УДК 004.89

ПРИМЕНЕНИЕ АЛГОРИТМОВ МАШИННОГО ОБУЧЕНИЯ ДЛЯ ПРЕВЕНТИВНОЙ ДИАГНОСТИКИ ЗАБОЛЕВАНИЙ

Симонов И.А.
(Федеральное государственное автономное образовательное учреждение высшего образования «Национальный исследовательский университет ИТМО»)
Научный руководитель – к.техн.н, Клечиков А.В.
(Федеральное государственное автономное образовательное учреждение высшего образования «Национальный исследовательский университет ИТМО»)

В работе представлены основные этапы, методы и результаты автоматизации превентивной диагностики заболеваний по данным электронных медицинских карт с использованием технологий машинного обучения. Работа содержит обобщенные промежуточные результаты исследования в рамках магистерской программы «Управление государственными информационными системами».

Одна из значимых проблем здравоохранения — повышение качества диагностики заболеваний. Задача диагностики, прогнозирования течения заболевания, выбора стратегии и тактики лечения требуют учета совокупности имеющейся информации о пациенте, без чего медицинские решения носят приблизительный, «неточный» характер.

Процесс постановки диагноза является непростой задачей даже для опытного врача. Результаты анализов могут быть неоднозначными: отклонение одного показателя от нормы может свидетельствовать о наличии сразу нескольких заболеваний. Необходимо учитывать и влияние человеческого фактора. Основные причины врачебных ошибок на стадии диагностики:

1. Наличие симптомов или синдромов, симулирующих другие заболевания;
2. Наличие атипичных симптомов часто встречающихся заболеваний;
3. Наличие симптоматики заболевания, встречающегося как казуистика (собрание конкретных случаев, выясняющих известную форму болезни);
4. Наличие проявлений нескольких заболеваний у одного пациента.

Для повышения качества медицинского обслуживания Санкт-петербургским медицинским информационно-аналитическим центром была сформулирована задача реализации подхода, который по данным, поступающим в процессе лечения (динамическим), сможет прогнозировать наличие исследуемого заболевания у пациентов.

Актуальность данной задачи, заключается в том, что её решение сможет помочь перейти от общего к персонализированному курсу лечению. В частности, индивидуальный подход к каждому пациенту позволяет в рамках дальнейшего лечения пациента опираться не только на общие рекомендации для таких случаев, но и учитывать его личные особенности. Кроме того введение этого и подобных подходов в качестве систем поддержки принятия решений позволит в автоматическом режиме обрабатывать большие объёмы информации, работать в качестве системы раннего предупреждения сразу же после получения новых анализов, тем самым снижая нагрузку на медицинский персонал и снижая вероятность ошибки из-за человеческого фактора.

Данная задача относится к задачам классификации. Для исследования было рассмотрено несколько различных алгоритмов классификации, чтобы получить наиболее точный результат классификации.

Были рассмотрены следующие алгоритмы:

1. Random forest
2. KNeighbors
3. Логистическая регрессия
4. SVC
5. Наивный байесовский
6. Градиентный бустинг над решающими деревьями

Значения гиперпараметров для каждого из классификаторов были определены с функции GridSearchCV поиска по сетке гиперпараметров. Таким образом представленные выше значения оптимальные для каждого из классификаторов. Также в исследование были включены ансамбли для проверки гипотезы о том, что каждый из классификаторов ошибается на разных данных, и их композиция позволит скрыть недостатки друг друга. Набор переменных для алгоритма классификации, при котором классификаторы показывают наилучшие результаты, составлялся через поиск по матрице корреляций. Все вычислительные эксперименты для данной работы проводились с помощью языка программирования python (версия 3.6) и следующих библиотек для него (numpy, matplotlib, pandas, seaborn, scikit-learn, scipy, collections, datetime).

В качестве перспективы работы и исследования планируется проведение тестирования работы прикладных решений на основе реальных данных с проведением статистического анализа.

В работе представлены обобщенные промежуточные результаты исследования в рамках магистерской программы «Управление государственными информационными системами».