DOI: 10.15593/2499-9873/2020.3.03
УДК 519.2+51-74

А.А. Окунев

Пермский государственный национальный
исследовательский университет, Пермь, Россия

ИСПОЛЬЗОВАНИЕ ФУНКЦИОНАЛЬНОЙ ПРЕДОБРАБОТКИ ДАННЫХ ПРИ ПРОГНОЗИРОВАНИИ ПАРАМЕТРОВ ВИБРАЦИИ НЕФТЕПЕРЕКАЧИВАЮЩИХ АГРЕГАТОВ

Посвящена алгоритму функциональной предобработки данных, который может быть использован для уменьшения результирующей ошибки при решении задач прогнозирования с помощью построения нейросетевых моделей. Описанный в статье алгоритм был разработан в рамках построения системы прогнозирования параметров вибрации нефтеперекачивающих агрегатов, применяемой для прогнозирования развития дефектов.

Автор анализирует существующие подходы к вибродиагностике и приходит к необходимости рассмотрения поставленной задачи как задачи долгосрочного прогнозирования, а не задачи классификации, как это принято при решении аналогичных задач. Причина данного решения заключается в отсутствии размеченных данных.

Основные идеи решения задачи долгосрочного прогнозирования следующие: нейросетевая модель принимает на вход и выдает на выход характеристики измеряемых величин в периоды времени, время разбивается на периоды, для каждого периода строится своя нейросетевая модель, причем период следующей шкалы в целое число раз больше периода предыдущей шкалы, шкалы с меньшим периодом используются для краткосрочного прогнозирования, а с большим – для долгосрочного.

Для повышения качества прогнозирования применяется функциональная предобработка данных. Она заключается в том, что построенная по алгоритму последовательность функций применяется к входу модели прогнозирования, чтобы повысить коэффициент корреляции между входом и выходом.

Поскольку наблюдаемые временные ряды нестационарны, возможны изменения распределений измеряемых величин и видов зависимостей между ними. Следовательно, исходный алгоритм предобработки был модифицирован: в него добавлены шаги, обеспечивающие устойчивость предобработки (минимизируется разница результатов ее работы на разных множествах).

Устойчивость обеспечивается с помощью двух вариантов предварительного отбора функций предобработки. Первый из них заключается в том, что его проходят функции, для которых разница коэффициентов корреляции между входом и выходом модели на двух непересекающихся подмножествах обучающего множества минимальна. Второй вариант отбора заключается в том, что его проходят функции, повышающие коэффициент корреляции на обоих подмножествах.

Проведенные на данных с двух агрегатов эксперименты показали, что применение функциональной предобработки данных в подавляющем большинстве случаев приводит к уменьшению ошибки прогнозирования. Более чем в половине случаев применение модификации алгоритма, обеспечивающей устойчивость, позволяет получить меньшую ошибку на тестирующем множестве, чем при использовании исходного алгоритма.

Ключевые слова: нейронные сети, прогнозирование, предобработка данных, вибрационная диагностика, математическое моделирование, машинное обучение, интеллектуальные системы, дефекты промышленного оборудования, многомерные временные ряды, построение устойчивых моделей.

A.A. Okunev

Perm State University, Perm, Russian Federation

FUNCTIONAL DATA PREPROCESSING APPLICATION TO OIL-TRANSFER PUMPS VIBRATION PARAMETERS FORECASTING

This work describes functional data preprocessing algorithm. This algorithm provides a way to reduce error in forecasting problems solution. The algorithm is a part of oil-transfer pumps vibration parameters forecasting system that enables pump failures dynamics forecasting.

The author analyses existing approaches to vibration monitoring and decides to solve failure-forecasting problem as a long-term forecasting problem although researchers usually solve such problems with classification methods. An insufficiency of labeled data is the main reason of such a decision.

Main ideas of the problem solution are the following. Neural network model takes and calculates periodical metered values characteristics. Time is split into periods using scales with different periods. We use shorter periods for short-term forecasting and longer periods for long-term forecasting.

Functional data preprocessing provides a way to increase forecasting quality. Preprocessing key idea is following. Functions sequence transforms one of model's inputs in order to increase correlation between input and output.

Metered values distributions and dependencies between values can be variant because of observed time series nonstationarity. Author decided to modify original preprocessing algorithm to solve a nonstationarity problem.

Idea of the modification is to add steps that provide preprocessing robustness i.e. allow to reduce difference between preprocessing results on different datasets. Preliminary preprocessing functions selection provides robustness. There are two variants of preliminary selection. The first one is following: function with the least difference between correlations between input and output in data subsets pass the selection. The second one is following: functions that increase correlation on both subsets pass the selection.

Experiments on two pumps data prove the hypothesis that data preprocessing in vast majority of cases allows to decrease forecasting error. Modified algorithm often has less test error than original one.

Keywords: neural networks, forecasting, data preprocessing, vibrational diagnostics, mathematic modeling, machine learning, artificial intelligence, equipment failures, multivariate time-series, robust models building.

Введение

В процессе обслуживания промышленного оборудования необходима возможность точного определения его состояния, в том числе зарождающихся дефектов, степени изношенности и деградации. Данную возможность обеспечивает мониторинг состояния машин, заключающийся в непрерывном сборе и анализе данных о состоянии компонентов оборудования [1]. Мониторинг состояния машин позволяет осуществить переход от устранения негативных последствий сбоев к обслуживанию по состоянию [2]. В обеих приведенных работах определение состояния оборудования выполняется интеллектуальными системами, построенными на основе нейронных сетей.

Анализ параметров вибрации, заключающийся в извлечении необходимой информации из вибрационных сигналов, является одним из основных подходов к диагностике роторных машин [3].

Рассмотрим решение задачи прогнозирования развития дефектов нефтеперекачивающих агрегатов на основе их параметров вибрации. Задача была решена в ходе работ по построению системы вибродиагностики в рамках НИОКР по заказу ООО НПП «ТИК». Требование заказчика заключалось в том, что данная система на основе исторических данных должна быть способна определять состояние оборудования и прогнозировать динамику изменения его состояния. Идеи, лежащие в основе решения данной задачи, представлены в работе [4].

1. Описание контролируемой системы

Контролируемая система представляет собой множество агрегатов, включающих в себя двигатель, насос и другие элементы. К каждому элементу агрегата подключаются датчики, снимающие вибрационный сигнал.

Значения виброускорения снимаются с заданным шагом (10 с – 10 мин) в течение 2 с, после чего выполняется аналого-цифровое преобразование сигнала с частотой 30 кГц. Среднеквадратичные значения (СКЗ) виброскорости и виброперемещения вычисляются путем аппаратного интегрирования с шагом 2 с.

Кроме СКЗ виброскорости, виброускорения и виброперемещения на основе снятого сигнала программным путем вычисляются применяемые для диагностики величины – контрольные показатели (КП), например частота проявления тел качения, частота проявления рабочего колеса машины, гармоники оборотной части и т.д.

2. Анализ существующих подходов к решению задач вибродиагностики

В работах [1] и [2] описаны системы вибродиагностики, работающие со спектрами сигналов. Эти системы основаны на нейронных сетях ARTMAP. В работе [3] рассмотрено распознавание дефектов подшипников роторной машины с использованием многослойных персептронов. В работе [5] представлена система, которая позволяет при помощи формирования трендов вейвлет-коэффициентов на основе АЧВС вибрационных сигналов определять факторы вибрации, вызывающие возрастание коэффициентов полос спектра.

Общими чертами приведенных выше подходов является то, что исследователи при обучении опираются на достаточные по объему и размеченные экспертами наборы данных и используют алгоритмы, позволяющие строить модели с дискретными выходами, т.е. модели, подходящие только для решения задач классификации.

Подходы, в основе которых лежит предположение о наличии размеченного набора данных достаточного объема, не могут быть применены для решения рассматриваемой задачи вследствие отсутствия у заказчика подобного набора данных. Основная причина отсутствия такого набора данных заключается в том, что разметка данных экспертами является дорогостоящей процедурой, так как она требует остановки, разбора и простоя агрегата.

3. Переход к задаче прогнозирования

Ввиду того, что поставленную задачу не удалось свести к задаче классификации, было решено воспользоваться эвристикой: рост СКЗ виброскорости напрямую связан с развитием дефектов, а превышение им заданного экспертами и(или) нормативами предельного значения (уставки) является признаком критического состояния агрегата [1].

В ходе решения задачи автор делает предположение, что СКЗ виброскорости может быть спрогнозировано по предыдущим и текущим значениям СКЗ и КП, вычисленным на основе снятого вибрационного сигнала. Таким образом, поставленная изначально задача вибродиагностики сводится к задаче определения времени превышения уставки, для решения которой обучается модель прогнозирования СКЗ по предыдущим и текущим СКЗ и КП – модель прогнозирования значений одной из компонент многомерного временного ряда.

Временной ряд $F = \left\{ \left(CI_t^1,\ CI_t^2,\ \ldots,\ CI_t^m,\ \mathrm{RMSA}_t,\ \mathrm{RMSV}_t,\ \mathrm{RMSS}_t \right) : t \geq 0 \right\}$, где $CI_t^1,\ CI_t^2,\ \ldots,\ CI_t^m$ – значения контрольных показателей с номерами 1,2, …, m в моменты времени t; RMSA_t, RMSV_t, RMSS_t – СКЗ виброускорения, виброскорости, виброперемещения в момент времени t. Необходимо найти функцию f, такую, что выполняется равенство $\mathrm{RMSV}_{t+k} = f\left(CI_t^1, \ldots, CI_t^m, \mathrm{RMSA}_t, \mathrm{RMSV}_t, \mathrm{RMSS}_t, \ldots, CI_{t-L}^1, \ldots, \mathrm{RMSS}_{t-L} \right)$, где $L \geq 0, k > 0$.

4. Анализ существующих подходов к решению задач прогнозирования временных рядов

Традиционный подход к прогнозированию многомерных временных рядов при помощи нейронных сетей сводится к тому, что при обучении нейронной сети ей на вход подаются предыдущие и текущие значения компонент рассматриваемого временного ряда, а ее выходным значением является значение той компоненты временного ряда, для которой строится прогноз. Для прогнозирования на N шагов вперед строится N нейронных сетей. Проблема заключается в том, что начиная с некоторого N качество прогнозирования становится неприемлемым.

В статье [6] рассмотрен подход к долгосрочному прогнозированию, основанный на выделении тренда, предобработке данных и прогнозировании подаваемых на вход модели прогнозирования значений компонент рассматриваемого многомерного временного ряда. Необходимо, чтобы значения, подаваемые на вход модели, могли быть прогнозируемы с достаточно малой ошибкой для большого числа будущих периодов. Данный подход в описанном в работе виде неприменим для представленной в данной статье задачи, так как последовательности СКЗ и КП немонотонны и зашумлены, но можно применить идею предобработки данных с целью уменьшения ошибки прогнозирования.

В работах [7–10], посвященных прогнозированию одномерных временных рядов, утверждается, что применение предобработки данных, позволяющей удалить тренд и сезонность, позволяет уменьшить ошибку прогнозирования. Подходы, описанные в указанных статьях, неприменимы для решения задачи, которой посвящена настоящая статья, так как, как было сказано ранее, последовательности СКЗ и КП не являются ни монотонными, ни периодическими.

В статье [11] описан такой метод долгосрочного прогнозирования одномерных временных рядов, как множественная нейросетевая модель. Такая модель включает в себя определенное количество нейронных сетей, каждая из которых отвечает за прогнозирование значений временного ряда на заданное количество шагов вперед. Затем ее выходное значение добавляется к уже известным значениям ряда и полученные значения подаются на вход следующей нейронной сети. Данный метод применим только для одномерных временных рядов, так как в случае многомерных рядов необходимо решать задачу прогнозирования для каждой компоненты ряда, что приводит к накоплению и увеличению ошибки.

В статье [12] представлен алгоритм построения модели, основанной на нейронных сетях, позволяющей прогнозировать значения многомерных временных рядов на один шаг вперед, причем учитывается, что ряды могут быть зашумленными. Для уменьшения ошибки прогнозирования и для обеспечения устойчивости авторы данной работы обучают сети различной структуры на различных фрагментах обучающего множества и объединяют лучшие нейронные сети (среди дающих разные выходные значения при одинаковых входных значениях) в ансамбль. Описанный метод не может быть применен для решения рассматриваемой в настоящей статье задачи вследствие невозможности использования нескольких фрагментов при достаточно большой длительности периода.

В работе [13] строится ансамбль нейронных сетей, причем во всех сетях лаг между значениями определяющих и прогнозируемого показателей различен. Данный подход неприменим для рассматриваемой задачи ввиду отсутствия возможности использования достаточно большого количества периодов для построения каждой из сетей.

5. Построение модели прогнозирования

Условимся называть выход модели прогнозируемым показателем, а входы, которые вычисляются на основе СКЗ и КП, – определяющими. Горизонт прогнозирования решено принять равным суткам. Будем считать, что показатель превышает уставку, если его среднее значение за заданный промежуток времени больше, чем уставка.

Значение прогнозируемого показателя в будущие периоды времени вычисляется несколькими нейросетевыми моделями, каждая из которых соответствует отдельному периоду, т.е. выходы моделей составляют непрерывный фрагмент ряда значений прогнозируемого показателя.

Обучающая выборка формируется из данных о работе агрегатов, находящихся на предприятиях, и о работе стенда, на котором имитируется развитие дефекта начиная от нормального состояния до критического.

6. Многократное разбиение времени

Поскольку компоненты рассматриваемого многомерного временного ряда зашумлены из-за особенностей наблюдаемого процесса, а прогнозирование является долгосрочным и уставка должна быть превышена на протяжении целого периода, может быть применен такой подход, как многократное разбиение времени с различными шкалами. Он заключает-

ся в том, что время разбивается на равные периоды, причем период каждой следующей шкалы больше периода предыдущей в целое число раз. Шкалы с меньшим периодом могут быть применены для получения краткосрочного прогноза, с большим – для долгосрочного.

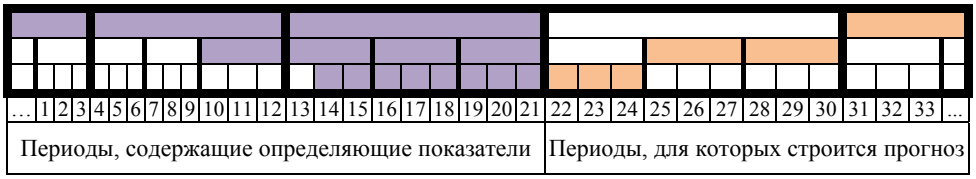

Рис. 1. Многократное разбиение периодов времени

На рис. 1 изображено разбиение времени с тремя шкалами и отмечено, значения показателей за какие периоды подаются на вход модели и значения в какие будущие периоды эта модель позволяет прогнозировать.

Кроме того, необходимо разбивать временную шкалу на периоды так, чтобы они могли пересекаться и становиться длиннее, например период вместо 60 мин становится равным 75 мин. Данная модификация необходима для того, чтобы можно было учесть возможные локальные экстремумы на границах периодов.

Как говорилось ранее, прогнозирование выполняется несколькими моделями, причем каждая модель соответствует одному из будущих периодов. Необходимо отметить, что периоды, для которых строится прогноз, могут находиться на разных временных шкалах (см. рис. 1).

7. Вычисление определяющих и прогнозируемых показателей

Важной деталью является то, что модель прогнозирования принимает на вход не сами значения СКЗ и КП, а их характеристики, вычисленные в периоды времени. Следовательно, для каждого СКЗ и КП в конкретный период вычисляются среднее значение за период, среднее ± стандартное отклонение, среднее и максимальное скользящие средние за период. Применение скользящего среднего какого-либо СКЗ или КП с периодом, меньшим, чем период временной шкалы, дает возможность исключить резкие скачки, учесть тренды и экстремумы на границах периодов.

Пусть имеется величина (СКЗ или КП) V^k, для нее одна из характеристик вычисляется с помощью функции f_j. Тогда соответствующая компонента временного ряда имеет вид $g_j\left(V_t^k\right), g_j\left(V_{t-1}^k\right), \ldots, g_j\left(V_{t-L}^k\right)$, X_t^k – значения величины в текущий период, V_{t-l}^k, $l > 0$, – в предыдущие. Пусть имеется K величин, J функций вычисления характеристик величин, для прогнозирования используются данные за L прошлых периодов, тогда имеется $N^0 = KJ\left(L+1\right)$ определяющих показателей вида $g_j\left(V_{t-l}^k\right)$, где $1 \le j \le J$, $1 \le k \le K$, $0 \le l \le L$. Обозначим определяющие показатели как x_i, где $1 \le i \le N^0$, а прогнозируемый показатель для рассматриваемого будущего периода обозначим как y. Здесь и далее будет рассматриваться построение модели прогнозирования для одного заданного будущего периода.

В качестве прогнозируемого показателя было решено взять максимальное скользящее среднее СКЗ виброскорости за период, для которого строится прогноз. Это решение обосновывается тем, что основной задачей является прогнозирование момента превышения уставки и она должна быть превышена в течение периода времени, равного периоду скользящего среднего.

8. Отбор определяющих показателей модели

Известно полное множество определяющих показателей $X^0 = \left\{x_i,\ i \in \left[1; N^0\right]\right\}$. До построения модели необходимо выполнить отбор определяющих показателей и тем самым определить, какие из них несут полезную информацию.

Перед тем как начать его выполнение, отметим, что множество отобранных показателей пустое: $X = \varnothing$, по окончании процедуры отбора оно будет включать в себя N определяющих показателей, причем N неизвестно заранее.

Для каждого показателя x_i введем индикаторную переменную I_i, равную 1 в том случае, если данный показатель рассматривался для добавления, и 0 в противном случае.

Процедура отбора показателей включает в себя следующие шаги:

1. Исключить из дальнейшего рассмотрения определяющие показатели, такие, что $\sigma_j = 0$, где σ_j – стандартное отклонение показателя x_j.

2. Для каждого определяющего показателя x_j вычислить коэффициент корреляции r_j с прогнозируемым показателем *y*.

3. Исключить из дальнейшего рассмотрения определяющие показатели, такие, что $r_j < (\overline{r} - \sigma_r)$, где $\overline{r}$ – средний коэффициент корреляции, σ_r – стандартное отклонение.

4. Выбрать определяющий показатель с максимальным коэффициентом корреляции: $X = \left\{ x_j : j = \operatorname{argmax}(r_k), k \in \left[1; N^0\right] \right\}$ (на данном этапе множество *X* включает в себя один показатель и будет дополняться в дальнейшем).

5. Обучить нейронную сеть, на вход которой подаются значения показателей из *X*, найти ошибку обучения ε.

6. Из определяющих показателей выбрать самый некоррелированный с уже выбранными, т.е. выбрать определяющий показатель x_p такой, что $p = \operatorname{argmin}\left(\max\left(r_{ps}\right)\right)$, где r_{ps} – коэффициент корреляции между показателями x_p и x_s, $x_s \in X, x_p \notin X, I_p = 0$.

7. Обучить нейронную сеть, которая принимает на вход значения показателей из множества $X^+ = X \cup \left\{x_p\right\}$, найти ошибку обучения ε^+.

8. Если $\varepsilon^+ < \varepsilon$, принять $X = X^+$.

9. Принять $I_p = 1$.

10. Если остались нерассмотренные показатели $(\exists i\, I_i = 0)$, вернуться к п. 5.

9. Функциональная предобработка данных

В работе [6] сформулирована идея, применимая для решения данной задачи: использование функциональной предобработки данных может повышать качество прогнозирования. Такая же идея формулируется в работе [14], в которой приведен исходный алгоритм функциональной предобработки данных, и в последующей работе [15].

В предыдущих версиях алгоритма делаются следующие предположения:

1. Данные однородны в плане распределения наблюдаемых величин и вида зависимостей между ними.

2. Функции, с помощью которых выполняется предобработка, определяются на всем обучающем множестве и так же хорошо подходят для тестирующего множества.

Поскольку наблюдаемые временные ряды нестационарны, возможны ситуации, когда в течение периода, на котором формируется обучающее множество, меняется распределение одной из компонент ряда и/или вид зависимостей между ними. Следовательно, необходимо модифицировать исходный алгоритм и сделать предобработку более устойчивой, т.е. разница результатов ее работы на разных множествах должна быть минимальной. Для этого вводится такая модификация алгоритма, как дополнительный шаг, на котором выполняется предварительный отбор функций, после которого уже выбирается одна лучшая функция (заключительный шаг из исходного алгоритма). Методы предварительного отбора, позволяющего сделать предобработку более устойчивой, будут рассмотрены далее.

10. Используемые функции и их параметры

Для предобработки данных используются семейства функций: x^{α}, $e^{\alpha x}$, $\sin(\alpha x)$, $\cos(\alpha x)$, $\mathrm{ch}(\alpha x)$, $\mathrm{sh}(\alpha x)$, $\mathrm{th}(\alpha x)$, $\mathrm{arctg}(\alpha x)$, $\ln(x+\alpha)$, $\frac{1}{\left(1+e^{-\alpha x}\right)}$, $\frac{1}{\alpha(x+\beta)}$, где α и β – параметры. Обозначим множество семейств как *G*, семейства будем обозначать как g_j, для каждого семейства g_j $\left(g_j \in G, j \in \left[1;|G|\right]\right)$ задается множество параметров (векторов параметров, если их несколько) A_j.

Рассмотрим выбор значений параметра для семейства функций $g_j(\alpha, x)$ с одним параметром. Для семейств с двумя и более параметрами правила аналогичны. Итак, при выборе значений параметра необходимо обратить внимание на близость и сходство графиков функций $g_j(\alpha_1, x)$ и $g_j(\alpha_2, x)$ и на их наклоны на отрезке [–1; 1].

Необходимо рассматривать только те функции $g_j(\alpha_1, x)$ и $g_j(\alpha_2, x)$, соответствующие значениям параметров α_1 и α_2, при которых графики функций $g_j(\alpha_1, x)$ и $g_j(\alpha_2, x)$ отличаются друг от друга

в достаточной степени, т.е. средняя разность функций в точках из отрезка [–1; 1] была не меньше наперед заданной величины. Данная величина была принята равной 0,05–0,1.

Если на некотором отрезке, лежащем внутри отрезка [–1; 1] и имеющем длину, не меньшую заданной, значения функций близки к некоторой константе, то такую функцию необходимо исключить из рассмотрения, так как она делает различные значения показателей практически равными друг другу.

Отрезок [–1; 1] используется для единообразия (нейронные сети принимают на вход и выдают на выход значения из этого промежутка) и простоты подбора параметров. Благодаря переходу к этому отрезку можно подобрать параметры таких функций, как логарифм или обратная пропорциональность, так, чтобы они были определены на всем отрезке и в окрестностях его границ. Определенность в окрестностях границ за пределами отрезка необходима, так как на тестирующем множестве возможна ситуация, когда нормированный результат предобработки по модулю превысит 1. Исходя из практических соображений, было решено допустить выход за пределы на 20 % от длины отрезка.

Следуя указанным правилам, проанализировали различные значения α для всех вышеуказанных функций.

Для x^{α} следует взять значения параметра: 1/7; 1/5; 1/3; 2; 3.

Для $e^{\alpha x}$ следует взять значения параметра: 1,7; 1,6; 1,5; 1,4; 1,3; 1,2; 1,1; 0,95; 0,8; 0,65; 0,5; 0,35; 0,2 (и эти же значения с противоположным знаком).

Для $\cos(\alpha x)$ следует взять значения параметра: 3; 2,8; 2,65; 2,5; 2,35; 2,2; 2,05; 1,9; 1,75; 1,6; 1,45; 1,3; 1,15; 0,95; 0,7.

Для $\sin(\alpha x)$ следует взять значения параметра: 3; 2,8; 2,6; 2,3; 2,05; 1,75; 1,45; 1,25; 1,1; 0,95; 0,8; 0,65; 0,5; 0,38; 0,27; 0,15 (и эти же значения с противоположным знаком).

Для $\mathrm{ch}(\alpha x)$ следует взять значения параметра: 2; 1,9; 1,8; 1,7; 1,6; 1,5; 1,4; 1,3; 1,2; 1; 0,8; 0,6.

Для $\mathrm{sh}(\alpha x)$ следует взять значения параметра: 2; 1,9; 1,8; 1,7; 1,55; 1,4; 1,25; 1,1; 0,95; 0,75; 0,55; 0,35; 0,15 (и эти же значения с противоположным знаком).

Для $\mathrm{th}(\alpha x)$ следует взять значения параметра: 3; 2,2; 1,7; 1,35; 1,1; 0,9; 0,7; 0,55; 0,4; 0,25; 0,1 (и эти же значения с противоположным знаком).

Для arctg(αx) следует взять значения параметра: 3; 2,5; 2,1; 1,7; 1,4; 1,15; 0,95; 0,75; 0,55; 0,4; 0,25; 0,1 (и эти же значения с противоположным знаком).

Для $\ln(x+\alpha)$ следует взять значения параметра: 1,2; 1,25; 1,3; 1,35; 1,4; 1,45; 1,5; 1,6; 1,7; 1,8; 1,9; 2; 2,1; 2,2; 2,3; 2,4; 2,5; 2,6; 2,7; 2,8; 2,9; 3; 3,1; 3,2; 3,3; 3,4; 3,5.

Для $\frac{1}{e^{-\alpha x}+1}$ следует взять значения параметра: 4; 2,5; 1,7; 1,1; 0,5 (и эти же значения с противоположным знаком).

Для $\frac{1}{\alpha(x+\beta)}$ следует взять значения параметра α:

1) 0,6; 0,7; 0,9; 1,1; 1,3; 1,6; 2; 2,5; 3; 4; 5; 6; 7; 8 (при β = 1,1);

2) 0,6; 0,65; 0,7; 0,75; 0,8; 0,9; 1; 1,1; 1,2; 1,3; 1,4; 1,6; 1,8; 2; 2,5 (при β = 1,5);

3) 0,2; 0,25; 0,3; 0,35; 0,4; 0,45; 0,5; 0,55; 0,6; 0,7; 0,8; 0,9; 1,1; 1,2 (при β = 2).

Множество, включающее в себя все возможные преобразования, обозначим как G_0, $G_0 = \left\{ g_j\left(\alpha_{j_k}, x_i\right), g_j \in G,\ j \in \left[1; |G|\right],\ j_k \in \left[1; |A_j|\right] \right\}$.

Мощность данного множества $|G_0| = \sum_{j=1}^{|G|} |A_j|$.

11. Общий алгоритм предобработки данных

Рассмотрим общий алгоритм построения последовательности функций, используемой для функциональной предобработки данных. Он описан для *i*-го определяющего показателя. Для показателя x_i строится последовательность функций L_i, в которую входят непосредственно функции предобработки (описаны ранее) и функции нормировки, причем количество функций предобработки может быть ограничено сверху числом M ($M > 0$). Ограничение количества функций будет рассмотрено далее.

До выполнения непосредственно шагов алгоритма необходимо следующее:

1. Нормировать значения показателей – привести их к отрезку [–1; 1]. Далее подразумевается, что значения всех определяющих и прогнозируемого показателей нормированы, т.е. $y \in [-1;\ 1]$, $x_i \in [-1;1]$, $i \in [1, N]$.

2. Разделить обучающее множество T на два примерно равных по размеру непересекающихся подмножества T_1 и T_2.

3. Принять $t = 1$, где t – номер итерации алгоритма.

Алгоритм:

1. Вычислить коэффициент корреляции r_0 определяющего показателя x_i с прогнозируемым показателем y.

2. Преобразовать значения определяющего показателя x_i каждой из функций, т.е. вычислить $g_j(x_i),\ \forall j \in \left[1; |G_0|\right]$.

3. Для каждого результата преобразования $g_j(x_i)$ вычислить коэффициент корреляции r_j с прогнозируемым показателем y: $r_j = r_{y,g_j(x_i)}$;

4. Выполнить предварительный отбор функций предобработки, т.е. построить множество $G^* = \left\{g(x),\ g \in G_0, I_g = 1\right\}$, где I_g – индикаторная переменная, принимающая значение, равное 1, если заданные условия (будет рассмотрено далее) выполняются для функции g.

5. Проверить, есть ли функции, которые прошли предварительный отбор, т.е. проверить условие $G^* \neq \varnothing$.

а) если таких функций нет $(G^* = \varnothing)$, то построение последовательности для данного показателя можно считать завершенным;

б) если такие функции есть $(G^* \neq \varnothing)$, то выполнить выбор оптимальной функции предобработки. Оптимальной считается функция, для которой на обучающем множестве T выполняется условие: коэффициент корреляции между прогнозируемым и преобразованным с помощью нее определяющим показателями максимальный. Иными словами, найти функцию $g_s, s = \operatorname{argmax} r_j$.

6. Если выполняется условие $r_s > r_0, s = \operatorname{argmax} r_j$, то построение последовательности для данного показателя можно считать завершенным.

7. Применить выбранную функцию к значениям определяющего показателя и нормировать полученные значения (привести их к отрезку [–1; 1]): $x_i = n_t\left(g_s(x_i)\right),\ E(n_t) = [-1;1]$.

8. Добавить выбранные на предыдущих шагах функции к последовательности $L_{i_{2t-1}} = g_s(x), L_{i_{2t}} = n_t$.

9. Принять $t = t + 1$.

10. Если число функций предобработки не превысило ограничение $(t \le M)$, перейти к шагу 1. Иначе построение последовательности для данного показателя можно считать завершенным.

После того как последовательности функций построены, они могут быть применены к значениям показателей из обучающего и тестирующего множеств текущим значениям, подаваемым на вход модели прогнозирования, т.е. показатель x_i до подачи на вход модели преобразуется функцией, которая имеет вид $n_{t_i}\left(g_{t_i}\left(n_{t_i-1}\left(g_{t_i-1}\left(\ldots n_1\left(g_1\left(x_i\right)\right)\ldots\right)\right)\right)\right)$.

12. Обеспечение устойчивости с помощью предварительного отбора функций преобразования

Вариант 1. Выполнить следующие шаги:

1. Вычислить для всех функций разности коэффициентов корреляции на обоих подмножествах обучающего множества.

2. Выбрать заданный процент функций, для которых разность минимальна.

Вариант 2. Предварительный отбор проходят те функции, применение которых к значениям показателей увеличивает коэффициенты корреляции на обоих подмножествах обучающего множества.

Вариант 3. Предварительный отбор функций не выполняется. Считается, что все функции его проходят. Данный вариант описан в исходной версии алгоритма.

13. Ограничение количества функций предобработки

Необходимо рассмотреть следующие варианты длины последовательности функций: количество функций не ограничено, количество функций принимается равным 1, количество функций принимается равным 2, количество функций принимается равным 3.

Ограничение количества функций предобработки в последовательности может иметь смысл в плане уменьшения ошибки на тестирующем множестве, так как подобное ограничение можно рассматривать как простой аналог механизма ранней остановки.

14. Время выполнения

Функциональная предобработка данных может быть выполнена как до отбора определяющих показателей, так и после. Приведем аргументы за и против каждого из вариантов.

Если выполнять предобработку до отбора, то в ходе отбора не будут исключены из рассмотрения те показатели, которые несут полезную информационную нагрузку, но без повышения их коэффициента корреляции с прогнозируемым показателем не учитываются моделью. Если предобработку выполнять после отбора, то возможно исключение из рассмотрения полезных показателей. При этом можно оценить влияние предобработки на ошибку прогнозирования без учета того, как был выполнен отбор. Это возможно благодаря тому, что при каждом эксперименте выполняется предобработка одних и тех же показателей.

15. Эксперименты

В ходе экспериментов использовались оба описанных выше способа предварительного отбора функций, также последовательности функций строились и без предварительного отбора. Кроме того, проводилось сравнение результатов работы алгоритма с ограничением длины последовательности и без него (табл. 1–6). В табл. 1–6 приняты следующие обозначения способа предварительного отбора функций: «–» – его отсутствие, 1 – 1-й способ, 2 – 2-й способ. Во всех таблицах время приведено в минутах.

Эксперименты на стенде (предобработка до отбора)

Всего было сформировано 40 наборов данных.

Таблица 1

Уменьшение ошибки для экспериментов на стенде (предобработка до отбора)

A	*B*	*C*	В процентных пунктах		Относительное	
			среднее, %	максимальное, %	среднее, %	максимальное, %
–	–	12	1,18	1,23	16,87	100,00
–	1	13	1,10	4,14	11,59	48,47
–	3	14	0,79	1,95	8,08	21,77
1	–	12	0,80	1,80	9,04	21,94
1	1	18	1,12	2,79	11,41	27,40
1	3	9	0,90	2,09	9,53	24,40
2	–	16	0,97	2,26	9,88	20,87
2	1	16	0,93	3,23	9,83	37,79
2	2	13	1,19	2,85	11,99	26,76
2	3	17	1,54	4,51	15,69	50,31

Примечание: A – способ предварительного отбора функций; *B* – ограничение количества функций («–» означает отсутствие ограничения); *C* – количество наборов данных, когда ошибка стала меньше.

Таблица 2

Результаты прогнозирования для некоторых наборов данных (эксперименты на стенде, предобработка до отбора)

Длина периода	Период скользящего среднего	Перекрытие	Номер периода	Без предобработки, %	С предобработкой, %
120	30	14	1	6,78	7,24
120	30	14	2	9,10	7,31
120	15	14	2	10,87	7,80
180	30	14	2	10,65	8,17
180	30	14	3	6,79	8,02
240	15	14	3	8,79	6,76
240	30	14	2	10,85	6,34
360	30	29	2	7,04	5,69
360	30	29	4	8,54	4,25
60	15	14	6	9,72	8,80

Применение предобработки данных привело к уменьшению ошибки на тестирующем множестве для 32 наборов данных. Предобработка с применением предварительного отбора функций дает результат лучше, чем без него, для 30 наборов данных. При наличии ограничения длины последовательности функций (от 1 до 3) ошибка на тестирующем множестве меньше для 35 наборов данных.

Эксперименты на агрегате З-205А (предобработка до отбора)

Всего было сформировано 80 наборов данных.

Таблица 3

Уменьшение ошибки для экспериментов на агрегате Р-205А (предобработка до отбора)

A	*B*	*C*	В процентных пунктах		Относительное	
			среднее, %	максимальное, %	среднее, %	максимальное, %
–	–	18	1,32	4,79	9,02	25,28
–	1	36	1,39	3,69	10,10	31,81
–	2	24	1,05	3,55	7,31	22,56
1	–	20	1,10	3,63	8,70	28,25
1	1	32	1,48	3,98	10,29	23,62
1	2	26	0,96	3,21	7,35	18,48
2	–	32	1,28	4,69	8,89	26,89
2	1	42	1,50	4,72	10,85	26,48

Примечание: A – способ предварительного отбора функций; *B* – ограничение количества функций («–» означает отсутствие ограничения); *C* – количество наборов данных, когда ошибка стала меньше.

Таблица 4

Результаты прогнозирования для некоторых наборов данных (эксперименты на агрегате Р-205А, предобработка до отбора)

Длина периода	Период скользящего среднего	Перекрытие	Номер	Без предобработки, %	С предобработкой, %
120	15	14	1	7,03	5,79
120	15	14	2	11,02	8,72
180	15	14	4	16,61	13,91
180	30	29	1	10,96	7,47
180	30	29	5	15,20	13,50
240	15	14	1	12,20	8,52
240	30	29	5	15,20	12,86
360	30	14	3	14,89	11,53
360	30	29	1	12,49	10,36
60	15	14	2	6,72	6,05

Применение предобработки данных привело к уменьшению ошибки на тестирующем множестве для 66 наборов данных. Предобработка с применением предварительного отбора функции дает результат лучше, чем без него, для 56 наборов данных. При наличии ограничения длины последовательности функций (от 1 до 2) ошибка на тестирующем множестве меньше для 59 наборов данных.

Эксперименты на стенде (предобработка после отбора)

Всего было сформировано 40 наборов данных.

Таблица 5

Уменьшение ошибки для экспериментов на стенде (предобработка после отбора)

A	*B*	*C*	В процентных пунктах		Относительное	
			среднее, %	максимальное, %	среднее, %	максимальное, %
1	1	14	0,98	3,64	12,32	61,42
1	–	14	0,79	2,42	8,68	21,90
2	1	22	0,97	2,20	11,31	23,83
2	2	18	1,07	2,86	12,55	39,41
2	–	22	0,85	2,64	10,81	44,62
–	1	18	0,83	2,58	9,35	26,03
–	3	16	1,04	2,59	11,72	26,09
–	–	16	1,08	3,98	15,98	78,87

Примечание: A – способ предварительного отбора функций; *B* – ограничение количества функций («–» означает отсутствие ограничения); *C* – количество наборов данных, когда ошибка стала меньше.

Применение предобработки данных привело к уменьшению ошибки на тестирующем множестве для 36 наборов данных. Предобработка с применением предварительного отбора функции дает результат лучше, чем без него, для 28 наборов данных. При наличии ограничения длины последовательности функций (от 1 до 3) ошибка на тестирующем множестве меньше для 28 наборов данных.

Эксперименты на агрегате Р-205А (предобработка после отбора)

Всего было сформировано 80 наборов данных.

Таблица 6

Уменьшение ошибки для экспериментов на агрегате Р-205А (предобработка после отбора)

A	*B*	*C*	В процентных пунктах		Относительное	
			среднее, %	максимальное, %	среднее, %	максимальное, %
1	1	29	0,98	3,07	6,60	20,49
1	2	28	1,14	4,70	7,29	23,71
1	–	23	1,11	2,83	7,63	16,09
2	2	19	1,27	4,04	8,53	24,41
2	3	26	1,15	5,12	7,50	25,80
2	–	24	1,15	3,65	7,49	20,15
–	2	22	0,93	3,76	6,20	20,14
–	3	21	1,38	6,66	8,70	33,59
–	–	17	1,03	3,79	6,41	19,13

Примечание: A – способ предварительного отбора функций; *B* – ограничение количества функций («–» означает отсутствие ограничения); *C* – количество наборов данных, когда ошибка стала меньше.

Применение предобработки данных привело к уменьшению ошибки на тестирующем множестве для 59 наборов данных, предобработка с применением предварительного отбора функции дает результат лучше, чем без него, для 56 наборов данных. При наличии ограничения длины последовательности функций (от 1 до 3) ошибка на тестирующем множестве меньше для 67 наборов данных.

16. Анализ результатов

Из приведенных результатов можно сделать вывод, что применение функциональной предобработки данных в подавляющем большинстве случаев приводит к повышению качества прогнозирования. Более

чем в половине случаев применение предварительного отбора приводит к уменьшению ошибки на тестирующем множестве. В большинстве случаев ограничение количества функций предобработки в последовательности позволяет получить меньшую ошибку, чем без него.

Заключение

В работе представлены решение задачи долгосрочного прогнозирования и алгоритм функциональной предобработки данных. Приведено сравнение результатов его работы с разными параметрами и разными методами предварительного отбора функций. Продемонстрировано применение алгоритма на практике при работе с реальными данными. Эксперименты показали целесообразность применения функциональной предобработки данных в целом и модификаций, обеспечивающих устойчивость в частности.

Список литературы

1. Lin C.C. Intelligent vibration signal diagnostic system using artificial neural network / ed. by K. Suzuki // Artificial Neural Networks – Industrial and Control Engineering Applications. – London: IntechOpen Limited, 2011. – P. 421–440.

2. Knapp G.M., Javadpour R., HsuPin Wang. An ARTMAP neural network based machine condition monitoring system // Journal of Quality in Maintenance Engineering. – 2000. – Vol. 6, no. 2. – P. 86–105.

3. Khoualdia T., Lakehal A., Chelli Z. Practical investigation on bearing fault diagnosis using massive vibration data and artificial neural network // Big Data and Networks Technologies. BDNT 2019. Lecture Notes in Networks and Systems. – 2019. – Vol 81. – P. 110–116.

4. Гусев А.Л., Еремин И.В., Окунев А.А. Долгосрочное прогнозирование параметров вибрации нефтеперекачивающих агрегатов при помощи нейронных сетей // Нейрокомпьютеры и их применение: материалы XVI Всерос. науч. конф., г. Москва, 13 марта 2018 г. / ФГБОУ ВО МГППУ. – М., 2018. – С. 198–202.

5. Кацив С.Ш. Подход к прогнозированию развития дефектов гидроагрегата на основе использования искусственной нейронной сети // Научные труды Винницкого национального технического университета. – 2012. – № 1. – С. 1–6.

6. Longterm forecasting of solid waste generation by the artificial neural networks / A.M. Abdoli, M.F. Nezhad, R.S. Sede, S. Behboudian // Environmental Progress & Sustainable Energy. – 2012. – Vol. 31, no. 4. – P. 68–636.

7. Peter Zhang G., Qi Min. Neural network forecasting for seasonal and trend time series // European Journal of Operational Research. – 2005. – No. 160. – P. 501–514.

8. Solar radiation forecasting using ad-hoc time series preprocessing and neural networks / C. Paoli, C. Voyant, M. Muselli, M.L. Nivet // Emerging Intelligent Computing Technology and Applications. ICIC 2009. Lecture Notes in Computer Science. – 2009. – Vol. 5754. – P. 898–907.

9. The effects of pre-processing methods on forecasting improvement of artificial neural networks / A. Azadeh, M. Sheikhalishahi, M. Tabesh, A. Negahban // Australian Journal of Basic and Applied Sciences. – 2011. – Vol. 5, no. 6. – P. 570–580.

10. Time-series extreme event forecasting with neural networks at uber / N. Laptev, J. Yosinski, S. Smyl, Li. Li Erran // International Conference on Machine Learning, Sydney, Australia, 6–11 August 2017. – Sydney, Australia, 2017. – P. 1–5.

11. Nguyen H.H., Chan C.W. Multiple neural networks for a long term time series forecast // Neural Computing & Applications. – 2004. – Vol. 13, no. 1. – P. 90–98.

12. Ruta D., Gabrys B. Neural network ensembles for time series prediction // International Joint Conference on Neural Networks, Orlando, FL, USA, 12–17 August 2007. – Orlando, FL, USA, 2007. – P. 1204–1209.

13. Multi-stage algorithm based on neural network committee for prediction and search for precursors in multi-dimensional time series / S. Dolenko, A. Guzhva, I. Persiantsev, J. Shugai // Artificial Neural Networks – ICANN 2009. Lecture Notes in Computer Science. – 2009. – Vol. 5769. – P. 295–304.

14. Гусев А.Л., Черепанов Ф.М., Ясницкий Л.Н. Функциональная предобработка входных сигналов нейронной сети // Нейрокомпьютеры. Разработка и применение, 2013. – Т. 5. – С. 19–21.

15. Гусев А.Л., Гильманов А.Р., Окунев А.А. Предобработка статистических данных для повышения качества прогноза нейронной сети // Современная наука: актуальные проблемы теории и практики. Серия Естественные и технические науки. – 2018. – № 03. – С. 49–51.

References

1. Lin C.C. Intelligent Vibration Signal Diagnostic System Using Artificial Neural Network. Ed. by Suzuki K. *Artificial Neural Networks - Industrial and Control Engineering Applications*. London: IntechOpen Limited, 2011, pp. 421-440.

2. Knapp G.M., Javadpour R., Hsu Pin Wang. An ARTMAP neural network based machine condition monitoring system. *Journal of Quality in Maintenance Engineering*, vol. 6, no. 2, 2000. pp. 86-105.

3. Khoualdia T., Lakehal A., Chelli Z. Practical Investigation on Bearing Fault Diagnosis Using Massive Vibration Data and Artificial Neural Network. *Big Data and Networks Technologies. BDNT 2019. Lecture Notes in Networks and Systems*, vol. 81, 2019, pp. 110-116.

4. Gusev A.L., Eremin I.V., Okunev A.A. Dolgosrochnoe prognozirovanie parametrov vibratsii nefteperekachivaiushchikh agregatov pri pomoshchi

neironnykh setei [Long-term forecasting of vibration parameters of oil pumping units using neural networks]. Materialy XVI Vserossiiskoi nauchnoi konferentsii «Neirokomp'iutery i ikh primenenie. 13 March 2018. Moscow. Moskovskii gosudarstvennyi psikhologo-pedagogicheskii universitet, 2018, pp. 198-202.

5. Katsiv S.Sh. Podkhod k prognozirovaniiu razvitiia defektov gidroagregata na osnove ispol'zovaniia iskusstvennoi neironnoi seti. *Nauchnye trudy Vinnitskogo Natsional'nogo Tekhnicheskogo Universiteta*, 2012, no. 1, pp. 1-6.

6. 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*, 2012, vol. 31, no. 4, pp. 68-636.

7. G. Peter Zhang, Min Qi. Neural network forecasting for seasonal and trend time series. *European Journal of Operational Research*, 2005, no. 160, pp. 501–514.

8. Paoli C., Voyant C., Muselli M., Nivet M.L. Solar Radiation Forecasting Using Ad-Hoc Time Series Preprocessing and Neural Networks. *Emerging Intelligent Computing Technology and Applications. ICIC 2009. Lecture Notes in Computer Science*, 2009, vol. 5754, pp. 898-907.

9. Azadeh A., Sheikhalishahi M., Tabesh M., Negahban A. The Effects of Pre-Processing Methods on Forecasting Improvement of Artificial Neural Networks. *Australian Journal of Basic and Applied Sciences*, vol. 5, no. 6, 2011, pp. 570-580.

10. Laptev N., Yosinski J., Smyl S., Li Erran Li. Time-series Extreme Event Forecasting with Neural Networks at Uber. *International Conference on Machine Learning, 6–11 August 2017*. Sydney, Australia, 2017, pp. 1-5.

11. 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.

12. Ruta, D., Gabrys, B. Neural Network Ensembles for Time Series Prediction. *2007 International Joint Conference on Neural Networks. 12–17 August 2007, Orlando, FL, USA*, 2007, pp. 1204-1209.

13. Dolenko, S., Guzhva, A., Persiantsev, I., Shugai, J. Multi-stage Algorithm Based on Neural Network Committee for Prediction and Search for Precursors in Multi-dimensional Time Series. *Artificial Neural Networks – ICANN 2009. ICANN 2009. Lecture Notes in Computer Science*, 2009, vol. 5769, pp. 295-304.

14. Gusev A.L., Cherepanov F.M., Iasnitskii L.N. Funktsional'naia predobrabotka vkhodnykh signalov neironnoi seti [Functional preprocessing of neural network input signals]. *Neirokomp'iutery. Razrabotka i primenenie*, 2013, vol. 5, pp. 19-21.

15. Gusev A.L., Gil'manov A.R., Okunev A.A. Predobrabotka statisticheskikh dannykh dlia povysheniia kachestva prognoza neironnoi seti [Preprocessing of statistical data to improve the quality of the forecast by a neural network]. *Sovremennaia nauka: aktual'nye problemy teorii i praktiki. Seriia "Estestvennye i tekhnicheskie nauki"*, no. 3, 2018, pp. 49-51.

Статья получена: 13.07.2020

Статья принята: 01.09.2020

Сведения об авторе

Окунев Александр Анатольевич (Пермь, Россия) – аспирант, кафедры «Прикладная математика и информатика», Пермский государственный национальный исследовательский университет (614990, Пермь, ул. Букирева, 15, e-mail: info@psu.ru).

About the author

Alexander A. Okunev (Perm, Russian Federation) – Postgraduate Student, Department of Applied Mathematics and Computer Science, Perm State University (614990, Perm, Bukireva st., 15, e-mail: info@psu.ru).

Библиографическое описание статьи согласно ГОСТ Р 7.0.100–2018:

Окунев, А.А. Использование функциональной предобработки данных при прогнозировании параметров вибрации нефтеперекачивающих агрегатов / А. А. Окунев. – DOI 10.15593/2499-9873/2020.3.03. – Текст : непосредственный // Прикладная математика и вопросы управления = Applied Mathematics and Control Sciences. – 2020. – № 3. – С. 51–72.

Цитирование статьи в изданиях РИНЦ:

Окунев А.А. Использование функциональной предобработки данных при прогнозировании параметров вибрации нефтеперекачивающих агрегатов // Прикладная математика и вопросы управления. – 2020. – № 3. – С. 51–72. DOI: 10.15593/2499-9873/2020.3.03

Цитирование статьи в references и международных изданиях:

Cite this article as:

Okunev A.A. Functional data preprocessing application to oil-transfer pumps vibration parameters forecasting. *Applied Mathematics and Control Sciences*, 2020, no. 3, pp. 51–72. DOI: 10.15593/2499-9873/2020.3.03 (*in Russian*)