**УДК 51-76+616.9**

## ОЦЕНКА ПОГРЕШНОСТИ НЕЙРО-НЕЧЕТКОГО ПРОГНОЗИРОВАНИЯ ВРЕМЕННЫХ РЯДОВ ЗАБОЛЕВАЕМОСТИ

***Котин В.В., Догадов А.А.***

МГТУ имени Н. Э. Баумана, Москва, Россия

Работа посвящена применению нейро-нечетких сетей для прогнозирования показателя инцидентности и оценке точности такого прогноза. Применение прогнозирования является актуальным в вопросах оценки эпидемиологической обстановки и принятия управленческих решений в сфере здравоохранения. Для решения поставленной задачи используется инструмент ANFIS, представленный в системе «Matlab» и служащий для разработки и исследования гибридных сетей. В рамках работы реализованы две модели адаптивных систем нечеткого ввода-вывода и произведена оценена точности прогнозирования по трем критериям. Данные для обучения и тестирования моделей представляют собой эпидемиологические данные по заболеваемости скарлатиной детей в возрасте до 14 лет в городе Москве.

*КЛЮЧЕВЫЕ СЛОВА*: гибридные сети, нейро-нечеткое прогнозирование, временны**е** ряды, эпидемиология, заболеваемость, математическое моделирование.

## ОЦІНКА ПОХИБКИ НЕЙРО-НЕЧІТКОГО ПРОГНОЗУВАННЯ ЧАСОВИХ РЯДІВ ЗАХВОРЮВАНОСТІ

***Котін В.В., Догадов А.А.***

Робота присвячена застосуванню нейро-нечітких мереж для прогнозування показника інцидентності та оцінці точності такого прогнозу. Застосування прогнозування є актуальним в питаннях оцінки епідеміологічної обстановки та прийняття управлінських рішень у сфері охорони здоров'я. Для вирішення поставленої задачі використовується інструмент ANFIS, який представлений в системі Matlab для розробки і дослідження гібридних мереж. В рамках роботи реалізовані дві моделі адаптивних систем нечіткого введення-виведення і зроблена оцінена точності прогнозування за трьома критеріями. Дані для навчання і тестування моделей представляють собою епідеміологічні дані по захворюваності на скарлатину дітей віком до 14 років в місті Москві.

*КЛЮЧОВІ СЛОВА*: гібридні мережі, нейро-нечітке прогнозування, часові ряди, епідеміологія, захворюваність, математичне моделювання.

## ERROR ESTIMATION OF NEURO-FUZZY TIME SERIES FORECASTING OF MORBIDITY

***Kotin V.V., Dogadov A.A.***

The work is dedicated to application of the neuro-fuzzy networks for forecast of the incidence index and assessing the accuracy of the prediction. Application of the forecasting is relevant in the epidemiological studies and managerial decision-making in health care. To solve the problem, a tool ANFIS, presented in Matlab and serving for research and development of hybrid networks is used. Two models of adaptive systems of fuzzy input-output are developed and the accuracy of the prediction made by three criteria is evaluated. The epidemiological data on the incidences of scarlet fever among children under the age of 14 in Moscow have been used for training and testing the models.

*KEYWORDS*: hybrid networks, neuro-fuzzy forecasting, time series, epidemiology, morbidity, mathematical modeling.

**1. Введение.** Проблемы, связанные с популяционной динамикой инфекционных заболеваний человека традиционно занимали важное место в науках о жизни [1]. В настоящее время актуальность этих проблем приобретает совершенно новые характеристики. Не углубляясь в детальный анализ причин такого рода актуальности, ограничимся кратким их перечислением: потенциальная угроза возможных масштабных техногенных катастроф и террористических актов; растущая интенсивность миграционных потоков, как контролируемых, так и неконтролируемых; высокая скорость трансконтинентальных коммуникаций населения; тесная связь заболеваемости с экономическими условиями, демографией, социальной структурой, традициями и поведенческими особенностями отдельных сообществ.

**2. Основная часть.** В эпидемиологии для описания заболеваемости используют различные показатели, например, показатель инцидентности [2], который рассчитывается как

$$I = \frac{A}{N} R, \qquad (1)$$

где $A$ – число новых случаев болезни, выявленных в определенной группе населения, $N$ – численность рассматриваемой группы населения, $R$ – размерность (100000 для просантимилле).

Регистрируемое число новых случаев болезни подвержено влиянию аддитивного шума:

$$A_n = A^*_n + v_n,$$

где $A^*_n$ – истинное число новых случаев болезни, в месяце $n$, $A_n$ – зарегистрированное число новых случаев болезни в месяце $n$, $v_n$ – шум.

Шум $v_n$ обусловлен ошибками, связанными с неправильной постановкой диагноза, случаями, когда пациент не обращался за медицинской помощью и случай не был зарегистрирован, ошибками в учете и регистрации пациентов.

Рассмотрим задачу получения прогноза числа новых случаев болезни в месяц $n$, обозначаемого в дальнейшем $\tilde{A}_n$, и показателя инцидентности в месяц $n$, обозначаемого в дальнейшем $\tilde{I}_n$ и рассчитываемому по формуле (1).

Прогноз может быть основан на модели процесса [3] или на модели данных (статистические методы, регрессии, анализ временных рядов). Нами была исследована и проведена параметрическая SIR-модель эпидемий и проведена ее параметрическая идентификация.[4]. Рассмотренная модель связывает между величины $S$ – количество здоровых, подверженных риску заболеть, $I$ – количество больных переносчиков инфекции, $R$ – количество выздоровевших, получивших иммунитет.

Прогнозируемое значение числа новых случаев болезни $\tilde{A}_n$, можно выразить через величины, входящие в SIR-модель:

$$\tilde{A}_n = S_{n-1} - S_n.$$

Точность такого прогноза зависит от точности начальных условий, точности параметров модели, допущений, принятых в модели. Так как неизвестна величина шума, то вопрос о точности начальных условий и параметров модели остается открытым.

В основу прогнозирования также можно положить модели адаптивных систем нейро-нечеткого вывода ANFIS, реализованных в среде Matlab [5]. ANFIS представляет собой систему нечеткого вывода, функции принадлежности в которой подбираются с помощью алгоритма обратного распространения ошибки [6].

В качестве данных для обучения и тестирования модели нами рассматривалась внутригодовая динамика заболеваемости скарлатиной среди детей до 14 лет в городе Москве. Данные по инцидентности за 1996 – 2008 гг. (рис.1.) предоставлены кафедрой эпидемиологии и доказательной медицины Первого МГМУ им. Сеченова [7].

В работе было реализовано две модели адаптивных систем нечеткого ввода–вывода. Для обучения обеих моделей было использовано 75% данных (с мая 1996 года по июнь 2006), а для тестирования 25% (с июля 2006 по декабрь 2008). Число циклов обучения было принято, равное трем. Исследовалась ошибка прогноза обеих моделей при введении разного числа лингвистических термов с разными функциями принадлежности для каждой входной переменной.

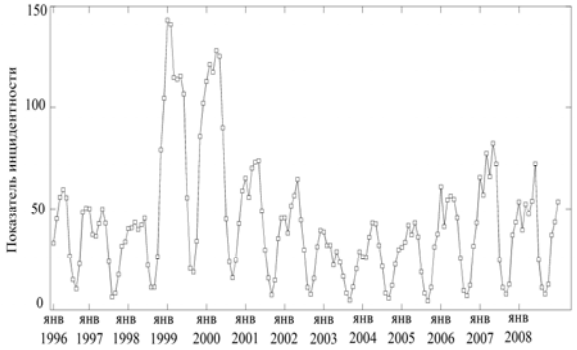


Рис.1. Данные по заболеваемости скарлатиной.

Первая модель имеет четыре нечетких входных переменных: показатели инцидентности за *i*-ый, *(i-1)*-ый, *(i-2)*-ой, *(i-3)*-ий месяц и одну выходную переменную, равную прогнозу значения показателя инцидентности в *(i+1)*-ый месяц. Во второй модели добавлена еще одна входная переменная – календарный номер месяца, на который осуществляется прогноз.

Ошибка прогноза обеих моделей исследовалась на данных для тестирования при помощи следующих критериев [8]:

1. Максимальное абсолютное отклонение

$$MAXD = \max\left(\left|\tilde{I}_n - I_n\right|\right),$$

где $n \in [1; T]$, $T$ – число прогнозов;

2. Корень средней квадратичной ошибки

$$RMSE = \sqrt{\frac{1}{T}\sum_{n=1}^{T}\left(\tilde{I}_n - I_n\right)^2};$$

3. Средняя абсолютная процентная ошибка

$$MAPE = \frac{1}{T}\sum_{n=1}^{T}\left|\frac{\tilde{I}_n - I_n}{I_n}\right| \cdot 100\%.$$

В таблице 1 приведены значения критериев точности первой модели, найденные для разного числа термов и вида функций принадлежности.

В таблице 2 приведены значения критериев точности второй модели, найденные для разного числа термов и вида функций принадлежности.

Таблица 1. Значения критериев точности первой модели.

| ***Вид функций принадлежности*** | ***2 терма*** | ***3терма*** |
|---|---|---|
| *Треугольная* | *MAXD=55,2 RMSE=16,4 MAPE=40,1 %* | *MAXD=97,6 RMSE=27,1 MAPE=71,1 %* |
| *Трапецеидальная* | *MAXD=67,5 RMSE=20,7 MAPE=66,3 %* | *MAXD=28,9 RMSE=14,2 MAPE=41,0 %* |
| *Обобщенная колоколообразная* | *MAXD=75,4 RMSE=22,4 MAPE=61,4 %* | *MAXD=58,9 RMSE=24,0 MAPE=101,6 %* |
| *Гауссова* | *MAXD=67,5 RMSE=18,3 MAPE =42,0 %* | *MAXD=56,5 RMSE=23,2 MAPE=99,0 %* |
| *Двойная гауссова* | *MAXD=69,4 RMSE=28,0 MAPE=90,8 %* | *MAXD=30,1 RMSE=13,6 MAPE=31,5 %* |
| *π-функция* | *MAXD=472,4 RMSE=92,6 MAPE=154,4 %* | *MAXD=32,3 RMSE=13,8 MAPE=35,0 %* |
| *Произведение 2 сигмоидальных функций* | *MAXD=67,7 RMSE=23,7 MAPE=93,9 %* | *MAXD=31,1 RMSE=13,6 MAPE=31,0 %* |

*Таблица 2. Значения критериев точности второй модели.*

| ***Вид функций принадлежности*** | ***2 терма*** | ***3 терма*** |
|---|---|---|
| *Треугольная* | *MAXD=35,2 RMSE=13,3 MAPE=35,6 %* | *MAXD=76,8 RMSE=21,8 MAPE=56,6 %* |
| *Трапецеидальная* | *MAXD=141,4 RMSE=40,0 MAPE=111,4 %* | *MAXD=33,1 RMSE=12,0 MAPE=26,2 %* |
| *Обобщенная колоколообразная* | *MAXD=101,9 RMSE=31,2 MAPE=107,1 %* | *MAXD=51,7 RMSE=21,4 MAPE=82,4 %* |
| *Гауссова* | *MAXD=41,0 RMSE=17,3 MAPE=58,1 %* | *MAXD=67,6 RMSE=26,7 MAPE=96,8 %* |
| *Двойная гауссова* | *MAXD=474,0 RMSE=111,4 MAPE=291,3 %* | *MAXD=28,1 RMSE=11,3 MAPE=24,8 %* |
| *π-функция* | *MAXD=242,0 RMSE=71,1 MAPE=190,3 %* | *MAXD=35,6 RMSE=13,3 MAPE=21,3 %* |
| *Произведение 2 сигмоидальных функций* | *MAXD=972,3 RMSE=213,5 MAPE=485,9 %* | *MAXD=32,5 RMSE=12,8 MAPE=25,0 %* |

**3. Выводы.** Таким образом, была опробована нейро-нечеткая система для прогнозирования показателя инцидентности и оценена ошибка прогнозирования. Наиболее точный по критериям *MAXD* и *RMSE* прогноз был получен при введении трех термов для каждой входной переменной и использовании двойной гауссовой функции принадлежности. Входными переменными являлись данные по заболеваемости за четыре предыдущих месяца и календарный номер месяца, на который производится прогнозирование. Наиболее точный прогноз по критерию *MAPE* был полчен при таком же наборе входных переменных и числе термов при введении π-функции принадлежности.

## ЛИТЕРАТУРА


1. Андерсон Р., Мэй Р. *Инфекционные болезни человека. Динамика и контроль*. Пер. с англ. Москва: Мир. – 2004. – 784 с.
2. *Клиническая эпидемиология с основами доказательной медицины. Руководство к практическим занятиям*. В.И. Покровский, Н.И. Брико (ред.). М: Гэотар–Медиа. – 2008. – 400 с.
3. Кравцов Ю.А. Случайность, детерминированность, предсказуемость. *Успехи физ. наук*. – 1989. – т. 158. – С. 92–121.
4. Догадов А.А., Котин В.В. Параметрическая идентификация SIR-модели внутригодовой динамики заболеваемости скарлатиной. *Научно-техническая конференция «Медико-технические технологии на страже здоровья». Сборник докладов*. М., Изд. НИИ РЛ МГТУ им. Н.Э.Баумана. – 2012. – С. 83–88.
5. Леоненков А.В. *Нечеткое моделирование в среде MATLAB и fuzzyTECH*. СПб. : БХВ-Петербург. – 2003. – 736 с.
6. http:www.mathworks.com/help/fuzzy/anfis-and-the-anfis-editor-gui.html [электронный ресурс].
7. Брико Н.И., Котин В.В. Ярынкина Т.А. Анализ периодичности и персистентности временных рядов заболеваемости. *Актуальные проблемы эпидемиологии на современном этапе*. – 2011. – С. 79–82.
8. Котин В.В., Ярынкина Т.А. Прогнозирование заболеваемости с использованием адаптивных и нечетких моделей. *Медицинская радиоэлектроника*. – 2012 – N11. – С. 13–22.
9. Дмитриев А.Н., Котин В.В. Моделирование временных рядов заболеваемости с использованием искусственных нейронных сетей. *Медицинская техника*. – 2013 – N1. – С. 35–37.
10. Вьюн В.И., Еременко Т.К., Кузьменко Г.Е., Михненко Ю.А. Об одном подходе к прогнозированию эпидемиологической обстановки по гриппу–ОРВИ с использованием временных рядов. *Математичні машини і системи*. – 2011. – N 2. – С. 131–136.