Исследование методов машинного обучения для прогнозирования инсульта

А.Р. Фасхутдинова
Казанский национальный исследовательский технический университет им. А.Н. Туполева
Казань, Россия
l.akalova@yandex.ru

Б.А. Гарафутдинов
Казанский национальный исследовательский технический университет им. А.Н. Туполева
Казань, Россия
fenef0@gmail.com

Д.Н. Григорьева
Казанский национальный исследовательский технический университет им. А.Н. Туполева
Казань, Россия
Daragrigoreva@icloud.com

В.В. Мокшин
Казанский национальный исследовательский технический университет им. А.Н. Туполева
Казань, Россия
vladimir.mokshin@mail.ru

В этой статье рассматриваются методы прогнозирования инсульта. Было показано, что существуют разные методы решения проблемы. В статье представлено описание методов машинного обучения для прогнозирования вероятности развития инсульта. Система позволяет провести быструю диагностику данного заболевания на основе малого количества входных параметров.

Ключевые слова— нейросеть, метод k-ближайших соседей, прогнозирование, система, случайный лес, метод опорных векторов, диагностика заболеваний

1. Введение

Проблема инсульта в последние годы становится все более актуальной. Ежегодно в мире мозговой удар настигает более 15 млн человек. В России каждый год регистрируется более 500 тысяч случаев острых нарушений мозгового кровообращения. Инсульт молодеет в последние годы: не менее 20% нарушений кровообращения отмечаются у больных моложе 50 лет[1].

По исследованиям Всемирной организации здравоохранения (ВОЗ) инсульт занимает второе место в мире среди причин смертности [2].

Предупреждение развития инсульта представляется одной из важных клинических задач текущего времени, которое необходимо решать в условиях высокой загруженности специализированных стационаров и дефицита специалистов.

Целью данной работы является разработка модели, способной достаточно быстро и точно выявлять риск развития инсульта на основе малого количества входных параметров.

2. Методы реализации

В современной медицине принято проводить классификацию инсульта по таким параметрам, как механизм нарушения кровообращения, причины, вызвавшие это нарушение и др [3]. К основным факторам риска развития инсульта, которые можно рассматривать относительно всех видов данного заболевания, причисляют возраст, пол, курение, сахарный диабет (диагностируется на основании среднего уровня глюкозы в крови), ожирение, заболевания сердца, стресс. Именно эти факторы будут использоваться для отбора признаков при построении нейронной сети.

Проблема обучения на несбалансированных данных является достаточно распространенной темой для исследований последних лет [4]. Наличие данной проблемы было учтено при создании архитектуры нейронной сети. Эффективность использования искусственного интеллекта при оценке риска развития сердечно-сосудистых заболеваний была обоснована в работах [5-7].

Таким образом, для проектирования системы были использованы следующие входные параметры, оценивающие состояние диагностируемого: X1 – пол (0 – женский, 1 – мужской) , X2 – возраст, лет, X3 – наличие гипертонии (0 – нет, 1 – да), X4 – наличие сердечно-сосудистых заболеваний (0 – нет, 1 – да), X5 – факт вступления в брак (0 – не вступал(а) ранее, 1 – вступал(а)), X6 – тип профессиональной деятельности (0 – гос. служба, 1 – отсутствие опыта работы, 2 – частный предприниматель, 3 – самозанятый, 4 -ребенок) X7 – тип места проживания (0 – загородная территория, 1 – город), X8 – средний уровень глюкозы в крови пациента, мг/дл, X9 – ИМТ, кг/м^2, X10 – отношение к курению (0 – бросил(а) курить, 1 – никогда не курил(а), 2 – курит). Целевая переменная: D1 – перенесен ли инсульт, – принимает значения от 0 до 1, соответствующие вероятности развития инсульта для диагностированного, представленной в десятичной форме.

Для обучения модели множество было разделено на обучающее и тестирующее в соотношении 80% и 20% соответственно. Выборка состоит из 5100 наблюдений и взята из открытых источников (www.kaggle.com).

A.Метод k-ближайших соседей

Метод k-ближайших соседей используется для решения задачи классификации [8]. Он относит объекты к классу, которому принадлежит большее количество из k его ближайших соседей. Для улучшения результатов классификации вводят взвешивание примеров в зависимости от их удаленности.

$$Q_j = \sum_{i=1}^{n_j} \frac{1}{D^2(x, a_{ij})}$$

где D — оператор вычисления расстояния, x – вектор признаков классифицируемого объекта, a_{ij} – i-й пример j-го класса. Точность классификатора K ближайших соседей составляет 94,5 %.

B. Лес случайных решений

Лес случайных решений реализуется так: пусть обучающая выборка из N образцов, размерность пространства признаков равна M, и задан параметр m как неполное количество признаков для обучения. Наиболее распространенный способ построения деревьев ансамбля- бэггинг. Происходит генерация случайной повторной подвыборки размером N из обучающей выборки. Строится решающее дерево, классифицирующее образцы данной подвыборки. Дерево строится до полного исчерпания подвыборки.

Точность метода случайный лес решений составляет 99%.

C. Метод опорных векторов (SVM)

Метод опорных векторов (SVM) – один из самых популярных алгоритмов обучения с учителем, который используется как для задач классификации, так и для задач регрессии. Цель алгоритма SVM – создать наилучшую линию или границу решения, которая может разделить n-мерное пространство на классы, чтобы мы могли легко поместить новую точку данных в правильную категорию в будущем. Эта граница наилучшего решения называется гиперплоскостью.

В пространстве R^n уравнение

$$\langle \omega^{\rightarrow}, x^{\rightarrow} \rangle \, b = 0$$

при заданных $\omega^{\rightarrow}$ и b определяет гиперплоскость – множество векторов $x^{\rightarrow} = (x_1, ..., x_n)$, принадлежащих пространству меньшей размерности R^{n-1}. Например, для R^1 гиперплоскостью является точка, а для R^2-прямая, для R^3 - плоскость. Параметр $\omega^{\rightarrow}$ определяет вектор нормали к гиперплоскости, а через

$$\frac{b}{\| \vec{\omega} \|}$$

выражается расстояние от гиперплоскости до начала координат. Точность метода опорных векторов составляет 99% [9-10].

D. Метод Gradient Boosting

Gradient Boosting состоит из трех основных компонентов:

Роль функции потерь заключается в оценке того, насколько хорошо модель делает прогнозы с заданными данными. Это может варьироваться в зависимости от проблемы. Например, если мы пытаемся предсказать вес человека в зависимости от некоторых входных переменных (задача регрессии), то функция потерь поможет нам найти разницу между предсказанным весом и наблюдаемым весом [11].

Точность метода опорных векторов составляет 97%.

Изучив четыре метода реализации нейронной сети для прогнозирования инсульта, был выбран метод опорных векторов SVM.

Таблица I. Точность методов обучения

Метод	Точность метода %
k-ближайших соседей	94.5%
Лес решений	99%
Метод опорных векторов SVM	99%
Gradient Boosting	97%

3. Заключение

В данной работе был проведен анализ существующих методов прогнозирования инсульта, предложены различные методы машинного обучения и выбран наиболее оптимальный метод. Далее был проведен отбор признаков с помощью различных методов и были получены результаты обучения по каждому методу. Был выбран наиболее оптимальный метод искусственной нейронной сети для прогнозирования инсульта метод опорных векторов SVM.

Литература

[1] Статистика инсульта [Электронный ресурс]. – Режим доступа: https://яокб.рф/новости/публикации-специалистов/угрожающая-статистика-инсульта/

[2] Научный центр неврологии [Электронный ресурс].-Режим доступа:https://www.neurology.ru

[3] Фадеев, П.А. Инсульт / П.А. Фадеев – М.: Мир и Образование, Оникс, 2008. – 20 с.

[4] He, H. Learning from Imbalanced Data / H. He, A. Garcia // IEEE transactions on knowledge and data engineering.–2009. – Vol. 21(9). – P. 1263- 1284.

[5] Yasnitsky, L.N. Artificial Intelligence and Medicine: History, Current State, and Forecasts for the Future / L.N. Yasnitsky // Current Hypertensio Reviews. – 2020. – Vol. 16(3). P. 210-215. DOI : 10.2174/1573402116666200714150953 https://pubmed.ncbi.nlm.nih.gov/

[6] Yasnitsky, L.N. The capabilities of artificial intelligence to simulate the emergence and development of diseases, optimize prevention and treatment thereof, and identify new medical knowledge / L.N. Yasnitsky, A. Dumler, F.M. Cherepanov // Journal of Pharmaceutical Sciences and Research. – 2018. – Vol. 10(9). – С. 2192- 2200.

[7] Ясницкий, Л.Н. Нейросетевая система экспресс-диагностики сердечно-сосудистых заболеваний / Л.Н. Ясницкий, А.А. Думлер, А.Н. Полещук, К.В. Богданов, Ф.М. Черепанов // Пермский медицинский журнал. – 2011. – Т. 28, № 4. – С. 77-86

[8] Метод k-ближайших соседей [Электронный ресурс]. – Режим доступа: https://wiki.loginom.ru/articles/k-nearest-neighbor.html

[9] Mokshin, A.V. Adaptive genetic algorithms used to analyze behavior of complex system / A.V. Mokshin, V.V. Mokshin, L.M. Sharnin // Communications in Nonlinear Science and Numerical Simulation. – 2019. – Vol. 71. – P. 174–186.

[10] Tutubalin, P.I. The Evaluation of the cryptographic strength of asymmetric encryption algorithms / P.I. Tutubalin, V.V. Mokshin // Second Russia and Pacific Conference on Computer Technology and Applications. – 2017. – Vol. 1. – P. 180–183.

[11] Gradient Boosting In Classification: Not a Black Box Anymore [Electronic resource]. — Access mode: https://blog.paperspace.com/gradient-boosting-for-classification/