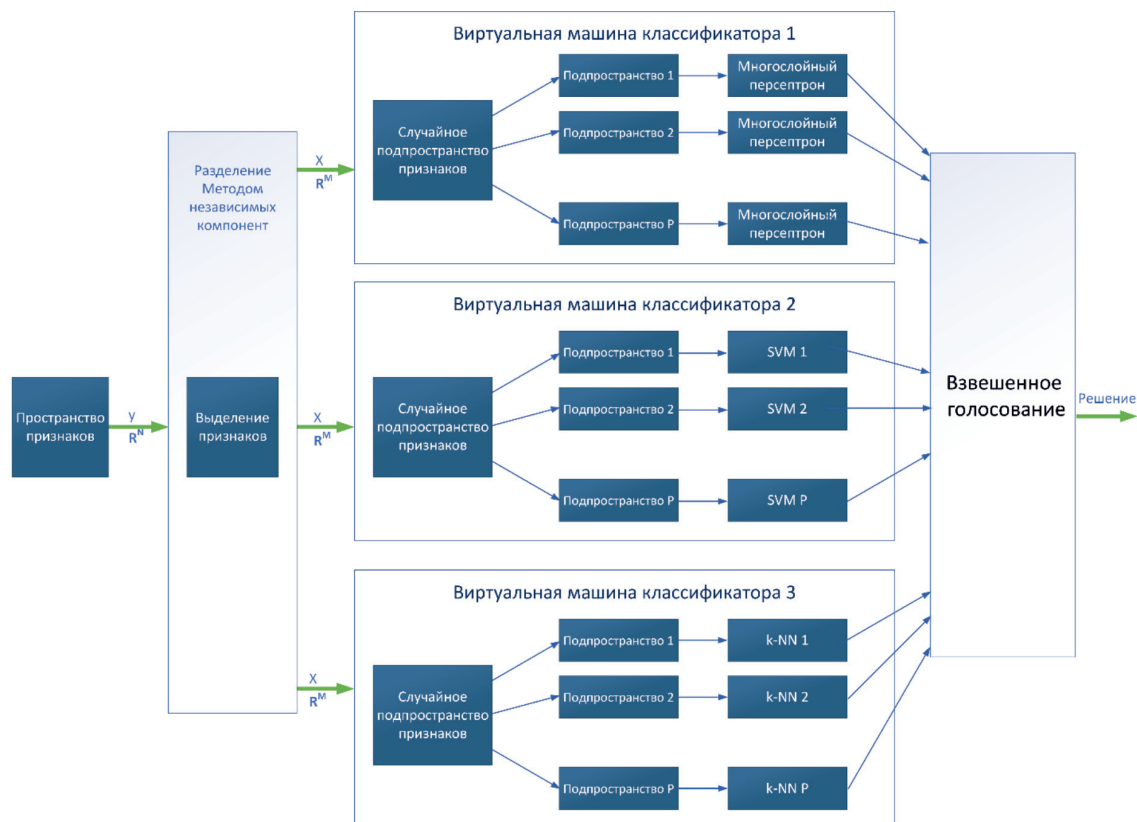


Рис. 3. Ансамбль классификаторов ЭПИ-активности

Для оценки эффективности предложенного подхода были обработаны контрольные записи ЭЭГ восьми пациентов с височной и экстратемпоральной эпилепсией. В рамках оценки использовалась распределенная вычислительная система, основанная на многоуровневой облачной архитектуре. Используя методику перекрестного исключения по одному, была проведена оценка точности, чувствительности, конкретности, а также ложноположительных и ложноотрицательных результатов: 0,97; 0,97; 0,95; 0,05 и 0,03 соответственно. Кроме того, в рамках проведения данного обследования удалось проанализировать возможности системы в рамках процедуры электроэнцефалографии.

Выводы

Разработанная платформа для распределенного облачного хранения и обработки данных, полученных в результате диагностических исследований, прошла процедуру апробации и предварительного тестирования. Можно отметить, что за счет гибкости архитектуры платформы и разделения данных по уровням, интеграция клиентских приложений и диагностических

устройств различных производителей не занимает много времени. С другой же стороны, решение легко подключается к существующим МИС и позволяет создать единое информационное пространство. С точки зрения структурирования и накопления информации был получен уникальный результат, так как платформа позволяет накапливать большие данные и обрабатывать их средствами машинного обучения.

На базе облачной платформы было проведено исследование и сравнительный анализ ряда методов классификации и реализован собственный алгоритм, основанный на комбинации классификаторов различных типов. За счет технологии виртуализации и подходов к обработке больших данных удалось достичь высокой скорости работы алгоритма. Более того, архитектура классификатора позволяет дополнять его новыми алгоритмами без существенной потери скорости работы и повышения вычислительной сложности. Кроме того, удалось достичь точности распознавания ЭПИ-паттернов ансамблем классификаторов равной 0,97, что значительно превышает результаты работы других алгоритмов.

Дальнейшие исследования ориентированы на расширение обучающих выборок за счет верификации результатов классификации высококвалифицированными специалистами и повышения эффективности работы ансамбля классификаторов ЭПИ-активности.

Работа выполнена при поддержке гранта РФФИ № 18-07 00 50.

Список литературы

1. Митькина П.А. Особенности хранения медицинской информации // Современные научные исследования и инновации. 2017. № 5 [Электронный ресурс]. URL: <http://web.snauka.ru/issues/2017/05/82546> (дата обращения: 27.07.2019).

2. Артеменко С.В. Нейросетевые алгоритмы обработки электроэнцефалограмм для диагностики эпилепсии: авто-

реф. дис. ... канд. техн. наук [Электронный ресурс]. URL: <https://www.bsu.by/ru/sm.aspx?guid=814633> (дата обращения: 26.07.2019).

3. Health Insurance Portability and Accountability Act [Электронный ресурс]. URL: https://en.wikipedia.org/wiki/Health_Insurance_Portability_and_Accountability_Act (дата обращения: 28.07.2019).

4. Knowles D., Ghahramani Z. Infinite sparse factor analysis and infinite independent components analysis. Independent component analysis and signal separation. Springer Berlin Heidelberg. 2017. P. 381–388.

5. Tang B., He H., Enn. Extended nearest neighbor method for pattern recognition [research frontier]. IEEE Comput Intell Mag. 2015. P. 52–60.

6. Mohammad-Parsa H., Dario P., Kost E., Hamid S. Random ensemble learning for EEG classification. Artificial Intelligence in Medicine. 2018. № 84. P. 146–158.