УДК 004.89, 616

МЕТОДЫ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА БИОМЕДИЦИНСКИХ ДАННЫХ

Е. В. Тимощенко, А. Ф. Ражков
(Учреждение образования «Могилевский государственный университет имени А. А. Кулешова», кафедра программного обеспечения информационных технологий)

Рассмотрены методы интеллектуального анализа на примере исследования биомедицинских данных. Разработано программное обеспечение для демонстрации возможностей этих методов с целью прогнозировании заболеваний сердечно-сосудистой системы, а также наличия заболевания по существующим симптомам у пациента. Обозначена перспектива использования разработки при преподавании дисциплин медико-биологического профиля.

Развитие информационных технологий привело к формированию понятия медицинская информатика. Большие данные медицинской информатики генерируются из различных источников и включают данные больницы, данные пациента, данные о заболевании и т.д. Эти данные могут быть обработаны с использованием методов интеллектуального анализа данных с целью прогнозирования заболеваний на этапе первичной диагностики. Интеллектуальный анализ данных – это процесс анализа, извлечения и предоставления данных в виде знаний, которые формируют взаимосвязь между доступными данными. Некоторые методы приобретения данных включают в себя ассоциацию, кластеризацию, классификацию и прогнозирование.

Применение методов интеллектуального анализа данных является ключевым подходом для извлечения знаний из больших данных о заболеваниях [3]. Некоторые из программных продуктов интеллектуального анализа данных, которые используются в настоящее время в здравоохранении, помогают прогнозировать заболевания на основе предыдущих данных, собранных по аналогичным симптомам, проводить диагностику заболеваний на основе данных пациентов; позволяют провести анализ затрат на лечение и потребности в ресурсах, а также предварительную обработку зашумленных, отсутствующих данных.

В связи с актуальностью темы исследования интересной показалась разработка программного обеспечения для интеллектуального анализа биомедицинских данных, которое может быть использовано для прогнозирования наличия болезни по исходным биомедицинским данным не только в профильных учреждениях здравоохранения, а также в процессе преподавания блока медико-биологических дисциплин в учебных заведениях непрофильного вуза.

При разработке использовались такие классификационные модели.

- Метод опорных векторов (Support Vector Machines, SVM). Основан на принципе построения оптимальных гиперплоскостей, разделяющих предположительно линейно разделяемые данные.

- Метод Байеса (Naive Bayes). Основан на применении теоремы Байеса с дополнительным предположением о независимости различных свойств рассматриваемой модели. Метод популярен благодаря его вычислительной и относительно хорошей прогностической эффективности.

- Логистическая регрессия (Logistic Regression). Данный алгоритм классификации использует сигмоидную функцию $g(x) = \frac{1}{1+e^{-x}}$ в качестве функции активации и позволяет дать вероятностную оценку принадлежности объекта каждому классу [1].

- Дерево решений (Decision Tree) – алгоритм машинного обучения (класс методов), который решает задачи прогнозирования так, как решал бы их человек. В общем случае – это k-ичное дерево с решающими правилами в нелистовых вершинах (узлах) и некотором заключении о целевой функции в листовых вершинах (прогнозом). Здесь решающее правило – это некоторая функция от объекта, позволяющая определить, в какую из дочерних вершин нужно поместить рассматриваемый объект [2].

Кроме того, были использованы: Random Forest (Decision Forest) – алгоритм машинного обучения с учителем, представляющий собой набор деревьев решений, при котором все деревья решений должны быть разными; LightGBM – платформа для градиентного бустинга, которая использует алгоритм обучения на основе дерева; XGboost – алгоритм машинного обучения, основанный на дереве поиска решений и использующий фреймворк градиентного бустинга.

Для разработки программного обеспечения было решено использовать язык программирования Python, используя библиотеки Pandas для работы с CSV-файлами, Numpy, Sklearn, предоставляющая использование алгоритмов, и Tkinter для создания графического интерфейса. Исходные данные для анализа были взяты из открытых источников в виде двух CSV-файлов, содержащих информацию о симптомах пациентов и последующем диагнозе, данные биомедицинского характера и о наличии сердечно-сосудистого заболевания.

В результате разработано два приложения для прогнозирования предрасположенности к сердечно-сосудистым заболеваниям по биомедицинским данным пациента и прогнозирования наличия заболевания по существующим симптомам у пациента.

Программы позволяют не только продемонстрировать возможности методов интеллектуального анализа для дальнейшего определения, находится ли пациент в группе риска при сердечно-сосудистых заболеваниях, и осуществления прогноза наличия заболевания при существующих у него симптомах. Также программы дают возможность их применения в диагностике и прогнозировании любых других заболеваний, основываясь при этом на новых исходных биомедицинских данных.

Так как разработанное программное обеспечение может быть легко модифицировано под конкретные запросы пользователя и содержит достаточное количество разнообразных данных для анализа, представляется перспективным его использование при преподавании дисциплин медико-биологического профиля в учреждениях образования непрофильного обучения в качестве, например, одного из программных компонентов виртуального лабораторного практикума. Это может способствовать развитию критического мышления, выработке навыков и умений практического использования получаемой информации, в том числе и дистанционно [4].

Литература

1. Воронцов, К. В. Лекции по линейным алгоритмам классификации [Электронный ресурс] / К В. Воронцов. – Электрон. текстовые дан. – 2009. – Режим доступа: http://www.machinelearning.ru/wiki/images/6/68/voron-ML-Lin.pdf. – Дата доступа: 29.01.2020.

2. Деревья решений и алгоритмы их построения [Электронный ресурс]. – Электрон. журн. – 2014. – Режим доступа: http://datareview.info/article/derevya-resheniy-i-algoritmyi-ih-postroeniya. – Дата доступа: 29.01.2020.

3. Ражков, А. Ф. Применение методов интеллектуального анализа биохимических данных при преподавании дисциплин медико-биологического профиля / А. Ф. Ражков, Е. В. Тимощенко // Научные стремления – 2019: материалы Междунар. науч.-практ. молодежной конференции в рамках Междунар. науч.-практ. инновационного форума «INMAX'19». Ч. 1. – Минск : Лаборатория интеллекта, 2019. – С. 89–90.

4. Юревич, Ю. В. Из опыта использования технологии дистанционного обучения в процессе получения дневной формы образования / Ю. В. Юревич, Е. В. Тимощенко // Качество подготовки специалистов в техническом университете : проблемы, перспективы, инновационные подходы : материалы IV Междунар. науч.-метод. конф., редкол. : А. С. Носиков (отв. ред.) [и др.]. – Могилев : МГУП, 2018. – С. 182–184.