Черепанов Ф.М. Выявление аномальных наблюдений в обучающем множестве посредством нейросетевой модели. // Проблемы информатики в образовании, управлении, экономике и технике: Сб. статей XIV Междунар. научно-техн. конф. – Пенза: ПДЗ, 2014. – С. 210-213.

УДК 004.8

ВЫЯВЛЕНИЕ АНОМАЛЬНЫХ НАБЛЮДЕНИЙ В ОБУЧАЮЩЕМ МНОЖЕСТВЕ ПОСРЕДСТВОМ НЕЙРОСЕТЕВОЙ МОДЕЛИ

Ф.М. Черепанов

IDENTIFICATION OF OUTLIERS IN THE TRAINING SET THROUGH NEURAL NETWORK MODELS

F.M. Cherepanov

Аннотация. Предложен способ выявления аномальных данных в обучающем множестве, основанный на сочетании нейросетевого моделирования и статистических методов анализа, позволяющий, в отличие от известных способов, учесть особенности нейросетевой модели, используемой для обработки этих данных.

Ключевые слова: нейронные сети, выбросы, анализ данных, аномальные наблюдения.

Abstract. We propose a method for identification of outliers in the training set, based on the combination of neural network modeling and statistical methods of analysis, which allows, in contrast to the known methods take into account the features of the neural network model used for the processing of these data.

Keywords: neural networks, outliners, data analysis, abnormal observations.

Во всём мире заболевания сердечнососудистой системы (ССС) являются одной из основных причин смертности населения и составляют до 50 %, что является серьёзной проблемой. Для успешного лечения заболеваний ССС необходимы точные методы дифференциальной диагностики с возможностью определения уровня тяжести.

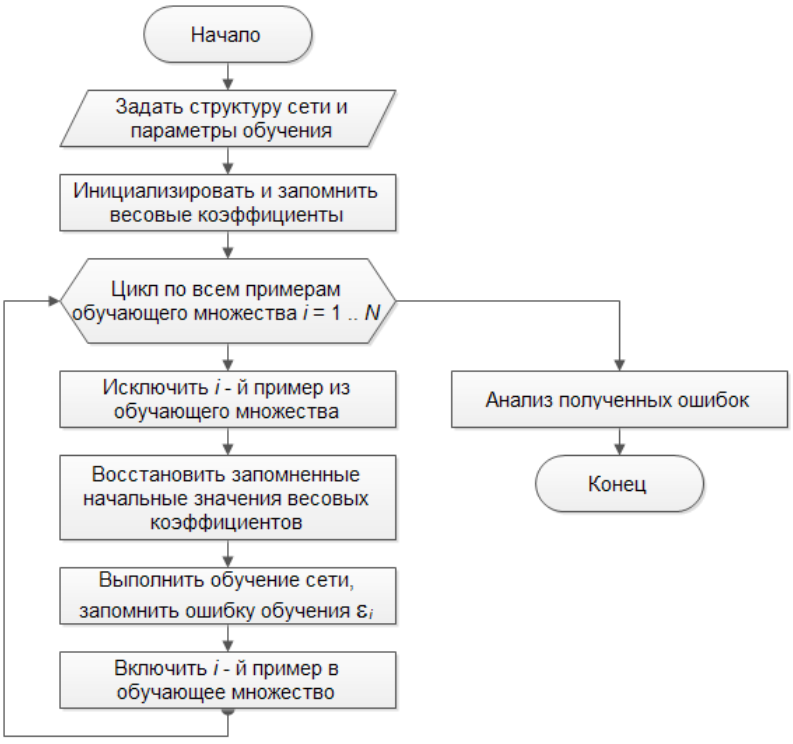
Для решения этой задачи наряду с разработкой новых приборов и методов традиционной медицинской диагностики всё активнее развиваются интеллектуальные диагностические системы, в том числе основанные на аппарате нейронных сетей, точность работы которых напрямую зависит от качества обрабатываемых эмпирических данных.

Присутствие аномальных наблюдений в статистических данных является весьма распространенным явлением, отрицательно влияющим на качество их последующей обработки и использования в качестве обучающего множества в нейросетевых системах. Если в простейших случаях выбросы обнаруживаются сравнительно легко и даже видны «на глаз», то в более сложных многомерных зависимостях при больших объемах информации выявление и исключение выбросов представляет собой непростую математическую проблему.

В регрессионном анализе для выявления выбросов или аномальных наблюдений применяются методы на основе анализа удаленных остатков и его модификации: стьюдентизированные остатки, DFFITS, стандартизованные DFFITS и др. Они определяются как стандартизованные остатки регрессионной модели,

т.е. разница между фактическим значением и прогнозируемым, для соответствующих наблюдений, полученных при поочередном исключении соответствующих наблюдений из анализа. При этом используются различные регрессионные модели, которые, однако, могут не соответствовать более точной модели той же предметной области, полученной при помощи нейронной сети.

В настоящей работе согласно [1, 2] развивается способ выявления аномальных данных в обучающем множестве, основанный на сочетании нейросетевого моделирования и статистических методов анализа, позволяющий, в отличие от известных способов, учесть преимущества нейросетевых моделей, используемых для обработки этих данных. Предлагаемый подход основан на методе анализа удаленных остатков, но вместо регрессионных моделей используются нейросетевые, и заключается в поочередном исключении примеров из обучающего множества, и последующей оценки изменения погрешностей искусственной нейронной сети, обученной на этих обучающих выборках, при прочих равных условиях. На рисунке представлена блок-схема общего алгоритма поочередного исключения примеров. На базе вышеприведенного можно реализовать ряд алгоритмов для выявления аномальных наблюдений, которые отличаются последним шагом, а именно – процедурой анализа влияния удаления каждого из примеров на показатели нейросетевой модели, полученные в результате работы алгоритма. Ниже приведены две такие процедуры – «Анализ ошибки обучения» и «Анализ суммарной ошибки».



Алгоритм выявления аномальных наблюдений

Анализ ошибки обучения основывается на том факте, что при отсутствии выброса в обучающем множестве сети будет легче выявить закономерности предметной области, она быстрее и лучше обучится, при этом среднеквадратичная ошибка, вычисляемая после завершения процесса обучения по формуле

$$E = \frac{1}{JQ} \sum_{q=1}^{Q} \sum_{j=1}^{J} (y_{qj} - d_{qj})^2 ,$$

будет меньше.

Здесь y_{qj} – значение j-го выхода ИНС для q-го обучающего примера; d_{qj} – желаемое значение j-го выхода для q-го обучающего примера; J – число нейронов в выходном слое; Q – количество примеров в обучающем множестве.

Как показал опыт [3–5], метод анализа ошибки обучения сети дает хорошие результаты на небольших обучающих множествах, в которых встречается не больше одного выброса. Для выборок, содержащих более ста элементов, эффективнее другой алгоритм, который назовем методом анализа суммарной ошибки. Идея этого метода основана на том факте, что если q-й пример является выбросом, то после обучения сети на обучающем множестве, не содержащем этого примера, её ошибка обобщения

$$e_q = \sum_{j=1}^{J} \left| y_{qj} - d_{qj} \right|$$

будет больше других. Если в обучающей выборке, помимо примера с номером q, имеются и другие выбросы, то, несмотря на это, ошибка для q-го аномального наблюдения по-прежнему будет больше ошибок примеров без аномалий.

Предложенные в работе методы обнаружения аномальных данных использовались при построении нейросетевой системы диагностики заболеваний ССС [5].

Библиографический список

1. Ясницкий Л.Н. Введение в искусственный интеллект. – М.: Академия, 2005. – 176 с.

2. Черепанов Ф.М., Ясницкий Л.Н. Нейросетевой фильтр для исключения выбросов в статистической информации // Вестник Пермского университета. Серия: Математика. Механика. Информатика. – 2008. – № 4. – С. 151–155.

3. Ясницкий Л.Н., Бондарь В.В., Полещук А.Н., Федорищев И.Ф., Черепанов Ф.М. и др. Пермская научная школа искусственного интеллекта и ее инновационные проекты. – М.; Ижевск: НИЦ «Регулярная и хаотическая динамика», 2008. – 75 с.

4. Ясницкий Л.Н., Данилевич Т.В. Современные проблемы науки. – М.: БИНОМ. Лаборатория знаний, 2008. – 294 с.

5. Yasnitsky L.N., Bogdanov K.V., Cherepanov F.M., Makurina T.V., Dumler A.A., Chugaynov S.V., Poleschuk A.N. Diagnosis and Prognosis of Cardiovascular Diseases on the Basis of Neural Networks // Biomedical Engineering. – 2013. – Т. 47. – № 3. – С. 160–163.

Черепанов Федор Михайлович
Пермский государственный гуманитарно-педагогический университет, г. Пермь, Россия
E-mail: e-mail: fe-c@yandex.ru

Cherepanov Fedor Mikhailovich
Perm State Humanitarian Pedagogical University, Perm, Russia