Лазебная Е.А., доц.
Белгородский государственный технологический университет им. В.Г. Шухова

ПОРЯДОК ПРОВЕДЕНИЯ ПРЕДВАРИТЕЛЬНОЙ ОБРАБОТКИ ДАННЫХ, СОСТАВЛЯЮЩИХ ПРОГНОЗНЫЙ ФОН ПРИ ПРОГНОЗИРОВАНИИ С ПОМОЩЬЮ ВРЕМЕННЫХ РЯДОВ

l_el_a@mail.ru

Задача формирования эффективной территориально-отраслевой системы подготовки специалистов, востребованных существующим на рынке труда спросом, требует создания качественного информационного обеспечения в виде как краткосрочного, так и долгосрочного, постоянно уточняемого прогноза. Осуществить эффективные прогнозные оценки невозможно без использования адекватных изучаемым процессам математических моделей, опирающихся на ретроспективные данные и количественно оценивающих существующий спрос на специалистов. Важным этапом в построении математической модели прогнозирования востребованности специалистов на региональном рынке труда является предварительная обработка данных, составляющих прогнозный фон, которая выполняется с целью повышения качества временного ряда, что в конечном итоге повлечет за собой повышение точности результата прогноза, представляющего собой оценку будущей востребованности специалистов. При этом предлагаемые подходы должны учитывать необходимость проведения адаптации прогнозной модели к условиям неполных и нечетких данных в виду особенностей функционирования элементов системы – рынка труда и рынка образовательных услуг в нечетких условиях.

Ключевые слова. *Предварительная обработка данных, временные ряды, прогнозирование востребованности специалистов, снижение противоречивости наборов временного ряда.*

Введение. Моделирование временного ряда представляет собой формализованную процедуру, позволяющую по прошлым значениям ряда вычислять будущие значения прогнозируемого показателя на заданный период упреждения. Конечной целью формирования временного ряда является его подготовка к использованию для построения математической модели прогнозируемого процесса, которая и будет описывать распознаваемую ситуацию с заданным условием оптимизации [1].

Формализованное представление математической модели, описывающей востребованность специалистов на региональном рынке труда, этапы адаптивного построения математической модели прогнозирования и их особенности подробно рассмотрены в [2–3]. Возможность моделирования зависимости величины сегмента рынка труда для отдельной профессиональной группы от экономических показателей обеспечивается посредством выявления и анализа основных групп факторов, описывающих, существующий на региональном рынка труда совокупный спрос на специалистов. На основе анализа, проведенного с помощью когнитивной модели, в качестве основных факторов для прогнозирования востребованности специалистов определены следующие: показатель потенциального спроса, показатель реализованного спроса, показатель резервного спроса и показатель степени трудоустройства выпускников [4].

Методология. Прогнозирование востребованности специалистов на региональном рынке труда на основе временных рядов основывается на многофакторной регрессионной модели [5]. Предварительная обработка данных, составляющих прогнозный фон, следует общей концепции проведения исследований с помощью методов интеллектуального анализа данных Data Mining, включая методы предварительной обработки данных, классификации и регрессии [6–7]. Аппарат нечетких множеств и нечеткой логики используется для учета специфики функционирования регионального рынка труда относительно сложившейся на нем профессионально-квалификационной структуры в нечетких условиях и при неполных данных [8–9]. Оценка эффективности проведенной предобработки данных временного ряда проводится с помощью соотношения, в теории непрерывных функций называемого константой Липшица.

Основная часть. Предварительная обработка данных включает в себя несколько шагов.

Шаг 1. Из всей совокупности показателей регионального рынка труда, описывающих совокупный спрос на специалистов, определяется набор факторов, значения которых могут быть вычислены на основании собранных сведений. Эти значения представляют собой прогнозный фон, на основе которого будет строиться временной ряд.

В результате получена группа факторов, значения которых определены в интервале вре-

мени T с периодичностью один год: $\mathrm{G}' = \left\{\mathrm{G}_1', \mathrm{G}_2', \dots, \mathrm{G}_{\mathrm{k}'}'\right\}^{\mathrm{T}}$, где k'– определяет количество полученных факторов. Поскольку объект исследования – региональный рынок труда функционирует в нечетких условиях и при неполных данных, то для каждого из факторов определена своя глубина погружения в историю (т.е. временной период, в течение которого определены его значения): $R' = \left\{R_1', R_2', \dots, R_{k'}'\right\}^{\mathrm{T}}$ (табл. 1).

Таблица 1

Сведения, полученные на шаге 1 предварительной обработки данных

Фактор	$(G_1')^{\mathrm{T}}$	$(G_2')^{\mathrm{T}}$	…	$\left(\mathrm{G}_{\mathrm{k}'}'\right)^{\mathrm{T}}$
Глубина погружения	$(R_1')^{\mathrm{T}}$	$(R_2')^{\mathrm{T}}$	…	$\left(R_{k'}'\right)^{\mathrm{T}}$

Шаг 2. Из полученного множества факторов G' для отбора наиболее значащих факторов, участвующих в построении модели прогнозирования, на данном шаге используется набор правил вывода. При этом необходимо оптимальным образом учитывать мнение эксперта относительно степени влияния на результат прогноза факторов из множества G', глубину погружения каждого фактора в историю и значения коэффициентов парной корреляции факторов.

С одной стороны набор правил рассматривает возможность исключения некоторых малозначащих факторов для сохранения бо́льшего размера глубины погружения всего временного ряда. С другой стороны, набор правил рассматривает возможность сокращения размера глубины погружения всего временного ряда для учета в модели наиболее важных факторов из множества G', для которых R_l' в интервал времени T не является максимальной, где $R_l' \in R', l = 1, \dots, k'$. В результате выполнения этого шага из общей совокупности факторов G' получен поднабор факторов $\mathrm{G} = \{\mathrm{G}_1, \mathrm{G}_2, \dots, \mathrm{G}_{\mathrm{k}}\}^{\mathrm{T}}$ и оптимальная глубина погружения $R_{\text{оптим}}$. При этом k определяет количество полученных факторов в множестве G, которые будут учтены в модели прогнозирования, величина $R_{\text{оптим}}$ будет одинаковой для всех факторов множества G (табл. 2).

Таблица 2

Сведения, полученные на шаге 2 предварительной обработки данных

Фактор	$(G_1)^{\mathrm{T}}$	$(G_2)^{\mathrm{T}}$	…	$(G_k)^{\mathrm{T}}$
Глубина погружения	$R_{\text{оптим}}$			

Шаг 3. Выполняется обработка данных в зависимости от требований к форме получения результата прогноза:

а) для получения на выходе прогнозной модели в качестве результата информации о характере динамики изменения исследуемого процесса трудоустройства на период упреждения (т.е. оказывается достаточным прогнозировать только знак приращения), дальнейшую предобработку данных временного ряда необходимо провести по правилу:

$$GK_{j\,t_i} = \begin{cases} 1, \Delta G_{j\,t_i} > 0 \\ 0, \Delta G_{j\,t_i} = 0 \\ -1, \Delta G_{j\,t_i} < 0 \end{cases}, \qquad (1)$$

где $\Delta G_{j\,t_i} = G_{j\,t_{i+1}} - G_{j\,t_i}$, $t_i \in T$, $i = 1, \dots, R_{\text{оптим}} - 1$, $j = 1, \dots, k$, k – количество факторов в множестве G. Полученный в результате предобработки ряд $\left\{GK_{t_1}, GK_{t_2}, \dots, GK_{t_{R_{\text{оптим}}-1}}\right\}^{\mathrm{T}}$ будет сохранять основную информацию о характере и последовательности изменений процессов трудоустройства, но такой переход сопряжен и с потерей части информации. Поэтому использовать его можно только при соответствующих требованиях к виду получаемого результата.

б) для получения на выходе прогнозной модели значения, определяющего величину динамики изменения исследуемого процесса востребованности специалистов на период упреждения, необходимо в качестве значений факторов модели прогнозирования использовать не конкретные значения соответствующих им показателей в каждый отдельный временной период, а их приращения за последовательные временные периоды. В связи с этим предобработка исходных данных будет заключаться в следующем преобразовании: от данных $G = \{G_1, G_2, \dots, G_k\}^{\mathrm{T}}$, где k – количество факторов в множестве G перейдем к $R_{\text{оптим}} - 1$ разностям этого ряда: $\Delta G_{j\,t_1}$, $\Delta G_{j\,t_2}, \dots, \Delta G_{j\,t_{R_{\text{оптим}}-1}}$, где $\Delta G_{j\,t_i} = G_{j\,t_{i+1}} - G_{j\,t_i}$, $t_i \in \mathrm{T}, i = 1, \dots, R_{\text{оптим}} - 1$, $j = 1, \dots, k$, k – количество факторов в множестве G.

Шаг 4. Для увеличения размера временного ряда необходимо брать приращения не только за последовательные, а за все возможные комбинации периодов $t_i \in T$, что позволит получить временной ряд, размер которого вычисляется по формуле:

$$R_{\text{макс}} = \frac{R_{\text{оптим}}}{2} \cdot \left(R_{\text{оптим}} - 1\right) \qquad (2)$$

В связи с этим предобработка исходных данных будет заключаться в следующем преобразовании: от данных $G = \{G_1, G_2, \dots, G_k\}^{\mathrm{T}}$, где k – количество факторов в множестве G перейдем к $R_{\text{макс}} - 1$ разностям этого ряда: $\Delta G_{j\,t_1}$,

$\Delta G_{j\,t_2}, \dots, \Delta G_{j\,t_{R_{макс}-1}}$, где $\Delta G_{j\,m} = G_{j\,t_i} - G_{j\,t_r}$, для всех $i < r$, где $i = 2, \dots, R_{оптим}$, $r = 1, \dots, R_{оптим} - 1, m = 1, \dots, R_{макс} - 1$.

При проведении такой обработки данных сведения о динамике не теряются, однако при этом необходимо учитывать в модели уменьшающуюся степень достоверности таких рядов данных, а также степень устаревания данных. Для чего в модель введены коэффициенты достоверности и устаревания: $K_{устар} = \frac{1}{i-1}$, и $K_{дост} = \frac{1}{i-r}$ для всех $i < r$, где $i = 2, \dots, R_{оптим}$, $r = 1, \dots, R_{оптим} - 1$. Пример вычисленных значений коэффициентов достоверности и устаревания при $R_{оптим}$=10 приводится в табл. 3. Произведение коэффициентов достоверности и устаревания показывает, что их использование позволит учесть в модели неравнозначность тех рядов данных, которые получены после проведения предобработки исходных данных.

Таблица 3

Сведения, полученные на шаге 4 предварительной обработки

№ набора	*i*=2..10	*r*=1..9	$i - r$	$K_{дост}$	$K_{устар}$	$K_{дост} \cdot K_{устар}$
1	2	1	1	1	1	1
2	3	1	2	0,5	0,5	0,25
...	…	…	…	…	…	
	9	8	1	1	0,125	0,125
	10	1	9			
...	…	…	…	…	…	
	10	8	2	0,5	0,125	0,0625
$R_{макс}$=45	$R_{оптим}$=10	9	1	1	0,111	0,111

Шаг 5. При решении задачи нахождения аппроксимируемой функции, описывающей распознаваемую ситуацию с заданным условием оптимизации, должны учитываться такие характеристики временного ряда, как полнота, равномерность, противоречивость и повторяемость [10]. Для получения возможности анализа этих характеристик предлагается проведение кластеризации по значениям независимых переменных, что позволит создать определенные правила, с помощью которых в дальнейшем можно относить объекты к различным классам или к одному классу. При этом объекты группируются, исходя из их сходства, или близости [5]. Полнота выборки, представляющей собой наборы временного ряда, определяется обеспеченностью классов обучающими наборами. Равномерность выборки показывает, насколько равномерно распределены наборы по классам, а повторяемость - показатель, характеризующий количество одинаковых наборов в рамках одного класса. Противоречивыми считаются наборы временного ряда, описывающие одинаковые ситуации (значения независимых переменных которых имеют сходство), но зависимая переменная имеет разные значения [10].

Естественно, что чем больше в обучающей выборке присутствует наборов, для которых входные векторы близки друг к другу, а выходные далеки (противоречивость) и чем ниже полнота задания временного ряда, тем труднее провести процесс построения математической модели прогнозирования. Поэтому основными требованиями к временному ряду являются характеристики непротиворечивости и полноты задания его значений. Решение задачи снижения противоречивости наборов временного ряда для исключения из неё противоречивых и резко выделяющихся из всех остальных данных на данном шаге проводится с помощью кластерного анализа, который проводится в 2 этапа: разделение наборов временного ряда на классы и устранение противоречивости данных.

Для разделения наборов временного ряда на классы, количество которых заранее известно использован метод *K*-средних [5], в основе которого использован алгоритм, представляющий собой итерационную процедуру. На каждой итерации происходит изменение границ классов и смещение их центров. В результате минимизируется расстояние между элементами внутри классов. Остановка алгоритма производится тогда, когда границы классов и расположения центроидов не перестанут изменяться от итерации к итерации.

В результате выполнения этого этапа получено распределение наборов временного ряда по классам на основе значений независимых переменных.

Для устранения противоречивости применяется искусственное сближение выходных значений зависимых переменных для наборов временного ряда, размещенных в одном классе, значения независимых переменных которых имеют сходство. Рассмотрим 2 способа решения задачи устранения противоречивости данных временного ряда для отдельно взятого класса, в котором определено *m* наборов данных.

1 способ Выходное значение зависимой переменной c_v' v-го набора отдельного класса (где v=1..m, m – количество наборов класса) будет рассчитываться как среднее выходных значений всех m наборов, размещенных в этом классе, взвешенное по функции от расстояния до входного v-го набора значений класса:

$$c_v' = \frac{\sum_{r=1}^{m} c_v \cdot \lambda_{v\,r}}{\sum_{r=1}^{m} \lambda_{v\,r}} \quad (3)$$

Здесь $\lambda_{v\,r}, (\,0 \le \lambda_{v\,r} \le 1)$ – весовые коэффициенты, вычисленные с помощью специальной взвешивающей функции. Роль взвешивающей функции может выполнять функция от расстояния между входными векторами, удовлетворяющая следующим условиям:

- существовать и быть неотрицательной на всем множестве возможных значений расстояния;
- убывать с увеличением расстояния;
- в зависимости от некоторого параметра α изменять скорость убывания. Параметр α задает степень упрощения исходной выборки.

Одной из наиболее известных и широко применяемых функций, удовлетворяющих перечисленным условиям, является функция Гаусса [10], которую и предлагается использовать в качестве взвешивающей. Таким образом, весовые коэффициенты в формуле (3) будут вычисляться следующим образом:

$$\lambda_{v\,r} = e^{-\left(\frac{\|A_v - A_r\|}{\alpha}\right)^2} \quad (4)$$

где r, v – номера наборов отдельного класса; r, v =1..m, m – количество наборов класса; A_v, A_r – сами наборы (включая значения только независимых переменных); $\|A_v - A_r\|$ – мера расстояния в многомерном пространстве (Евклидово расстояние);α > 0 − параметр, задающий ширину (отклонение) функции и определяющий ее влияние.

Функция Гаусса принимает свое максимальное значение, равное единице, при $A_v = A_r$ и убывает при удалении A_v от A_r. Таким образом, в формуле (4) коэффициент $\lambda_{v\,v} = 1$ (это максимальный коэффициент), $\lambda_{v\,r} \approx 0$, если $\|A_v - A_r\| > \alpha$.

В результате будет получено искусственное сближение выходных значений наборов, входные значения которых близки между собой. При использовании такого подхода обработки временного ряда количество наборов остается прежним, но противоречивость при этом несколько устранена.

2 способ Можно провести усреднение выходных значений наборов внутри каждого класса с учетом коэффициентов устаревания данных и достоверности данных, введенных в рассмотрение на шаге 4, используя методы усреднения, учитывающие частоту, например как средняя арифметическая взвешенная. При этом количество наборов временного ряда сократится до количества выделенных классов. Это может быть неплохо только в том случае, если в результате данные временного ряда останутся достаточно полными, т.е. для каждого класса есть выходное значение. В этом случае, задача определения неизвестных значений параметров отпадает – при получении нового набора решается задача классификации, определяющая к какому классу из существующих он больше всего подходит.

Шаг 6. Оценка эффективности проведенной предобработки данных временного ряда проводится с помощью соотношения (5), в теории непрерывных функций называемого константой Липшица [11], которая для пары наборов значений двух независимых факторов одного класса A_v, A_r и значений зависимых факторов C_v, C_r, характеризует сложность наборов следующим образом:

$$L_{v\,r} = \frac{\|C_v - C_r\|}{\|A_v - A_r\|} \quad (5)$$

Сложность воспроизведения всего временного ряда может быть получена расчетом среднего или максимального и минимального значений $L_{v\,r}$ для всех пар наборов. Применение соотношения (5) с целью оценки обучающей возможности временного ряда обсуждалось в литературе и показало свою практическую значимость [11].

Выводы. При моделировании временного ряда, используемого для оценки востребованности специалистов, в работе была учтена возможность возникновения ряда характерных трудностей, затрудняющих моделирование, а также приведены подходы к их устранению, а именно:

– развитие экономических процессов и явлений происходит непрерывно, но реально исследовать можно лишь дискретные по времени значения показателей рынка труда. Так как в исследованиях в качестве временного интервала выбран один год, то выборка содержит сравнительно немного элементов (небольшую глубину погружения в историю). Предложен подход увеличения размера существующей выборки на исследуемом интервале Т, на основе которой выполняется моделирование;

– поскольку характерной чертой временного ряда является существенность порядка наблюдения, то в модель введены коэффициенты устаревания и достоверности;

– экономические ряды динамики часто являются сильно автокоррелированными. Это учитывается при формировании группы наиболее значимых факторов в наборе правил вывода.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. Афанасьев В.Н., Юзбашев М.М. Анализ временных рядов и прогнозирование: Учебник. М.: Финансы и статистика, 2001. 228 с.

2. Лазебная Е.А. Методы и алгоритмы решения задачи прогнозирования в системе управления планированием подготовки специалистов // Приборы и системы. Управление, контроль, диагностика. Ежемесячный научно-технический журнал. 2014. № 11. С. 65–71.

3. Лукашин Ю.П. Адаптивные методы краткосрочного прогнозирования временных рядов. М.: Финансы и статистика, 2003. 415 с.

4. Лазебная Е.А., Лазебная И.А. Задачи и информационное наполнение системы прогнозирования потребности в трудовых ресурсах // Содействие профессиональному становлению личности и трудоустройству молодых специалистов в современных условиях: сб. материалов V Междунар. заочная науч.-практ. конф., Белгород : Изд-во БГТУ, 2013. С. 22–28.

5. Большаков А.А., Каримов Р.Н. Методы обработки многомерных данных и временных рядов. М.: Горячая линия-Телеком, 2007. 522 с.

6. Барсегян А.А., Куприянов М.С., Степаненко В.В. Холод И.И. Методы и модели анализа данных: OLAP и Data Mining: учеб. пособие. СПб.: БХВ-Петербург, 2004. 331 с.

7. Чубукова И.А. Data Mining: учеб. пособие. М.: БИНОМ. Лаборатория знаний, 2006. 324 с.

8. Гаврилова Т.А., Хорошевский В.Ф. Базы знаний интеллектуальных систем: учеб. пособие для вузов. СПб.: Питер, 2001. 384 с.

9. Баллод Б.А., Елизарова Н.Н. Методы и алгоритмы принятия решений в экономике. СПб.: Финансы и статистика, 2009. 224 с.

10.Тарасенко Р.А., Крисилов В.А. Предварительная оценка качества обучающей выборки для нейронных сетей в задачах прогнозирования временных рядов // Труды Одесского политехнического университета. 2001. Вып.1. С. 90–93.

11.Царегородцев В.Г. Предобработка обучающей выборки, выборочная константа Липшица и свойства обученных нейронных сетей / Нейроинформатика и ее приложения: сб. материалов X Всеросс. семинара // Красноярск, 2002. С.146–150.

Lazebnaya E.A.

THE ORDER OF A PRE-PROCESSING THE DATA THAT MAKE UP THE HISTORICAL DATA IN FORECASTING TIME SERIES

The task of creating an effective territorial and sectoral system of training of specialists requires the creation of high-quality information support in the form of both short and long term, continually refines the forecast. Implement effective forward-looking assessment is not possible without adequate study the process of mathematical models based on historical data and measure the existing demand for specialists. An important step in building a mathematical model of forecasting the demand for professionals in the regional labor market is a pre-processing the data, which is performed to improve the quality of the time series that eventually will lead to improve the accuracy of the forecast is an estimate of future demand for specialists. Proposed approach must take into account the need for adaptation of a predictive model to the conditions of incomplete and unclear data referring to elements of the functioning of the system - the labor market and the education market in fuzzy conditions.

Key words: *Pre-processing of data, time series, forecasting demand for professionals, reduced of the contradictory sets of time series.*

Лазебная Елена Александровна, доцент кафедры информационных технологий.
Белгородский государственный технологический университет им. В.Г. Шухова.
Адрес: Россия, 308012, Белгород, ул. Костюкова, д. 46.
E-mail: l_el_a@mail.ru