

Нейросетевой алгоритм выявления и удаления выбросов в зашумленных наборах данных

Л.Н. Ясницкий^{1,2*}, Е.Г. Плотникова²

¹Пермский государственный национальный исследовательский университет, Пермь, Россия

²Национальный исследовательский университет «Высшая школа экономики», Пермь, Россия
*yasn@psu.ru

Аннотация. Выбросы в статистических данных, являющиеся результатом ошибочно собранной информации, часто становятся препятствием для успешного применения метода моделирования, основанного на машинном обучении, во многих предметных областях. Наличие выбросов в наборах данных снижает точность моделей машинного обучения, а в некоторых случаях делает применение этих методов невозможным. Существующие в настоящее время методы выявления выбросов ненадежны. Они принципиально не способны выявлять некоторые типы выбросов, тогда как наблюдения, не являющиеся выбросами, эти методы нередко классифицируют как выбросы. Недавно появившиеся нейросетевые методы выявления выбросов лишены указанного недостатка, однако они не универсальны, поскольку способность нейросетей выявлять выбросы зависит как от архитектуры самой нейросети, так и от решаемой задачи. Цель настоящего исследования состоит в разработке алгоритма создания и применения нейронных сетей, способных правильно выявлять выбросы вне зависимости от решаемой задачи. Эта цель достигается благодаря использованию свойства некоторых, специально созданных нейронных сетей демонстрировать наибольшие ошибки обучения на тех наблюдениях, которые являются выбросами. Использование этого свойства, а также выполнение серии вычислительных экспериментов и обобщение их результатов с помощью математической формулы, являющейся модификацией следствия из теоремы Арнольда – Колмогорова – Хехт-Нильсена, позволили достичь поставленной цели. Применение разработанного алгоритма оказалось особенно эффективным при решении задач прогнозирования и управления взаимозависимыми теплофизическими и химико-энерготехнологическими процессами обработки рудного сырья, происходящими на действующих серийных металлургических предприятиях, где присутствие выбросов в статистических данных практически неизбежно, а без их выявления и исключения построение приемлемых по точности моделей машинного обучения вообще невозможно.

Ключевые слова: выброс в данных, теорема Арнольда – Колмогорова – Хехт-Нильсена, теплофизический и химико-энерготехнологический процессы, обработка рудного сырья, нейронная сеть, ошибка обучения

Для цитирования: Ясницкий Л.Н., Плотникова Е.Г. Нейросетевой алгоритм выявления и удаления выбросов в зашумленных наборах данных // Прикладная информатика. 2024. Т. 19. № 5. С. 88–100. DOI: 10.37791/2687-0649-2024-19-5-88-100

A neural network algorithm for identifying and removing outliers in noisy data sets

L. Yasnitsky^{1,2*}, E. Plotnikova²

¹Perm State National Research University, Perm, Russia

²National Research University Higher School of Economics, Perm, Russia
yasn@psu.ru

Abstract. Outliers in statistical data, which are the result of erroneously collected information, are often an obstacle to the successful application of machine learning methods in many subject areas. The presence of outliers in training data sets reduces the accuracy of machine learning models, and in some cases, makes the application of these methods impossible. Currently existing outlier detection methods are unreliable. They are fundamentally unable to detect some types of outliers, while observations that are not outliers are often classified as outliers by these methods. Recently emerging neural network methods for outlier detection are free from this drawback, but they are not universal, since the ability of neural networks to detect outliers depends both on the architecture of the neural network itself and on the problem being solved. The purpose of this study is to develop an algorithm for creating and using neural networks that can correctly detect outliers regardless of the problem being solved. This goal is achieved by using the property of some specially created neural networks to demonstrate the largest training errors on those observations that are outliers. The use of this property, as well as the implementation of a series of computational experiments and the generalization of their results using a mathematical formula, which is a modification of the consequence of the Arnold – Kolmogorov – Hecht-Nielsen theorem, made it possible to achieve the stated goal. The use of the developed algorithm turned out to be especially effective in solving the problems of forecasting and controlling interdependent thermophysical and chemical-energy-technological processes of processing ore raw materials, occurring at existing serial metallurgical enterprises, where the presence of outliers in statistical data is almost inevitable, and without their identification and exclusion, the construction of neural network systems that are acceptable in accuracy models are generally impossible.

Keywords: outlier in data, Arnold – Kolmogorov – Hecht-Nielsen theorem, thermophysical and chemical-energy-technological processes, neural network, training error

For citation: Yasnitsky L., Plotnikova E. A neural network algorithm for identifying and removing outliers in noisy data sets. *Prikladnaya informatika*=Journal of Applied Informatics, 2024, vol.19, no.5, pp.88-100 (in Russian). DOI: 10.37791/2687-0649-2024-19-5-88-100

Введение

Несмотря на впечатляющие успехи методов моделирования на основе машинного обучения, существует довольно большой пласт практически важных задач, для решения которых применение этих прогрессивных методов остается недоступным. Это задачи моделирования и управления сложными

взаимозависимыми теплофизическими и химико-энерготехнологическими процессами, встречающимися на серийных производственных предприятиях металлургического профиля. Сюда же относятся некоторые задачи медицинской диагностики, экономического моделирования и многие другие случаи, когда статистические выборки содержат большое количество выбросов.