ПРЕДОБРАБОТКА СТАТИСТИЧЕСКИХ ДАННЫХ ДЛЯ ПОВЫШЕНИЯ КАЧЕСТВА ПРОГНОЗА НЕЙРОННОЙ СЕТИ

PREPROCESSING OF STATISTICAL DATA TO IMPROVE THE QUALITY OF THE FORECAST BY A NEURAL NETWORK

A. Gilmanov
A. Gusev
A. Okunev

Summary. The article describes the method of functional preprocessing of statistical data to improve the forecast obtained with the help of neural networks. We consider a fairly wide range of functions that can be used to pre-process statistical data. The advantage of neural networks for forecasting using data preprocessing is shown, in terms of forecast stability. The forecast error is considered as a random variable for which: statistical estimates for the mathematical expectation and for the standard deviation are calculated, and a selective coefficient of variation is calculated to determine the most stable forecast model.

Keywords: functional preprocessing, forecast, stability of the neural network model, coefficient of variation.

Гильманов Артур Ринатович
Пермский государственный национальный исследовательский университет
arturinhog@yandex.ru
Гусев Андрей Леонидович
Д.т.н., К.ф.-м.н., профессор, Пермский государственный национальный исследовательский университет
alguseval@mail.ru
Окунев Александр Анатольевич
Аспирант, Пермский государственный национальный исследовательский университет
alexander2510@mail.ru

Аннотация. В статье описывается метод функциональной предобработки статистических данных для улучшения прогноза, получаемого с помощью нейронных сетей. Рассматривается достаточно широкий набор функций, который может быть использован для предобработки статистических данных. Показано преимущество нейронных сетей для прогноза с использованием предобработки данных, в смысле устойчивости прогноза. Ошибка прогноза рассматривается как случайная величина, для которой вычисляются: статистические оценки для математического ожидания и для стандартного отклонения, а также вычисляется выборочный коэффициент вариации для определения наиболее устойчивой модели прогноза.

Ключевые слова: функциональная предобработка, прогноз, устойчивость нейросетевой модели, коэффициент вариации.

В настоящее время в различных предметных областях задачи прогнозирования решаются с помощью прогнозирования многомерных временных рядов. Примерами таких областей являются экономика (прогнозирование макроэкономических показателей), медицина (показатели заболеваемости и смертности), промышленность (вибродиагностика) и т.д. Для решения подобных задач активно разрабатываются методы, использующие нейронные сети. Примерами могут служить методы, описанные в работах [1–4].

Существующие методы предполагают, что статистических данных достаточно много, чтобы построить качественную модель прогноза. Кроме того, эти методы ориентированы на прогнозирование величин только для одного объекта или одной территории. Но возможны ситуации, когда наблюдения ведутся для нескольких объектов в течение небольшого количества временных периодов. Решение подобных задач может быть выполнено с помощью метода экстраполирования ошибки нейронной сети, который рассмотрен в работах [5] и [6].

В методе экстраполирования ошибки нейросети задачу прогнозирования авторы настоящей статьи сформулировали следующим образом. Показатель y может быть спрогнозирован по известному определяющему вектору X и неизвестному определяющему Z. На первом этапе этого метода выполняется процедура сжатия информационного пространства. В результате этого исследователь получает сжатое множество наблюдений, на котором чаще всего можно построить приемлемую нейросетевую модель прогноза, то есть модель с удовлетворительной ошибкой прогноза. Получая ошибку прогноза на полном множестве наблюдений и экстраполируя ошибки нейросети для прогнозного периода, можно получить удовлетворительный общий прогноз, как сумму псевдо прогноза и экстраполированной ошибки нейронной сети.

Однако и этот метод экстраполирования ошибки не всегда дает положительные результаты в смысле общей средней ошибки прогноза. Что же можно предпринять в такой ситуации, когда прогнозирование необходимо по ряду объектов или территорий?

Как известно, корреляция между определяющими показателями и прогнозируемым показателем тесно связана с качеством прогнозирования при использовании регрессионных многомерных моделей. На практике при отборе наиболее значимых предикторов исследователи часто ориентируются на их коэффициенты корреляции с прогнозируемыми показателями. Исходя из этого, предположим, что выполнив определенную предобработку данных, можно увеличить коэффициенты корреляции между определяющими показателями и прогнозируемым показателем, если для модели прогноза используются нейронные сети. По естественному предположению это может привести к повышению качества прогнозирования, то есть к снижению общей средней ошибки прогноза. Подобный подход имеет смысл особенно в том случае, если статистические данные сильно зашумлены или их количество недостаточно для построения качественной неросетевой модели прогноза при помощи стандартных способов.

Рассмотрим функциональную предобработку данных, увеличивающую коэффициенты корреляции, которые авторы реализовали в виде программного продукта. Множество всех наблюдений предварительно делится на обучающее множество и тестирующее множество, согласно каноническому построению модели прогнозирования с помощью нейронных сетей. Только на обучающем множестве выполняется поиск оптимального способа предобработки с использованием следующих функций: x^{α}, $e^{\alpha x}$, $ln(\alpha x)$, $sin(\alpha x)$, $cos(\alpha x)$, $sh(\alpha x)$, $ch(\alpha x)$, $tg(\alpha x)$, $arcsin(\alpha x)$, $arccos(\alpha x)$, $arcsh(\alpha x)$, $arcch(\alpha x)$,$arctg(\alpha x)$, $arcth(\alpha x)$ и сигмоидная функция, где α принадлежит некоторому, заранее известному интервалу. Например, для сигмоиды α принадлежит интервалу от –10 до +10. Каждый интервал для α разбивается с некоторым шагом h, например, $h = 0{,}01$ или $h = 0{,}001$. Множество функций обозначим $F(\alpha, x)$. Пусть всего имеется N функций.

Алгоритм предобработки статистических данных заключается в следующем пошаговом итеративном процессе.

Шаг 1. Вычисляется корреляция (начальная корреляция) каждого из определяющих показателей x_i с прогнозируемым показателем y.

Шаг 2. Значения показателей x_i при помощи нормировки приводятся к интервалу [2; 102] или к интервалу [2; 3], что важно для использования всего спектра функций.

Шаг 3. Каждый определяющий показатель преобразуется каждой функцией из множества $F(\alpha, x)$. Получим N преобразований.

Шаг 4. Для каждого из N преобразований каждого определяющего показателя находится коэффициент корреляции с прогнозируемым показателем y.

Шаг 5. Для каждого x_i определяется функция преобразования из возможных N преобразований, для которой коэффициент корреляции по отношению к прогнозируемому показателю y будет наибольшим.

Шаг 6. Выполняется сравнение наибольшего коэффициента корреляции для конкретного определяющего показателя x_i с коэффициентом корреляции на предыдущей итерации алгоритма или начальным коэффициентом корреляции в случае первой итерации. Если коэффициент корреляции не увеличился или увеличился незначительно, то показатель исключается из дальнейшей предобработки. Незначительность определяется исследователем, например, может быть сравнение с величиной равной 0,001.

Шаг 7. Далее с помощью функции, определенной на 5 шаге, преобразуется x_i (т.е. происходит предобработка) и осуществляется переход к шагу 2 алгоритма. При этом считаем преобразование новым определяющим показателем.

Шаг 8. Алгоритм завершается, когда все x_i будут исключены из итерационного процесса на 6 шаге.

После выполнения алгоритма (найдены последовательности преобразующих функций для всех определяющих показателей из обучающего множества) в этой же последовательности преобразующие функции применяются к тем же показателям на тестирующем множестве. Далее следует построение модели прогноза по классической схеме для нейронных сетей. По сути, теперь исследователь имеет дело с одинаково преобразованными определяющими показателями, как в обучающем множестве, так и в тестирующем множестве.

В качестве примера рассмотрим прогнозирование путем построения нейронной сети с достаточно простой структурой.

Для проверки корректности метода предобработки данных из статистического справочника [7] был случайным образом выбран показатель «Заболеваемость», 5 показателей качества среды обитания и 4 показателя качества медицинского обслуживания. Все показатели анализировались за 10 лет (2005–2014 годы) по 77 территориям. Задача заключалось в том, чтобы по показателям качества среды обитания и качества медицинского обслуживания предыдущего года спрогнозировать показатель заболеваемости следующего года по 77 территориям. После сдвига данных с лагом в один год оказалось, что нейросетевую модель прогноза можно построить

по 8 годам (616 наблюдений — обучающее множество) и проверить её качество по 2014 году (77 наблюдений — тестирующее множество).

Определим, что под ошибкой прогноза будем подразумевать среднюю ошибку по территориям. А под ошибкой прогноза по территории будем подразумевать модуль разности между прогнозным и фактическим показателями заболеваемости по территории, разделенный на размах показателя заболеваемости и умноженный на 100%. Размах — разница между наибольшим и наименьшим значениями показателя заболеваемости за 8 лет (т.е. по обучающему множеству).

Сначала используем классический метод построения нейронной сети для прогноза на исходных статистических данных. Подберем архитектуру нейронной сети, показывающей наиболее стабильные результаты (в смысле результирующей ошибки). Такая нейронная сеть оказалась с достаточно простой структурой: 9 нейронов на входном слое, 2 нейрона на скрытом слое, 1 нейрон на выходном слое, активационные функции — гиперболический тангенс, применялся алгоритм обучения упругого распространения. Каждый раз при обучении нейронной сети проводилось 1000 итераций, а результат прогноза фиксировался в виде средней ошибки по 77 территориям. Количество таких экспериментов было равно 100. Среднюю ошибку можно рассматривать как случайную величину O_1, которая наблюдалась 100 раз. Эта случайная величина имела среднее выборочное равное $\overline{o}_1 = 4,34$ и выборочное стандартное отклонение равное $\sigma_{O_1} = 0,30$. Следовательно, выборочный коэффициент вариации этой случайной величины равен $v_{O_1} = 0,07$.

Теперь используем классический метод построения нейронной сети для прогноза на исходных данных с предобработкой. Подберем архитектуру нейронной сети, показывающей наиболее стабильные результаты (в смысле результирующей ошибки). Такая нейронная сеть оказалась с ещё более простой структурой: 9 нейронов на входном слое, 1 нейрон на скрытом слое, 1 нейрон на выходном слое, активационные функции — гиперболический тангенс, применялся алгоритм обучения упругого распространения. Каждый раз при обучении нейронной сети проводилось 1000 итераций, а результат прогноза фиксировался в виде средней ошибки по 77 территориям. Количество таких экспериментов было также равно 100. Среднюю ошибку вновь можно рассматривать как случайную величину O_2, которая наблюдалась 100 раз. Эта случайная величина имела среднее выборочное равное $\overline{o}_2 = 2,95$ и выборочное стандартное отклонение равное $\sigma_{O_2} = 0,19$. Следовательно, выборочный коэффициент вариации этой случайной величины равен $v_{O_2} = 0,06$.

В результате предобработки ошибка на тестирующем множестве в среднем снизилась с 4,34% до 2,95%, то есть на 1,39%. Этот факт позволяет говорить о корректности метода предобработки исходных статистических данных, позволяющим снизить общую среднюю ошибку прогноза. Причем, сравнение коэффициентов вариации средней ошибки нейронной сети позволяет утверждать, что предобработка статистических исходных данных обеспечивает большую устойчивость прогноза.

Вывод: рассмотренный метод позволяет уменьшать ошибку прогнозирования при условии недостаточности и/или зашумленности статистических данных. Также метод может быть использован, когда средняя ошибка прогноза при применении стандартных методов прогнозирования не устраивает исследователя.

ЛИТЕРАТУРА

1. Ruta D., Gabrys B. Neural Network Ensembles for Time Series Prediction // Neural Networks, 2007. IJCNN2007. International Joint Conference on. 2007. pp. 1204–1209.
2. Collotta M., Pau G. An Innovative Approach for Forecasting of Energy Requirements to Improve a Smart Home Management System Based on BLE // IEEE Transactions on Green Communications and Networking, 2017. pp. 112–120.
3. Nguyen, H.H., Chan, C. W. Multiple neural networks for a long term time series forecast // Neural Computing & Applications, Vol. 13, No. 1, 2004. pp. 90–98.
4. Abdoli A.M., Nezhad M. F., Sede R. S., Behboudian S. Longterm forecasting of solid waste generation by the artificial neural networks // Environmental Progress & Sustainable Energy, Vol. 31, No. 4, 2012. pp. 68–636.
5. Gusev A.L., Okunev A. A. Forecasting with incomplete set of factors determining the predicted factor. Neural network error extrapolation method // International Journal of Applied Mathematics and Statistics, Vol. 56, No. 5, 2017. pp. 48–52.
6. Гусев А.Л., Окунев А. А. Методы сжатия информационного пространства при прогнозировании в условиях неполноты информации // Материалы XV Всероссийской научной конференций «Нейрокомпьютеры и их применение». Москва. 2017. С. 190–191.
7. Регионы России. Социально-экономические показатели 2015. Статистический сборник. Росстат, 2015.

© Гильманов Артур Ринатович (arturinhog@yandex.ru),
Гусев Андрей Леонидович (alguseval@mail.ru), Окунев Александр Анатольевич (alexander2510@mail.ru).
Журнал «Современная наука: актуальные проблемы теории и практики»