Исследование методов машинного обучения для диагностики женского здоровья

Е.А. Скачкова[1], В.В. Мокшин[1]

[1]*Казанский национальный исследовательский технический университет им. А.Н. Туполева – КАИ, Карла Маркса 10, Казань, Россия, 420111*

Аннотация

В данной статье был проведен обзор существующих методов решения проблемы (деревья решений, нейросетевое обучение), выявлены их достоинства и недостатки. Предложенный метод отличается от существующих тем, что для выявления наиболее значимых признаков, влияющих на оценку вероятности получения заболевания, был осуществлен отбор признаков на основе различных методов (FR – forward regression, BR – backward regression, КМ – корреляционный метод, ПГА – параллельный генетический алгоритм), а также сравнение и выявление эффективных методов машинного обучения (метод 1 – регрессия, метод 2 – корреляционный метод, метод 3 – алгоритм Левенберга – Марквардта, метод 4 – байесовская регуляризация обратного распространения, метод 5 – градиентный спуск). Также был применен кластерный анализ, в результате которого выделилось три кластера. В результате сравнения был выбран метод обучения байесовской регуляризации обратного распространения, так как он показал наименьшую ошибку распознавания при относительно небольшом времени обучения, а также как метод отбора признаков был выбран генетический алгоритм как лучший метод для факторов 1 и 3, а также метод forward regression для фактора 2. Таким образом, была обучена нейронная сеть для прогнозирования рака шейки матки у женщин.

Ключевые слова

Машинное обучение, нейронная сеть, параллельный генетический алгоритм, регрессия, корреляция, кластерный анализ

1. Введение

В последние годы все большее внимание уделяется медицинским исследованиям в сочетании с машинным обучением. Следовательно, все больший процент научно-исследовательских групп готовы сделать свои наборы данных общедоступными в Интернете. За исключением исходных данных, самое раннее исследование набора данных можно отследить до чувствительного к стоимости классификатора, точность которого в отношении данных шейки матки только что с небольшим отрывом превышала базовый уровень [1]. Вскоре для прогнозирования риска рака шейки матки были предложены два усовершенствованных подхода с использованием машины опорных векторов (SVM), включая устранение рекурсивных признаков SVM и анализ главных компонентов SVM (SVM-PCA) [2]. Недавно был представлен новый подход к выбору признаков под названием Firefly Algorithm вместе с классификатором случайного леса. Аналогичным образом Sherif [3] использовал метод синтетической передискретизации меньшинства (SMOTE) для уменьшения количества объектов на основе классификации случайного леса.

2. Предлагаемый метод реализации

Для решения задачи необходимо сформировать входные и выходные данные и выбрать нейронную сеть, аппроксимирующую функцию передачи. Многослойный персептрон также известен как искусственная нейронная сеть. Он имитирует принцип работы человеческого мозга с использованием базового нейронного блока. Несколько нейронов, связанных весами, образуют

MLP в определенных иерархических структурах. Предположим, что существует L скрытых слоев, и каждый скрытый слой имеет нейроны Li. Вес для каждого соединения обозначается как wljk, где j относится к j-му нейрону в (l - 1) -м слое, k относится к k-му нейрону в l-м слое, b представляет смещение. Вход можно рассматривать как 0-й скрытый слой, а выход можно рассматривать как (L + 1) -й скрытый слой.

Вес обновляется от конца модели к началу; поэтому их также называют сетью обратного распространения ошибок (BP). Алгоритм BP - это итеративный алгоритм, основанный на градиентном спуске.

Необходимо на основе входных данных сделать вывод об их значимости, то есть провести отбор значимых признаков методом параллельного генетического алгоритма, forward regression, backward regression, корреляционным методом. Далее необходимо провести обучение с нормированными значениями и оценить коэффициент корреляции и отношение стандартной ошибки к среднему. Обучение проводится пятью методами. Метод 1 – регрессия, метод 2 – корреляционный метод, метод 3 – алгоритм Левенберга – Марквардта, метод 4 – байесовская регуляризация обратного распространения, метод 5 – градиентный спуск.

В ходе применения генетического алгоритма результатом является то, что признак с наибольшим значением функции адекватности будет оптимальным, а входные признаки, которым будет соответствовать единичный ген особи ω, будут использоваться для построения регрессионного уравнения для *j*-го результативного признака [4].

Проведя отбор признаков, выявляем существенные для каждого результативного показателя параметры и проводим обучения различными методами для выявления наиболее точного метода обучения. Выборка делится на обучающую (80%) и тестовую (20%), проводится обучение и денормируются данные.

Для повышения достоверности результатов, в работе применяется кластерный анализ на основе метода k-mean. Далее обучение проводится для каждого кластера различными методами.

3. Заключение

В данной работе был проведен анализ существующих методов прогнозирования рака шейки матки, предложены различные методы машинного обучения и выбран наиболее оптимальный метод байесовской регуляризации обратного распространения. Далее был проведен отбор признаков с помощью различных методов и были получены результаты обучения по каждому методу и был выбран генетический алгоритм как лучший метод для факторов 1 и 3, а также метод forward regression для фактора 2.

4. Литература

[1] Fatlawi, H.L. Enhanced classification model for cervical cancer dataset based on cost sensitive classifier / H. Fatlawi // Int. J. Comput. Tech. – 2017. – Vol. 4(4). – P. 1-8.

[2] Wu, W. Data-driven diagnosis of cervical cancer with support vector machine-based approache / W. Wu, H. Zhou // IEEE Access 5. – 2017. – P. 25189-25195.

[3] Abdoh, S.F. Cervical cancer diagnosis using random forest classifier with SMOTE and feature reduction techniques / S. Abdoh, M. Rizka, F. Maghraby // IEEE Access 6. – 2018. – P. 59475-59785.

[4] Mokshin, V.V. A parallel genetic algorithm of feature selection for analysis of complex system / V.V. Mokshin, I.R. Saifudinov, P.I. Tutubalin, L. M. Sharnin // Proceedings of ITNT. – 2018. – P. 739-799.