УДК 004.04

МЕТОДЫ И АЛГОРИТМЫ КОМПЬЮТЕРИЗИРОВАННОЙ СИСТЕМЫ ПРОГНОЗИРОВАНИЯ ПОКАЗАТЕЛЕЙ НАРОДОНАСЕЛЕНИЯ В УСЛОВИЯХ ДОНЕЦКОЙ ОБЛАСТИ

Привалов М.В., Стихарь А.Г.
Донецкий национальный технический университет

Abstract.
M. V. Privalov, A. G. Stihar *Methods and algorithms of computer system of the population prediction in Donetsk area*. Article describes experience of application of neural networks and decision trees for task of the population prediction using dynamic sequences achieved from Central administrative board of statistics in Donetsk area. Prediction of the population is shown that lowest error rate could

Введение

В последнее время происходит регионализация всех областей общественной жизни. Управление развитием региона, в частности планирование его бюджета, требует знания перспективной численности и особенностей возрастной структуры населения. Специфические черты общественного развития, разная степень проявления социально-экономических проблем, в свою очередь, создают как прямое, так и опосредованное влияние на формирование рождаемости, смертности, миграционных процессов, требуют дифференцированного подхода к обоснованию направлений улучшения демографической ситуации в стране.

Демографические прогнозы, как было сказано выше, чаще всего используются в качестве основы для планирования. Например, при оценке потребностей страны или региона в новых рабочих местах, учителях, школах, врачах, медицинских сестрах, городском жилье или продуктах питания, необходимо иметь сведения о численности населения, которому будут нужны услуги. Таким образом, демографические прогнозы служат отправной точкой для большинства прогнозов о будущих потребностях.

Постановка задачи

Входная информация, используемая для прогнозирования, имеет вид временного ряда, т.к. представляет собой упорядоченную по времени (по годам) последовательность значений некоторых переменных величин. Каждое отдельное значение переменной называется отсчётом временного ряда. Тем самым, временной ряд существенным образом отличается от простой выборки данных.

Прогнозирование временных рядов заключается в построении модели для предсказания будущих событий основываясь на известных событиях прошлого (ретроспекция), предсказания будущих данных до того как они будут измерены [1].

Все вышесказанное иллюстрирует рис. 1:

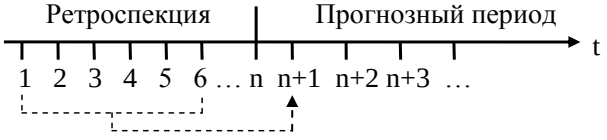

Рисунок 1 – Иллюстрация процесса прогнозирования

Пусть заданы n дискретных отсчетов $\{y(t_1), y(t_2),..., y(t_n)\}$ в последовательные моменты времени $t_1, t_2,..., t_n$. Тогда задача прогнозирования состоит в предсказании значения $y(t_{n+1})$ в некоторый будущий момент времени t_{n+1}:

$$y_t \to (y_1, y_2,..., y_n) \xrightarrow{F} y_{n+1},$$

где F – функциональный преобразователь, который, в нашем исследовании, представляет собой аппарат нейронных сетей и деревьев решений

В работе, для решения задачи прогнозирования с помощью нейронной сети, была выбрана радиально-базисная сеть (RBF) [2], на вход которой подавался многомерный временной ряд, а результатом прогнозирования являлось значение члена временного ряда в требуемый момент времени.

Для повышения качества прогноза производилась предварительная (препроцессорная) обработка информации, т.к. обычно нейронные сети плохо работают с величинами из широкого диапазона значений, встречающихся во входных данных. Для исключения этого нежелательного явления данные необходимо отмасштабировать в диапазоне [0..1]. Масштабирование входных данных производилось следующим образом:

$$X_S = Sc \cdot X_U + Of, \quad (1)$$

$$Sc = \frac{T_{\max} - T_{\min}}{R_{\max} - R_{\min}}, \quad (2)$$

$$Of = T_{\min} - Sc \cdot R_{\min}, \quad (3)$$

где X_S, X_U – соответственно, отмасштабированные и исходные входные данные;

$T_{\min} = 0, T_{\max} = 1$ – максимум и минимум целевой функции;

$R_{\max}, R_{\min}$ – максимум и минимум входных данных.

Использованная в работе радиально-базисная нейронная сеть (рис. 2) представляла собой сеть с одним скрытым слоем:

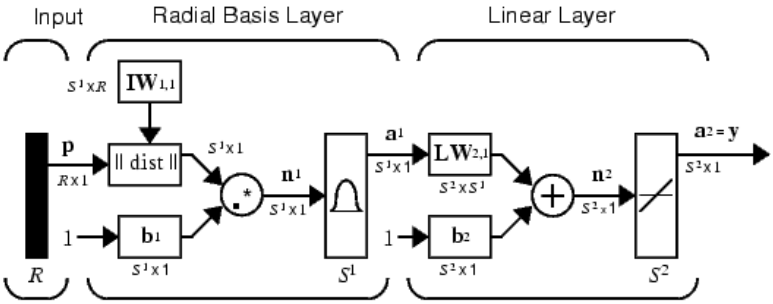

Рисунок 2 – Радиально-базисная нейронная сеть

где, R – число элементов входного вектора, S^1 – число нейронов в скрытом слое, S^2 – число нейронов в выходном слое.

Скрытый слой осуществлял преобразование входного вектора X с использованием радиально-базисных функций (RBF). Практически используются различные радиально-базисные функции, в данной же работе была применена наиболее часто употребляемая функция – Гауссиан (рис. 3).

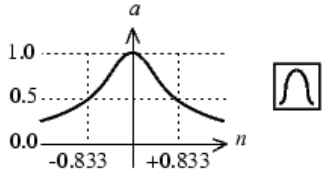

Рисунок 3 – Передаточная функция нейронов скрытого слоя (Гауссиан)

Эта функция имеет максимум, равный 1, при n = 0 и плавно убывает при увеличении n, достигая значения 0.5 при n = ±0.833.

Радиально-базисный нейрон имел следующий вид, представленный на рис. 4:

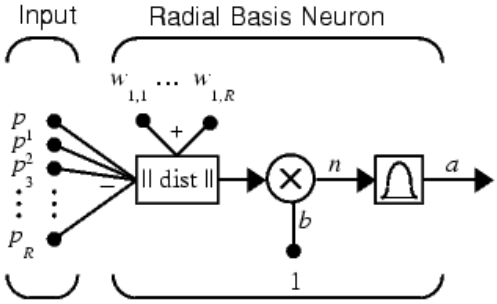

Рисунок 4 – Вид радиально-базисного нейрона

В первом каскаде нейрона вычислялось Евклидово расстояние между входным вектором $\overline{P}$ и $\overline{W}$. Далее полученное расстояние умножалось на константу b:

$$n = \|w - p\| \cdot b \quad (4)$$

Второй каскад реализовывал активационную функцию:

$$a = \exp(-n^2), \quad (5)$$

Выходной слой сети представлял линейный сумматор, а выход сети описывался выражением:

$$u = \sum_{k=1}^{N} w_k \varphi_k(X) \quad (6)$$

где w_k – вес, связывающий выходной нейрон с k-ым нейроном скрытого слоя.

Модель, представленная в виде дерева решений, является интуитивно понятной и упрощает понимание решаемой задачи. В работе использован масштабируемый алгоритм деревьев решений – SLIQ[3]. Выбор был обоснован тем, что SLIQ относится к классу регрессионных решающих деревьев (т.е. целевая переменная имеет непрерывные значения) и позволяет в процессе построения дерева определить значимые переменные, а также имеет возможность специальной обработки пропущенных значений.

С каждой вершиной, которая не является терминальным узлом (листом), связано некоторое значение, а каждому ребру, выходящему из узла, также соответствуют некоторое значение, которое является результатом вычисления выражения. Вычисления проводятся, начиная с корня и двигаясь к потомкам, до листа. Каждый лист имеет значение целевой переменной.

Построение дерева производилось следующим образом. Пусть имеется таблица данных X, в которой n атрибутов (к каждому столбцу атрибута X^i, прикреплен столбец с индексами ind_X^i), и Y – целевая переменная. На первом шаге необходимо выполнить сортировку каждого непрерывного атрибута, причем независимо друг от друга, поэтому существует необходимость хранить индексы для каждого непрерывного предиктора.

X							У
ind_X^1	X^1	ind_X^2	X^2	...	ind_X^n	X^n	

В результате получили $X^1 = \{x_1, x_2, ..., x_n\}$ – отсортированные значения числового атрибута X^1. Так как, любое значение между x_i и x_{i+1} разделит множество на те же самые два подмножества, необходимо исследовать только $n-1$ возможное разбиение. Середина каждого интервала x_i и x_{i+1} считается возможной точкой разбиения (т.е. возможным узлом).

В алгоритме SLIQ каждый узел дерева решений имеет ровно двух потомков. На каждом шаге построения дерева правило вида $x_i \leq c$, формируемое в узле, делит заданное обучающую выборку на две части – часть, в которой выполняется правило (потомок – right) и часть, в которой правило не выполняется (потомок – left), все вышеописанное иллюстрирует рис. 5.

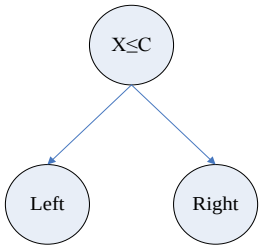

Рисунок 5 – Пример решающего дерева

Для выбора оптимального правила была использована суммарно-квадратичная функция оценки качества разбиения, которая определяется как минимизация SSE (7):

$$\min : SSE = SSE_{left} + SSE_{right}, \quad (7)$$

где SSE каждой части предполагаемого разбиения определялось по формуле:

$$SSE = \sum (y_i - \bar{y})^2 \quad (8)$$

Экспериментальные исследования и анализ результатов. Эксперименты проводились на ПЭВМ с математическими моделями RBF нейронной сети и решающего дерева, с целью выбора эффективной, для данной задачи модели. В качестве обучающей, использовалась выборка, содержащая шестнадцать различных демографических показателей, тем или иным образом влияющих на численность населения Донецкой области, таких как, например, рождаемость, смертность, миграция, средняя продолжительность жизни, выбросы в атмосферу вредных веществ, заболеваемость инфекционными заболеваниями, в том числе СПИДом и др. Выборка содержала наблюдения за ряд лет, начиная с 1950 до 2008 года [4].

В ходе исследования для моделирования нейронных сетей и решающих деревьев использовалась система математического моделирования MATLAB 7.1.

Результаты экспериментов приведены на рис. 6 – 7, где сплошной красной линией обозначена обучающая выборка, а синими кружками отмечены спрогнозированные значения.

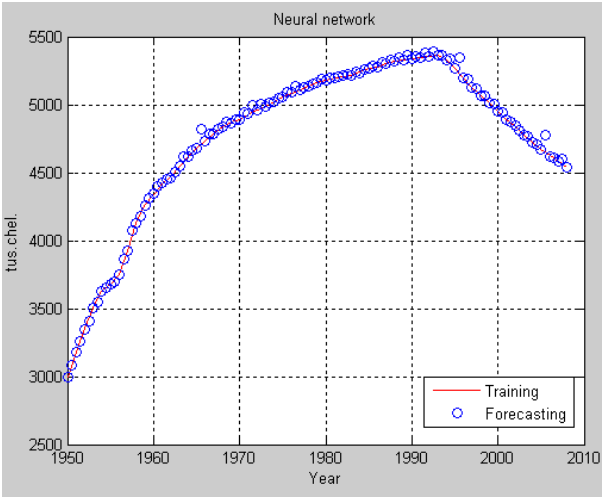

Рисунок 6 – Прогнозирование с использованием аппарата нейронных сетей

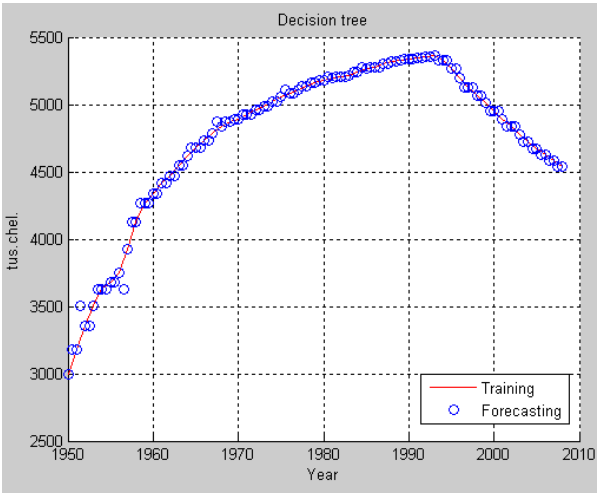

Рисунок 7 – Прогнозирование с использованием метода решающих деревьев

Построенная нейронная сеть имеет 16 входов, 50 нейронов в скрытом слое и один выходной нейрон. На вход подается обучающая выборка. Формируется сеть следующим образом: изначально первый слой не имеет нейронов. Сеть моделируется и определяется вектор входа с самой большой погрешностью, добавляется нейрон с радиально-базисной функцией активации и весами, равными вектору входа, затем вычисляются весовые коэффициенты линейного слоя, чтобы не превысить средней допустимой квадратичной ошибки, которая в нашем случае принималась равной 0,0001.

Расчет структуры решающего дерева определяется параметрами иерархической нелинейной регрессионной модели для матрицы независимых переменных X (обучающая выборка) и вектора значений зависимой переменной y (численность населения Донецкой области). Выходная переменная определяет бинарное дерево решений, в котором промежуточные узлы делятся ветвями на 2 возможных решения. В качестве условия выбора направления перехода выступает ограничение на значение независимой переменной.

Для выбора конкретной математической модели из двух, проведена оценка погрешности прогнозирования. Статистические характеристики качества каждой модели приведены в таблице 1:

Таблица 1

Статистические характеристики качества модели

	Аппарат нейронных сетей	Решающие деревья
Средняя квадратичная ошибка (SSE)	0,0113	0,0284
Средняя абсолютная ошибка (MAE)	0,0051	0,0069
Среднеквадратичная ошибка (MSE)	9,6714e-005	2,4310e-004
Среднеабсолютная процентная ошибка (MAPE)	0,51%	0,69%

Из сравнительной оценки показателей моделей можно сделать вывод, что построенные модели (как модели нейронных сетей, так и решающих деревьев) хорошо аппроксимируют фактические данные, т.е. они вполне отражают демографические тенденции, определяющие численность населения Донецкой области. Однако целесообразно использовать аппарат нейронных сетей для построения прогнозов высокого качества, т.к. для этой модели погрешность прогнозирования минимальна. В дальнейших исследованиях возможно применения метода главных компонент для отбора значащих факторов и с учетом этого, исследование точности результатов прогнозирования.

Литература

[1] Єріна А. М. Статистичне моделювання та прогнозування: Навч. посіб ник. / – К.: КНЕУ, 2001. – 170 с.

[2] Lendasse A. Approximation by radial basis function networks application.

[3] Manish Mehta, Rakesh Agrawal, Jorma Rissanen SLIQ: A Fast Scalable Classifier for Data Mining. / IBM Almaden Research Center, – 15 с.

[4] Населення Донецької області у 2005 році: [Демографічний щорічник. Головне управління статистики в Донецькій області / Відповідальний за випуск: Рак С.В.] – Донецьк: 2006. – 187 с.