

С.Г. КЛИМАНОВ, Д.Д.ЦАРЕВА

Национальный исследовательский ядерный университет МИФИ, Москва, Россия

АНАЛИЗ ОБУЧАЕМОСТИ LSTM-МОДЕЛИ, НА ДАННЫХ СНЯТЫХ С ПАЦИЕНТА, И РАСЧЕТ ОБЪЕМА ОБУЧАЮЩЕЙ ВЫБОРКИ

Рассмотрена модель с долговременной и кратковременной памятью LSTM, на данных снятых с пациента. Проведена оценка обучаемости модели и её склонность к переобучению. Вычислен объем выборки, необходимый для дальнейшего обучения модели.

S.G. KLIMANOV, D.D. TSAREVA

National Research Nuclear University MEPHI (Moscow Engineering Physics Institute), Moscow, Russia

ANALYSIS OF LSTM MODEL TRAINABILITY ON PATIENT-DERIVED DATA AND CALCULATION OF REQUIRED TRAINING SAMPLE SIZE

A model with long- and short-term memory (LSTM) was tested using patient data. The model's learnability and susceptibility to overfitting were assessed. The sample size required for further training was calculated.

Одной из задач машинного обучения является проблема переобучения модели, приводящей к потере способности обобщать, т.е. модель запоминает «шум», а не необходимые для обучения данные. Также освещается вопрос об объеме обучающей выборки.

В работе для просчета необходимого объема выборки представлена модель с долговременной и кратковременной памятью LSTM, обучаемая на показателях состояния 1-го пациента. Возможность переобучения модели – одна из задач, для которой были вычислены минимальная тренировочная ошибка – наименьшее значение функции потерь (MSE), показывающая, что чем ниже ее значение, тем лучше модель предсказывает результат; минимальная валидационная ошибка, помогающая проверить производительность модели и ее способность обобщать; их разница – максимальный разрыв между ошибками, при больших и положительных значениях, указывающий на переобучение. На рис. 1 представлена динамика ошибок модели на тренировочном и валидационном наборах данных, ошибки постепенно снижаются, модель обучается. На рис.2 сначала видна необученность модели, однако с 10-го периода начинается переобучение.

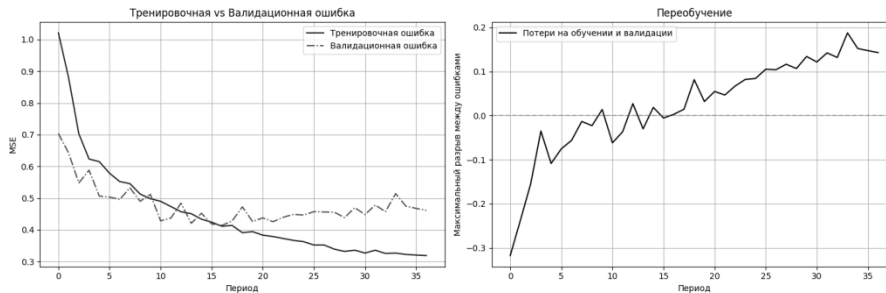


Рис. 1.

Рис.2.

Для определения объема выборки был использован двухвыборочный t-тест: $n=2*(Z_{1-\alpha/2}+Z_{\beta})^2/\delta^2$ [1], где n-объем выборки, $Z_{1-\alpha/2}$ –значение нормального распределения (уровень значимости $\alpha=0.05$), Z_{β} –квантиль для мощности $1-\beta$, $\beta=0.2$, δ –отношение разности средних значений на среднеквадратическое отклонение. В результате объем выборки равен 140 пациентам. Несмотря на умеренное переобучение с 20-го периода, модель успешно обучается.

Для корректной работы алгоритма объем выборки должен включать данные приблизительно от 140 пациентов.

Список литературы

1. Наркевич А.Н., Виноградов К.А. Методы определения минимально необходимого объема выборки в медицинских исследованиях / А.Н. Наркевич, К.А. Виноградов // Социальные аспекты здоровья населения [сетевое издание] – 2019. – 65(6):10.
2. B. Gholami, T.S. Phan, W.M. Haddad, A. Cason, J. Mullis, L. Price, J.M. Bailey. Replicating human expertise of mechanical ventilation waveform analysis in detecting patient ventilator cycling asynchrony using machine learning//Computers in Biology and Medicine. – 2018.
3. Hariharan, S., Karman, H. & Maheswari, D.U. Automated mechanical ventilator design and analysis using neural network// Sci Rep **15**, 3212 (2025)
4. Zeng Z., Hou Z., Li T., Deng L., Hou J., Huang X., Li J., Sun M., Wang Y., Wu Q., Zheng W., Jiang H., Wang Q. A Deep Learning Approach to Predicting Ventilator Parameters for Mechanically Ventilated Septic Patients. – 2021.
5. Yu, C., Liu, J. & Zhao, H. Inverse reinforcement learning for intelligent mechanical ventilation and sedative dosing in intensive care units// BMC Med Inform Decis Mak **19** (Suppl 2), 57 (2019).