Krassimir Markov, Vladimir Ryazanov,
Vitalii Velychko, Levon Aslanyan
(editors)

New Trends in Classification and Data Mining

I T H E A
SOFIA
2010

Krassimir Markov, Vladimir Ryazanov, Vitalii Velychko, Levon Aslanyan (ed.)

New Trends in Classification and Data Mining

ITHEA®

Sofia, Bulgaria, 2010

First edition

Recommended for publication by The Scientific Concil of the Institute of Information Theories and Applications FOI ITHEA

This book maintains articles on actual problems of classification, data mining and forecasting as well as natural language processing:

- new approaches, models, algorithms and methods for classification, forecasting and clusterisation. Classification of non complete and noise data;
- discrete optimization in logic recognition algorithms construction, complexity, asymptotically optimal algorithms, mixed-integer problem of minimization of empirical risk, multi-objective linear integer programming problems;
- questions of complexity of some discrete optimization tasks and corresponding tasks of data analysis and pattern recognition;
- the algebraic approach for pattern recognition - problems of correct classification algorithms construction, logical correctors and resolvability of challenges of classification, construction of optimum algebraic correctors over sets of algorithms of computation of estimations, conditions of correct algorithms existence;
- regressions, restoring of dependences according to training sampling, parametrical approach for piecewise linear dependences restoration, and nonparametric regressions based on collective solution on set of tasks of recognition;
- multi-agent systems in knowledge discovery, collective evolutionary systems, advantages and disadvantages of synthetic data mining methods, intelligent search agent model realizing information extraction on ontological model of data mining methods;
- methods of search of logic regularities sets of classes and extraction of optimal subsets, construction of convex combination of associated predictors that minimizes mean error;
- algorithmic constructions in a model of recognizing the nearest neighbors in binary data sets, discrete isoperimetry problem solutions, logic-combinatorial scheme in high-throughput gene expression data;
- researches in area of neural network classifiers, and applications in finance field;
- text mining, automatic classification of scientific papers, information extraction from natural language texts, semantic text analysis, natural language processing.

It is represented that book articles will be interesting as experts in the field of classifying, data mining and forecasting, and to practical users from medicine, sociology, economy, chemistry, biology, and other areas.

General Sponsor: Consortium FOI Bulgaria (www.foibg.com).

Printed in Bulgaria

ISBN 978-954-16-0042-9

C\o Jusautor, Sofia, 2010

НЕЙРОСЕТЕВАЯ АРХИТЕКТУРА НА ЧАСТИЧНЫХ ОБУЧЕНИЯХ

Николай Мурга

Аннотация: *рассматривается нейросетевая архитектура, обучение в которой происходит не с целью минимизации единого критерия качества, а с разбиением выборки данных на подмножества, для каждого из которых происходит обучение с целью минимизации своего критерия. Рассматривается применение сети к анализу и прогнозированию валютных пар EUR/GBP, EUR/USD, USD/CHF, USD/JPY.*

Ключевые слова: *нечёткая логика, вывод Такаги-Сугено, обучение нейронных сетей, кластеризация, прогноз значений валютных пар.*

ACM Classification Keywords: *G.1.0 Mathematics of Computing – NUMERICAL ANALYSIS – General – Error analysis; G.1.2 Mathematics of Computing – NUMERICAL ANALYSIS – Approximation – Least squares approximation; G.1.6 Mathematics of Computing – NUMERICAL ANALYSIS – Optimization - Gradient methods, Least squares methods; I.2.3 Computing Methodologies - ARTIFICIAL INTELLIGENCE - Deduction and Theorem Proving - Uncertainty, "fuzzy", and probabilistic reasoning; I.2.6 Computing Methodologies - ARTIFICIAL INTELLIGENCE – Learning - Connectionism and neural nets; I.5.3 - Computing Methodologies - PATTERN RECOGNITION - Clustering.*

Вступление

Данная работа посвящена модификации классического метода обучения нейросетевых архитектур, который базируется на минимизации критерия (критериев) качества для всей обучающей выборки данных. В противовес этому подходу, предлагается разбиение всей выборки на непересекающиеся подмножества, на которых и происходит минимизация критерия (критериев), зависящих лишь от значений точек данных подмножеств. В работе рассматривается нечёткая система, с механизмом нечёткого логического вывода Такаги-Сугено с тригонометрическими полиномами в «то»-части нечётких правил. Поставлены эксперименты для анализа свойств рассмотренной архитектуры на периодических зависимостях и случайных процессах. Произведён анализ приложения сети к дневным котировкам валютных пар EUR/GBP, EUR/USD, USD/CHF, USD/JPY за период с 25.03.2009 по 24.03.2010, которые были взяты из [1]. Анализировалась прогностическая способность сети на основании анализа значений критериев RMSE и MAPE.

1 Архитектура, метод обучения и анализ качества работы предлагаемой нейронной сети

Данный раздел является теоретическим. Он состоит из двух подразделов. В первом подразделе описываются: используемая в работе нейросетевая архитектура и метод её обучения; второй подраздел посвящён описанию и анализу критериев качества работы сети.

1.1 Описание архитектуры и метода обучения сети

Предлагаемая нечёткая (гибридная) нейронная сеть использует в качестве механизма нечёткого вывода механизм нечёткого вывода Такаги-Сугено (TS).

База нечётких правил TS выглядит следующим образом:

$$R_1: Если\ x_1 \in A_1^{(1)}, x_2 \in A_2^{(1)}, \dots, x_n \in A_n^{(1)}, то\ y_1 = \sum_{j=1}^{n} f_j^{(1)}(x_j)$$
$$\dots$$
$$R_K: Если\ x_1 \in A_1^{(K)}, x_2 \in A_2^{(K)}, \dots, x_n \in A_n^{(K)}, то\ y_K = \sum_{j=1}^{n} f_j^{(K)}(x_j) \quad (1)$$

где R_i - это -е нечёткое правило ($i = \overline{1,K}$); K - это количество нечётких правил; x_j – j-я компонента входного вектора; n – размерность входного пространства; y_i – выход i-го правила; A_j^i – значение лингвистической переменной x_j для правила R_i с симметричной функцией принадлежности.

Про зависимости $f_j^{(i)}(x_j)$ следует сказать, что для классической сети TSK ([2], [3]) они имеют фиксированный порядок и обычно либо линейные функции, либо – функции-константы. В данной работе данные зависимости имеют нефиксированный порядок и представляют собой тригонометрические полиномы.

Как отмечается в работе [3], тригонометрическим полиномом порядка M называется следующее выражение:

$$T_M(x) = \frac{a_0}{2} + \sum_{k=1}^{M}(a_k \cos kx + b_k \sin kx) \quad (2)$$

С учётом новых обозначений, вышеуказанные зависимости $f_j^{(i)}(x_j)$ можно записать в виде:

$$f_j^{(i)}(x_j) = A_j^{(i)} \cdot T_{M(i,j),j}^{(i)}(x_j) \quad (3)$$

Коэффициенты $a_{k,j}^{(i)}$, $b_{k,j}^{(i)}$ и $A_j^{(i)}$ находятся при помощи метода наименьших квадратов (используется метод скорейшего спуска [4]). Порядок $M(i,j)$, как уже не раз отмечалось, не фиксирован, в начале работы алгоритма он принимается равным 1 и начинает расти, пока не будет достигнута заданная точность либо заданное максимальное значение.

Однако новизна работы не в этом. В отличие от классической сети TSK, нахождение $a_{k,j}^{(i)}$, $b_{k,j}^{(i)}$ и $A_j^{(i)}$ выполняется не на всех $x_i = (x_{1i} \ \dots \ x_{ni}), l = \overline{1,L}$, а лишь на основании соответствующего подмножества. Это требует детального пояснения. Цель обучения классической сети TSK – минимизация функционала E:

$$E = \frac{1}{2}\sum_{l=1}^{L}(y_l - d_l)^2 \quad (4)$$

где y_l - реальный выход сети, а d_l – желаемое значение (стоит отметить, что обучающая выборка данных представляет собой набор пар $(x_l, d_l), l = \overline{1,L}$). Однако функционал (4) задаёт поиск некоторой «усреднённой» закономерности данных, что не всегда адекватно делать на реальных данных. В противовес такому подходу предлагается следующий. Производится кластеризация данных обучающей выборки. Вся обучающая выборка разбивается на K (где K - это количество кластеров) подмножеств по признаку принадлежности точек выборки определённому кластеру (наибольшему значению принадлежности). Таким образом, каждая (x_l, d_l) принадлежит некоторому множеству точек S_K – каждому отдельному кластеру i соответствует множество S_i. А из того, что каждый кластер i задаёт каждое правило i (компоненты центра кластера – центры функций принадлежности соответствующих входов для

данного правила), то каждому правилу i соответствует множество S_i. Задача обучения сети сводится к минимизации K функционалов вида:

$$E_{S_i} = \frac{1}{2}\sum_{l=1}^{L_i}\left(y_l^{(i)} - d_l^{(i)}\right)^2 \quad (5)$$

где L_i - количество точек обучающей выборки, принадлежащих i -му кластеру, а индекс (i) над компонентами y_l и d_l обозначает, что x_l принадлежит кластеру i.

Необходимо отметить тот факт, что в общем случае верно следующее неравенство:

$$\min E \neq \sum_{i=1}^{K} \min E_{S_i} \quad (6)$$

Схематически метод формирования описанной в данном разделе системы нечёткого логического вывода можно представить в следующем виде (см. Рис. 1).

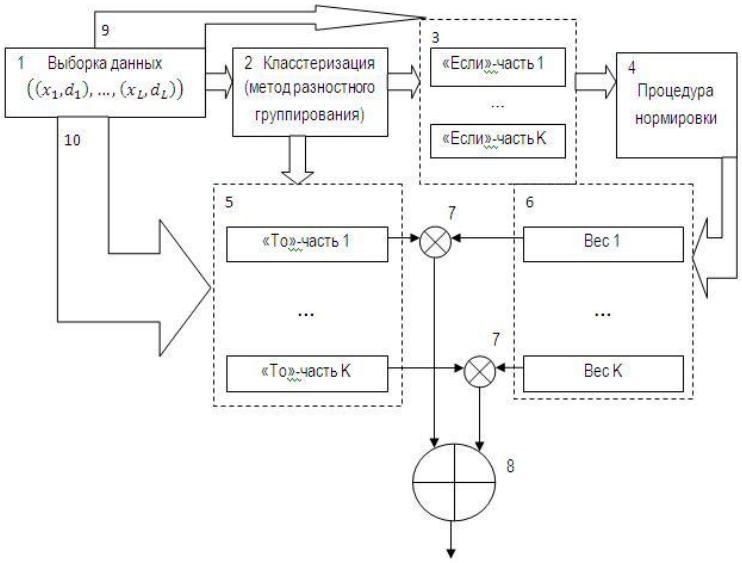


Рис. 1. Схематическое изображение предложенного метода

Описание схемы Рис. 1 следующее.

Этап обучения сети. Имеется выборка данных, состоящая и пар (x_i, d_i) (блок 1), компоненты x_i которой подаются на вход блока кластеризации (блок 2). В качестве метода кластеризации выбран метод разностного группирования со сферой влияния равной 0,5 ([2], [3]). В результате кластеризации получается K класстеров, центры которых становятся центрами функций принадлежности (выбраны функции принадлежности гауссовского вида) $\mu_{A_j^{(i)}}, i = \overline{1,K}, j = \overline{1,n}$. Параметры σ ([2], [3]) этих функций принадлежности выбираются одинаковыми для каждого отдельного x_j для всех правил. Так строится «Если»-часть нечётких правил (Блок 3). Для построения «То»-части нечётких правил, прежде всего, производится разбивка данных обучающей выборки на описанные ранее непересекающиеся подмножества S_i - разбиение происходит на основании максимальной принадлежности определённой точки выборки определённому кластеру. Далее, для каждого подмножества S_i, а следовательно, для

каждого правила i происходит обучение, целью которого является минимизация функционала E_{S_i}, который вычисляется по формуле (5) (Этот этап символизирует переход к блоку 5 из блока 2 и блок-переход 10). Этап обучения завершён.

Этап использования сети. Компоненты x_l пар (x_l, d_l) подаются при помощи перехода 9 на блок 3. Это символизирует расчёт пересечения значения термов для каждого правила. В реализации метода, использованного в работе, применяется пересечение в форме произведения значений функций принадлежностей термов. Так получаются веса нечётких правил. Однако для того, чтобы сумма весов правил была всегда равна 1, для каждого x_l выполняется процедура пересчёта весов (нормировки весов, что, де-факто, является просто делением каждого отдельного веса на сумму всех весов правил), что символизируют блоки 4 и 6. Рассчитываются значения y_l в «То»-части нечётких правил по полученным в результате обучения сети формулам (эти зависимости представлены символически в выражении (1)). Это символизирует блок 5, в который подаются x_l из блока 1 по блоку-переходу 10. Далее полученные y_l умножаются на полученные ранее нормированные веса нечётких правил и суммируются, что даёт выходы сети y_l. Символически это обозначено операциями-блоками 7 и 8. Далее, применяя анализ значений определённых функционалов качества, делается заключение о том, насколько значения y_l удалены от d_l, то есть, проводится анализ качества работы сети, что и является предметом экспериментальных исследований предложенных ниже в данной работе.

1.2 Анализ качества работы нейронной сети

Анализ качества работы сети в данной работе будет производиться на основе значений критериев RMSE и MAPE.

Критерий RMSE, при условии, что d_l - желаемый выход сети, а y_l - реальный выход сети, где $l = \overline{1,L}$; L - объём выборки данных вычисляется по формуле:

$$RMSE = \sqrt{\frac{1}{L}\sum_{l=1}^{L}(y_l - d_l)^2} \quad (7)$$

Данный критерий является оценкой абсолютного отклонения выдаваемых сетью значений от значений, которые она бы должна выдавать в идеале. Очевидным является тот факт, что критерий RMSE является чувствительным к масштабу данных, как и все критерии, оценивающие абсолютное отклонение. Для объяснения необходимо лишь привести следующих два примера: $\sqrt{\frac{1}{1}\sum_{l=1}^{1}(0{,}5-0{,}4)^2} = 0{,}1$, в то время, когда $\sqrt{\frac{1}{1}\sum_{l=1}^{1}(95-94)^2} = 1$. То есть, одно и тоже значение данного критерия может говорить как о высоком качестве работы сети, так и очень плохом, так как масштаб задаёт вес каждой отдельной цифры в числе.

Критерий MAPE при условии, что $d_l \geq 0$ - желаемый выход сети, а $y_l > 0$ - реальный выход сети, где $l = \overline{1,L}$; L - объём выборки данных вычисляется по формуле:

$$MAPE = \frac{1}{L}\sum_{l=1}^{L}\frac{|y_l - d_l|}{d_l} \quad (8)$$

Данный критерий является оценкой относительного отклонения выдаваемых сетью значений от желаемых значений. Критерий, в отличие от рассмотренного ранее, не чувствителен к масштабу чисел,

хотя и определяет косвенно количество верно распознанных цифр в числе в порядке убывания их значимости. Если значение критерия умножить на 100%, то будет получен средний процент значений отклонений значений выдаваемых сетью от желаемых значений.

Аббревиатуры критериев расшифровываются следующим образом. RMSE – root mean square error – корень квадратный из среднеквадратического отклонения. MAPE – mean absolute percentage error – средняя абсолютная процентная ошибка. Слово «абсолютная» вызывает недоумение – ведь критерий относительный, однако данное слово здесь обозначает, что берётся сумма модулей отклонений поделенных на соответствующие реальные значения, на что указывает слово «процентная». В работе будут использованы английские аббревиатуры из-за того, что они гораздо чаще используются в научной литературе, чем русские.

Ещё следует отметить такой момент, что, вследствие наличия у обоих критериев компоненты $y_i - d_i$, в экспериментах можно будет наблюдать некоторое подобие динамики критериев в зависимости от изменения параметров экспериментов. Однако необходимо изучение обоих критериев, так как они выполняют разные функции. Функция RMSE – описать то, насколько сеть хорошо обучена, а цель MAPE – описать то, насколько велико отклонение реальных значений, выдаваемых сетью от желаемых значений.

2 Экспериментальные исследования

Данный раздел посвящён экспериментальным исследованиям сети и состоит из двух подразделов. Первый подраздел является теоретическим исследованием свойств сети и свойств экспериментальной среды, которые влияют на качество функционирования сети. Второй подраздел посвящён приложению рассматриваемой нейросетевой архитектуры и метода обучения к задаче прогноза значений котировок валютных пар EUR/GBP, EUR/USD, USD/CHF, USD/JPY.

2.1 Теоретические исследования

Данный подраздел состоит из двух подразделов. Первый подраздел посвящён выявлению свойств рассматриваемой нейросети и среды эксперимента, которые влияют на качество работы сети путём приложения её к случайно сгенерированным данным. Второй раздел посвящён анализу сети на основании приложения её к периодическим зависимостям.

2.1.1 Экспериментальное исследование сети на случайных входных и выходных данных

Для выявления и анализа скрытых свойств функционирования сети проводилось её применение для анализа и поиска зависимостей в данных, в которых входы и выходы сети были случайными равномерно распределёнными величинами. Следует сразу заметить, что в данном случае, в отличие от последующих в данной работе исследований, значения обоих критериев (RMSE и MAPE) будут очень большими. И это обоснованно – нельзя искать закономерность там, где её нет. Отличие данного исследования от последующих в том, что для него не важны конкретные значения критериев, а важно – поведения данных значений в зависимости от изменений параметров экспериментальной модели.

Эксперимент построен следующим образом. Количество входов сети является параметром экспериментальной модели и изменяется от 2 до 4. Количество точек выборки – параметр эксперимента и принимает значения 10, 20, 30,…, 80, 90, 100. Значения точек выборок данных генерируются с помощью генератора случайных чисел (равномерное распределение); значения принадлежат интервалу – [0;1]. Делать выводы, проведя лишь одно исследование для конкретных значениях параметров эксперимента – неадекватно, в силу случайной природы исследуемых данных; в то же время, проведение огромного

числа опытов при конкретных значениях параметров эксперимента и расчёт их значений и дальнейший расчёт средних по значениям критериев – тоже неадекватно, так как с ростом количества рассмотренных вариантов падает влияние каждого отдельного варианта на среднее значений критериев. Из этих рассуждений для количества обучений и проверки сети при конкретных значениях параметров было выбрано число 10 – не один опыт и, в то же время, каждый отдельный опыт имеет значительное влияние на среднее критериев. Таким образом, как уже стало понятно из предшествующих объяснений, для каждых конкретных параметров экспериментов случайно генерируются данные выборок и количество таких выборок данных равно 10. Далее сеть обучается на данных выборках, считаются значения критериев RMSE и MAPE для каждой отдельной выборки, а после – происходит усреднение данных критериев по всем 10 реализациям. Деление выборок на обучающие и проверочные не проводилось из-за бессмысленности подобного деления для данного эксперимента. Параметры самой сети: максимальный порядок частичного описания – 3, допустимая погрешность – 0,00001. Выбор столь малых значений был обусловлен желанием изучить свойства сети на данном эксперименте так, чтобы минимизировать влияние самих параметров сети на результаты.

Результаты опытов представлены в следующей таблице.

Таблица 1. Результаты экспериментальных исследований на выборках данных, состоящих из случайных чисел

Количество точек в выборке	**Количество входов сети**					
	2		**3**		**4**	
	RMSE	**MAPE**	**RMSE**	**MAPE**	**RMSE**	**MAPE**
10	0,141	0,987	0,104	1	0,0586	0,49
20	0,326	5,36	0,154	2,24	0,216	2,53
30	0,22	4,2	0,174	3,27	0,158	3,67
40	0,166	3,78	0,126	2,93	0,117	2,35
50	0,147	4,04	0,141	3,19	0,157	3,12
60	0,145	3,8	0,115	2,77	0,111	3,37
70	0,11	2,98	0,116	3,72	0,116	3,85
80	0,117	3,46	0,113	3,46	0,097	4,21
90	0,0943	3,65	0,12	3,48	0,0854	3,6
100	0,109	4,26	0,0809	3,62	0,0926	3,64

Прежде всего, необходимо перечислить факты, которые следуют из приведённой выше таблицы данных. Значения критериев RMSE и MAPE говорят, что в то время, когда сеть достаточно хорошо настроена (следует напомнить, что данные из диапазона [0;1]) – об этом говорят значения критерия RMSE - сеть не делает ни одного точного прогноза и даёт очень большие ошибки – об этом говорят значения критерия MAPE. Вторым важным фактом является то, что при количестве точек выборки равном 10, значения критериев значительно отличаются от значений критериев при прочих значениях данного параметра

эксперимента – сеть относительно точно распознаёт некоторые ситуации. Третий факт - MAPE не проявляет никаких закономерностей при изменениях значений параметров эксперимента, кроме рассмотренного выше варианта. Четвёртый – значения критерия RMSE падают с ростом объёма выборки и ростом числа входов сети.

Объяснения первого, третьего и четвёртого факта лежат в следующих рассуждениях. Сеть обучается на критерии RMSE, а данный критерий характеризует отклонение точек от «некоторого тренда» в данных – и применение данного критерия адекватно для анализа случайных чисел. С ростом объёма выборки получаемый «тренд» становится ближе к «идеальному тренду» и, следовательно, описательное качество сети по данному критерию улучшается. С ростом числа входов сети – ситуация аналогична. Однако критерий MAPE для анализа работы сети на случайных числах применять нельзя, так как он является мерой отклонения реальных результатов работы сети от предложенной сетью нестатистической закономерности, а как можно требовать от сети такую закономерность, когда она не существует a priori. Второй факт говорит о том, что сеть при данных конфигурациях и количестве входов сети лишь запоминает все точки выборки – «зубрит» их, а не находит в них скрытые закономерности; однако и «зубрёжка» эта не всегда эффективна.

2.1.2 Экспериментальное исследование сети на периодических зависимостях выхода от входов

Главная цель проведения данного экспериментального исследования – определить: способна ли сеть описывать периодические зависимости.

Для данного эксперимента, по сравнению с экспериментом со случайными числами, на первый план выходят значения критериев RMSE и MAPE. Однако поведения критериев на периодических зависимостях тоже важны.

Эксперимент был построен следующим образом. Параметры эксперимента идентичны параметрам эксперимента со случайными числами и, наверное, нет смысла их детально описывать. Следует напомнить, что это – количество входов сети и число точек выборки данных. Деление выборки данных на обучающую и проверочную не производилось, поскольку для данного эксперимента важно исследование описательного свойства сети.

Отдельно следует рассмотреть то, как строились выборки данных. Прежде всего, определялось количество входов сети для исследования. Далее задавался шаг для построения сетки точек пространства. Следует отметить, что значение каждого входа сети принадлежало интервалу [0;1]. После этого, с учётом выбранного шага, строилась сетка, покрывающая единичный гиперкуб входных данных. Данная сетка дополнялась случайным количеством дублей случайно выбранных точек сетки. Из построенного дискретного пространства входных векторов случайным образом выбирались точки в количестве, задаваемым вторым параметром эксперимента – количеством точек выборки. Выход сети строился по формуле:

$$y_j = \sum_i f_i(x_{ij}) \tag{9}$$

где

$$f_i(x_{ij}) = rand_0 + \sum_{k=1}^{2}(rand_{2k-1} \cdot \cos kx_{ij} + rand_{2k} \cdot \sin kx_{ij}) \tag{10}$$

В данном эксперименте $rand_i$ - были случайными целыми числами, выбираемыми из диапазона [-5;5].

Аналогично эксперименту со случайными числами, опираясь на те же рассуждения, для конкретных значений параметров эксперимента вышеуказанные выборки строились по 10 раз, рассчитывались значения критериев качества сети RMSE и MAPE и в нижеприведённую таблицу записывались усреднённые значения критериев. Параметры сети аналогичны параметрам из эксперимента со случайными числами.

Следующая таблица содержит результаты проведения опытов.

Таблица 2. Результаты экспериментальных исследований на выборках данных периодических зависимостей

Количество точек в выборке	**Количество входов сети**					
	2		**3**		**4**	
	RMSE	**MAPE**	**RMSE**	**MAPE**	**RMSE**	**MAPE**
10	0,263	0,582	0,367	0,326	0,338	0,168
20	0,355	0,265	0,364	0,375	0,283	0,174
30	0,207	0,425	0,27	0,897	0,243	0,901
40	0,153	0,233	0,228	0,608	0,249	0,335
50	0,153	0,607	0,208	0,623	0,265	0,441
60	0,121	0,393	0,233	0,929	0,394	33,7
70	0,0888	0,43	0,197	0,498	0,211	0,804
80	0,103	0,189	0,143	0,711	0,255	0,635
90	0,0795	0,205	0,163	0,705	0,204	0,614
100	0,0682	0,88	0,154	0,472	0,19	1,19

Полученные результаты очень интересны и они ни в коем случае не говорят о неспособности сети описывать периодические зависимости. В постановке эксперимента был заложен один очень важный момент, который дал такие большие значения MAPE и о котором будет сказано чуть позже. Сейчас следует сразу сказать, что с устранением данного момента сеть давала значения критерия MAPE равное 10^{-2}, когда использовался вывод TS и 10^{-5}, когда использовался нечёткий вывод Белмана-Задэ. А момент этот – в покрытии точками выборки всего входного пространства. Прежде всего, введение в пространство дублей точек сетки придавало отдельным точкам выборки, в случае попадания в выборку двух одинаковых векторов – больший вес, а значит, сеть больше обучалась под эти точки, или вернее сказать точку выборки, чем под другие точки. И попадание в выборку нескольких таких точек и приводит к среднему значению MAPE равному, например, 33,7. Также, следует отметить, что при построении сетки для 2 входов, 3 и 4 входов использовался один и тот же шаг – 0,01, а количество точек выборки – тоже было одно и то же, и являлось параметром эксперимента. Таким образом, с ростом числа входов сети росла вероятность того, что будут набраны практически все точки из одного подпространства входного пространства и несколько точек относительно удалённых от данного пространства. А это обуславливает рост средних значений параметра MAPE в эксперименте с ростом числа входов сети. Этот факт является

очень поучительным – он говорит о том, что для корректного функционирования сети требуется как можно больше обучающих данных, которые как можно полнее описывают входное пространство. Ещё один нюанс, который следует объяснить, почему при 100 точках выборки с двумя входами средний MAPE получился равным 0,88, что означает, де факто, 88% ошибок, а проблема эта в наибольшей концентрации в выборке из 100 точек дублей, которые мешают верной работе сети.

В окончание подраздела следует ещё раз подчеркнуть тот факт, что подобная ситуация смоделирована искусственно и при правильном построении выборки, которая очень хорошо представляет определённую периодическую зависимость, были получены порядки точности 10^{-5}, при условии, что ограничение для алгоритма обучения сети на точность было равно 10^{-5}. Также следует отметить, что полагаться лишь на критерий RMSE как на единственный критерий качества работы сети нельзя, так как он описывает лишь то, насколько хорошо обучена сеть и насколько значения, выдаваемые сетью, близки к реальным значениям в среднеквадратическом смысле. Анализ качества работы сети всегда должен подкрепляться анализом критерия MAPE.

2.2 Применение сети к анализу данных котировок валют

Для проверки эффективности работы рассматриваемой нейронной сети относительно практических задач была выбрана задача прогноза котировок валют на основании их предыстории.

Постановка конкретной задачи, которая решалась в данной главе следующая. Имеется выборка данных дневных котировок валютных пар EUR/GBP, EUR/USD, USD/JPY, USD/CHF за период с 25.03.2009 по 24.03.2010. Необходимо произвести прогноз значения котировки на шаг вперёд на основании её предыстории.

Эффективность работы рассмотренной ранее нейросетевой архитектуры применительно к решению данной задачи оценивалась по двум критериям: RMSE и MAPE. Следует отдельно отметить, что, согласно алгоритму обучения сети, сеть настраивалась с целью минимизации, де-факто, критерия RMSE. Таким образом, критерий MAPE выступает в роли «независимого эксперта», то есть оценивает качество работы сети, при этом не участвуя в её обучении. Выходит, RMSE является критерием эффективности настройки сети, в то время, когда MAPE оценивает эффективность её приложения.

Эксперимент был поставлен следующим образом. Из указанных выше выборках данных котировок валют составлялись новые выборки данных для применения рассматриваемой нейросетевой архитектуры. Выбиралось количество периодов в предыстории, на основании которых необходимо сделать прогноз – в эксперименте это: 2, 3, 4 дня – это входы сети. Выходом сети являлось значение котировки в последующий момент времени. Указанные выборки данных котировок валют делились на обучающие и проверочные в соотношениях: 10:90, 20:80, 30:70, 40:60, 50:50, 60:40, 70:30, 80:20, 90:10 соответственно. Это означает, что, например, для котировки EUR/USD соотношение 10:90: 10% выборки – обучающая, а 90% выборки – проверочная. Сеть обучалась на обучающих выборках и фиксировались значения вышеуказанных критериев на обучающих и проверочных выборках при данных их соотношениях и при различном числе входов сети.

Результаты эксперимента для котировок EUR/GBP, EUR/USD, USD/JPY, USD/CHF представлены ниже.

Таблица 3. Значения критериев для валютной пары EUR/GBP

обуч:пров	Количество входов											
	2				3				4			
	RMSE обуч.	MAPE обуч.	RMSE пров.	MAPE пров.	RMSE обуч.	MAPE обуч.	RMSE пров.	MAPE пров.	RMSE обуч.	MAPE обуч.	RMSE пров.	MAPE пров.
10:90	0,002608	0,013772	0,006047	0,071061	0,002109	0,010416	0,007302	0,084142	0,001473	0,007201	0,006759	0,079349
20:80	0,000641	0,004802	0,000449	0,006671	0,002954	0,02282	0,001722	0,026742	0,000654	0,004785	0,000593	0,0081
30:70	0,000545	0,005095	0,000351	0,004864	0,000463	0,004413	0,000343	0,004695	0,000461	0,004444	0,000362	0,005079
40:60	0,000372	0,003944	0,000341	0,004363	0,000389	0,004317	0,00034	0,00437	0,000403	0,004289	0,000413	0,004969
50:50	0,000353	0,004268	0,000429	0,004994	0,000339	0,003984	0,000343	0,003943	0,000337	0,003889	0,000318	0,003563
60:40	0,000345	0,004362	0,000342	0,003619	0,000325	0,004067	0,000333	0,003463	0,000338	0,004351	0,000384	0,004155
70:30	0,000318	0,004367	0,000372	0,003392	0,000287	0,003816	0,000367	0,003277	0,000331	0,004526	0,000482	0,004493
80:20	0,000287	0,004209	0,00051	0,003865	0,000265	0,003797	0,000475	0,003548	0,000262	0,003702	0,000469	0,003473
90:10	0,000245	0,003676	0,000681	0,003759	0,000251	0,00384	0,000615	0,003481	0,000312	0,005037	0,001098	0,00593

Таблица 4. Графическое представление результатов экспериментов для валютной пары EUR/GBP

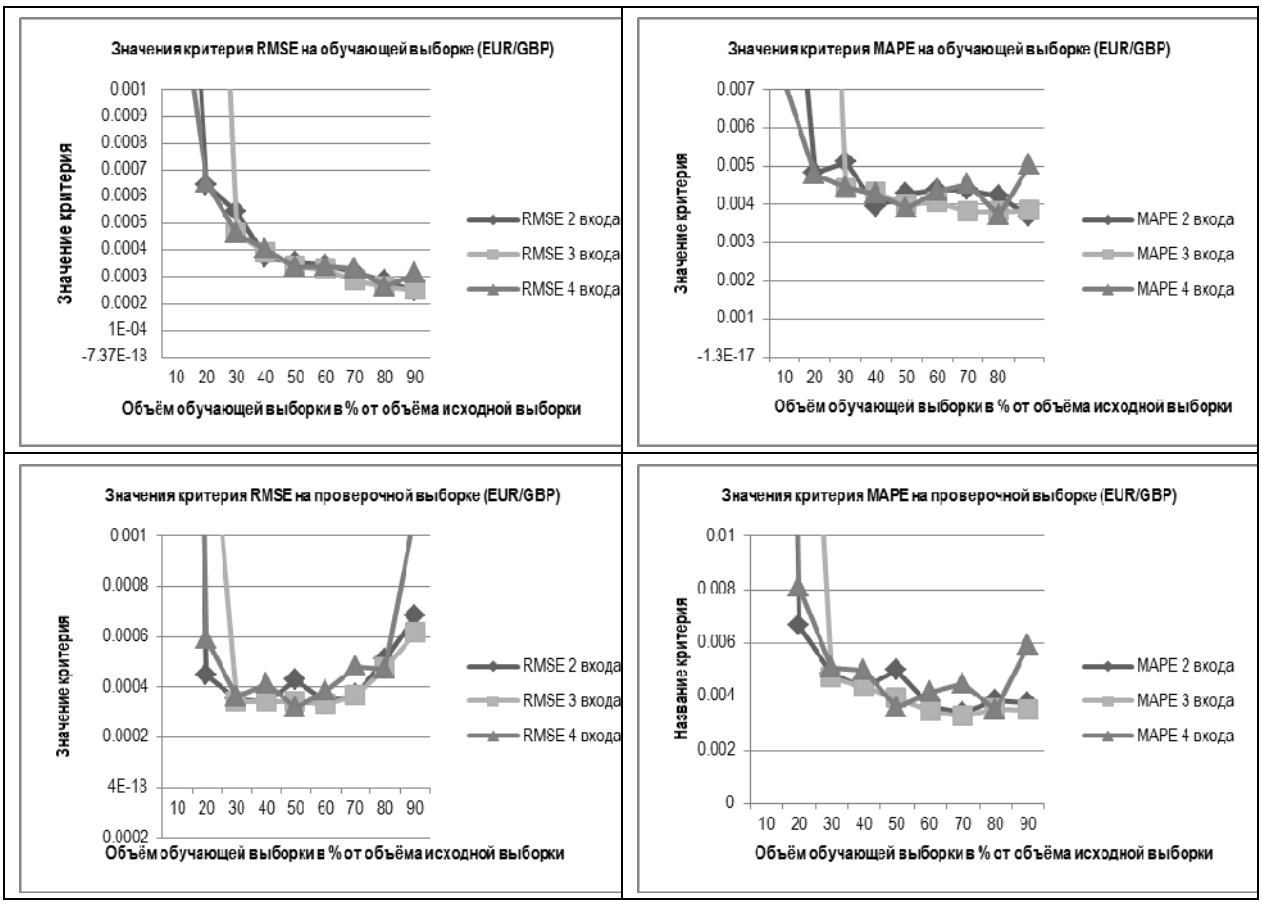


Таблица 5. Значения критериев для валютной пары EUR/USD

обуч:пров	Количество входов											
	2				3				4			
	RMSE обуч.	MAPE обуч.	RMSE пров.	MAPE пров.	RMSE обуч.	MAPE обуч.	RMSE пров.	MAPE пров.	RMSE обуч.	MAPE обуч.	RMSE пров.	MAPE пров.
10:90	0,002596	0,008119	0,02675	0,22567	0,001805	0,006454	0,03573	0,33629	0,001937	0,007164	0,038888	0,39704
20:80	0,001114	0,00528	0,004944	0,034266	0,001904	0,008783	0,013127	0,095096	0,001113	0,005631	0,003617	0,026909
30:70	0,001036	0,006263	0,003192	0,024584	0,000977	0,005755	0,002153	0,016265	0,000884	0,005047	0,002987	0,021669
40:60	0,000811	0,005485	0,002073	0,013896	0,000723	0,004796	0,001985	0,013074	0,000727	0,004817	0,003997	0,025057
50:50	0,000862	0,006316	0,00123	0,007417	0,001077	0,008568	0,00163	0,010054	0,001065	0,0083	0,002203	0,014389
60:40	0,000743	0,006218	0,000734	0,004903	0,000584	0,004721	0,000883	0,006112	0,000547	0,004388	0,000778	0,0052
70:30	0,000543	0,004711	0,000687	0,003757	0,000506	0,004413	0,000798	0,004429	0,000611	0,005103	0,000882	0,005125
80:20	0,000485	0,004515	0,00082	0,003721	0,00047	0,00436	0,001017	0,004823	0,00055	0,005184	0,001069	0,005254
90:10	0,000534	0,005175	0,001559	0,006103	0,00042	0,003925	0,00121	0,003785	0,000425	0,004014	0,00113	0,003547

Таблица 6. Графическое представление результатов экспериментов для валютной пары EUR/USD

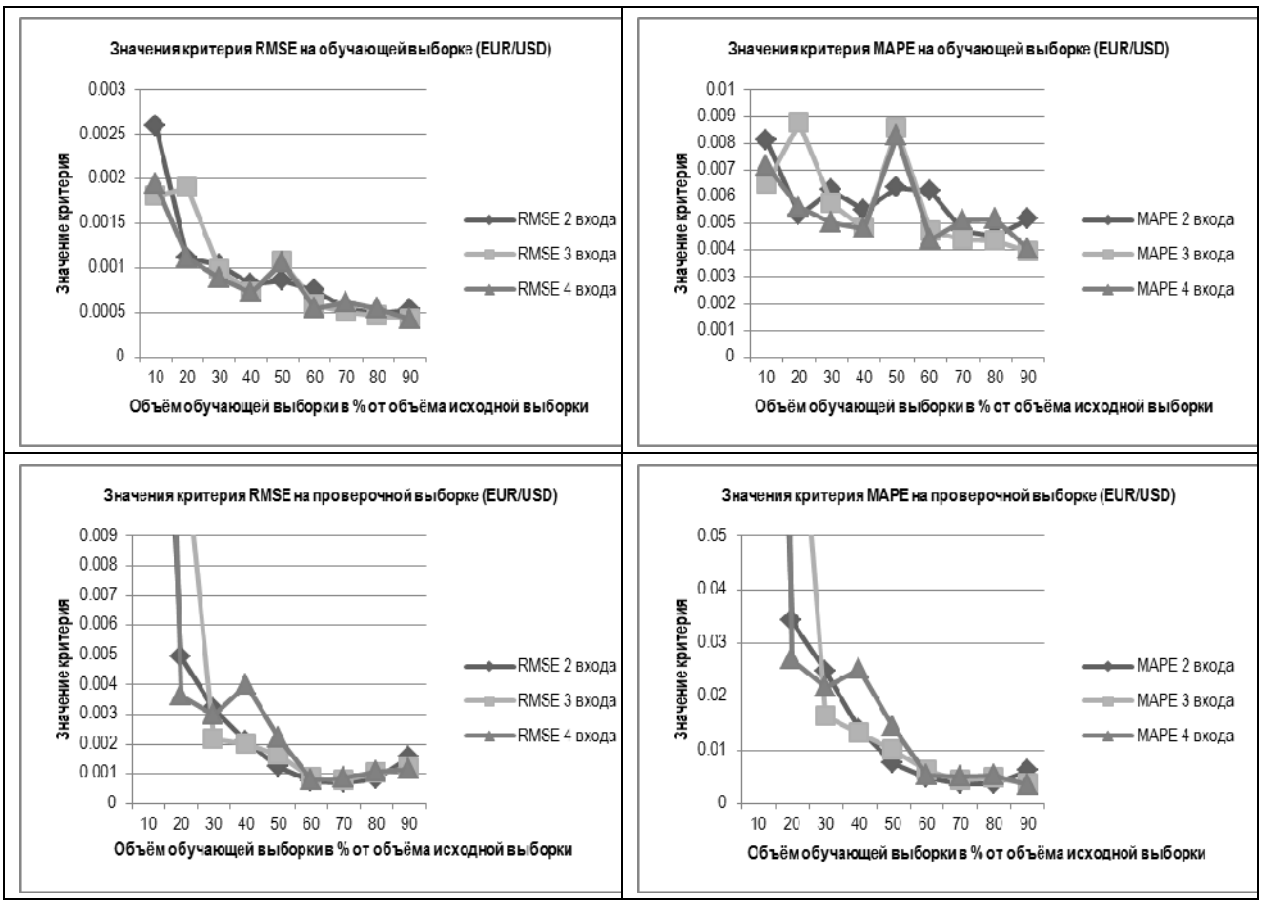


Таблица 7. Значения критериев для валютной пары USD/JPY

обуч:пров	Количество входов											
	2				3				4			
	RMSE обуч.	MAPE обуч.	RMSE пров.	MAPE пров.	RMSE обуч.	MAPE обуч.	RMSE пров.	MAPE пров.	RMSE обуч.	MAPE обуч.	RMSE пров.	MAPE пров.
10:90	0,10798	0,005251	0,45759	0,078326	0,11526	0,005468	1,1285	0,19462	0,15089	0,007299	1,8272	0,31217
20:80	0,08949	0,005902	0,32706	0,042931	0,094153	0,00644	0,11509	0,016595	0,11621	0,007547	0,96241	0,1484
30:70	0,071606	0,005814	0,052713	0,005479	0,099422	0,008136	0,14912	0,021071	0,078494	0,00636	0,16827	0,020172
40:60	0,067385	0,006201	0,42312	0,051221	0,077731	0,007054	0,65208	0,080762	0,061552	0,00561	0,72667	0,089607
50:50	0,051304	0,005456	0,063959	0,007649	0,050244	0,004964	0,14253	0,013533	0,049009	0,005052	0,056701	0,005995
60:40	0,052345	0,006029	0,06225	0,005453	0,045704	0,005369	0,051846	0,005141	0,04887	0,005757	0,068554	0,00615
70:30	0,04346	0,005542	0,054179	0,004602	0,041365	0,005181	0,054236	0,004717	0,050213	0,00674	0,085693	0,008241
80:20	0,04996	0,007199	0,069813	0,004987	0,044928	0,006461	0,069315	0,004746	0,049953	0,007028	0,059017	0,003899
90:10	0,042724	0,006117	0,089295	0,004181	0,039285	0,005608	0,08961	0,004133	0,12461	0,019378	0,53685	0,033844

Таблица 8. Графическое представление результатов экспериментов для валютной пары USD/JPY

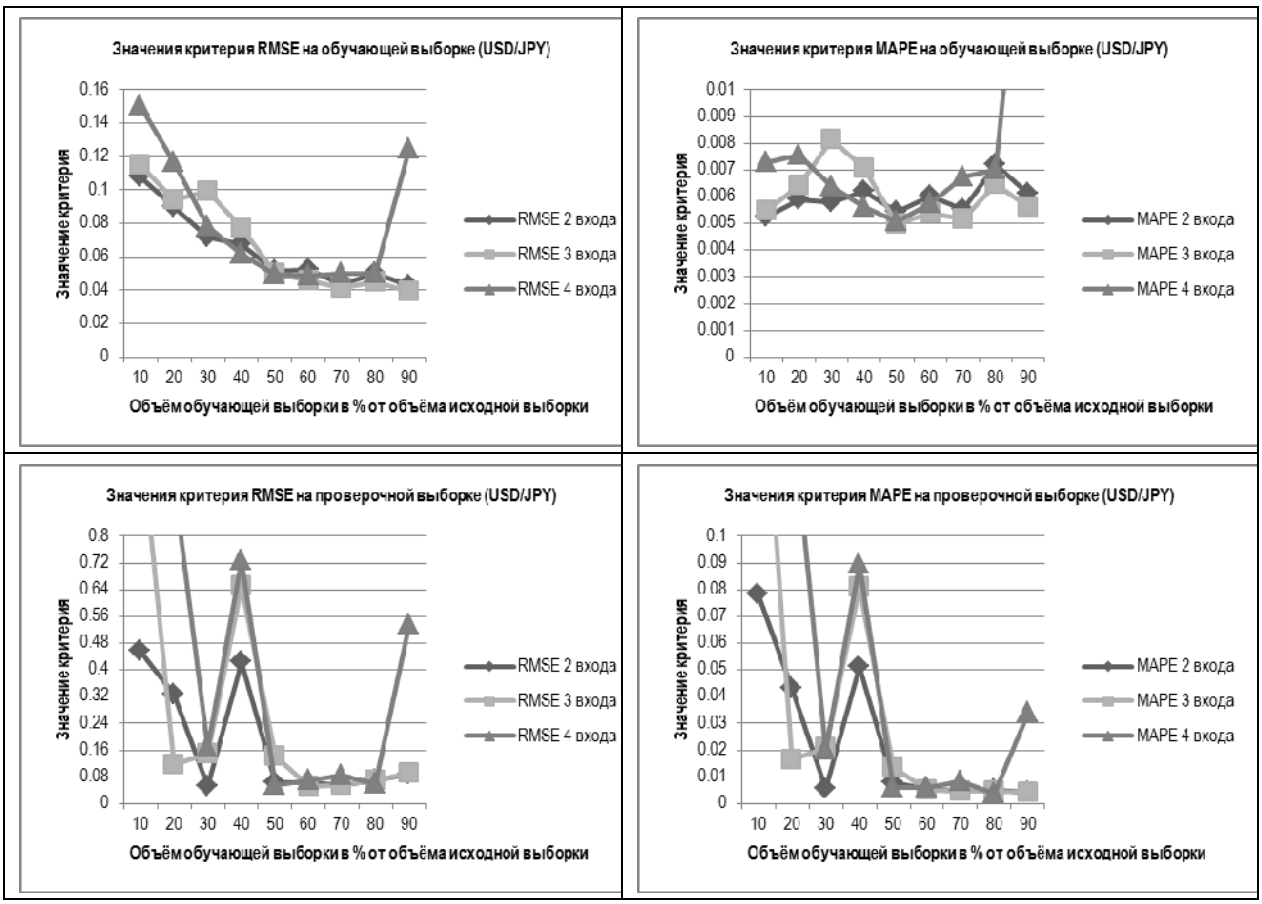
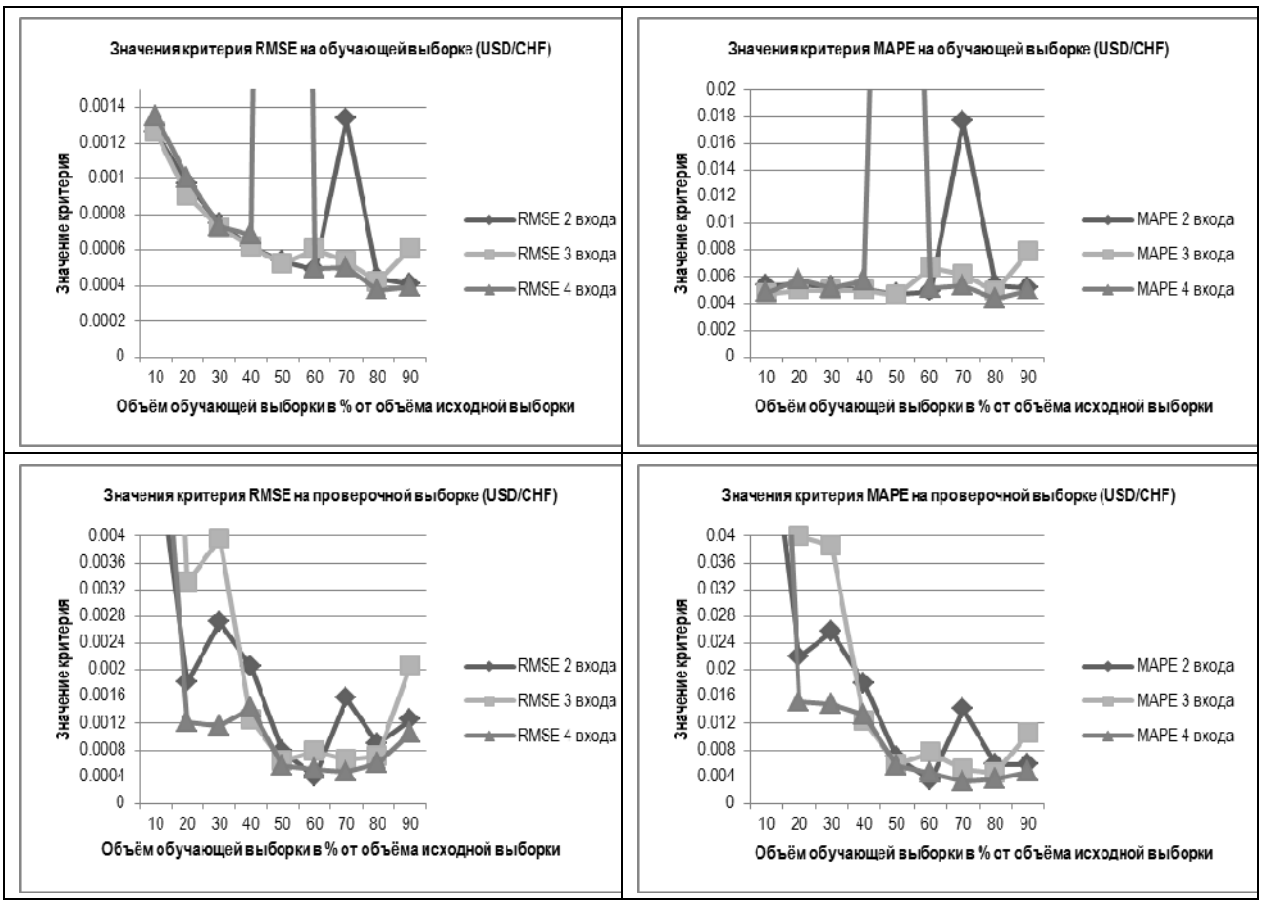


Таблица 9. Значения критериев для валютной пары USD/CHF

обуч:пров	Количество входов											
	2				3				4			
	RMSE обуч.	MAPE обуч.	RMSE пров.	MAPE пров.	RMSE обуч.	MAPE обуч.	RMSE пров.	MAPE пров.	RMSE обуч.	MAPE обуч.	RMSE пров.	MAPE пров.
10:90	0,00126	0,005325	0,00552	0,062745	0,001263	0,004744	0,01377	0,17941	0,00135	0,004777	0,007384	0,11671
20:80	0,000971	0,005424	0,00181	0,021933	0,000907	0,00501	0,003308	0,03995	0,001003	0,005767	0,001212	0,015209
30:70	0,000748	0,005233	0,002721	0,025682	0,000727	0,00496	0,003964	0,038628	0,000736	0,005187	0,001167	0,014883
40:60	0,000621	0,005054	0,002051	0,017899	0,000618	0,00504	0,00126	0,012279	0,000687	0,005728	0,001437	0,013268
50:50	0,000537	0,004729	0,000776	0,006943	0,000527	0,00464	0,000639	0,00588	0,014786	0,090375	0,000573	0,005672
60:40	0,000491	0,004889	0,000412	0,003576	0,000604	0,006677	0,000782	0,007651	0,000499	0,005168	0,000509	0,004593
70:30	0,001337	0,017609	0,001571	0,014083	0,000534	0,006209	0,000648	0,005141	0,000503	0,005384	0,000467	0,003291
80:20	0,000443	0,005329	0,000882	0,005893	0,000421	0,004925	0,000708	0,00454	0,000376	0,004403	0,000603	0,00368
90:10	0,000411	0,005218	0,001256	0,00584	0,00061	0,007912	0,002062	0,010478	0,000397	0,005001	0,001056	0,004748

Таблица 10. Графическое представление результатов экспериментов для валютной пары USD/CHF



Прежде чем переходить к анализу полученных результатов, следует описать факты, которые были получены в результате экспериментов. Во-первых, очевидно, что для обоих критериев на обучающих выборках происходит постепенное падение их значений. Во-вторых, значения критериев на проверочных выборках сначала падают, наступает «период относительной стабильности», а после – начинают расти. Как показывают эксперименты, период стабильности в среднем лежит в пределах соотношений обучающей и проверочной выборки 50:50 – 70:30 соответственно. В-третьих, для обоих критериев на обучающей и проверочной выборке при объёме обучающей выборки 10-20% от исходной выборки их значения заметно отличаются от значений при других объёмах обучающей выборки – они значительно больше. В-четвёртых, для критерия MAPE, при объёмах обучающих выборок больше 50%, для всех котировок на обучающих выборках порядок значений остаётся один и тот же и аналогичная ситуация наблюдается, если сравнивать между собой значения данного критерия для котировок на проверочных выборках, в то время, когда для RMSE для разных котировок, сравнивая его значения на соответствующих выборках и при соответствующих объёмах обучающих выборок он разный. В-пятых, оба критерия на одних и тех же выборках демонстрируют одинаковое поведение. В-шестых, отмечается незначительный рост качества работы сети с ростом числа входов.

Анализ результатов будет, фактически, объяснением выделенных фактов. Для объяснения всех последующих фактов необходимо, прежде всего, объяснить: почему при определённых соотношениях обучающей и проверочной выборок поведение и значения критериев значительно отличаются от поведения и значений при других соотношениях. Ответ лежит в следующем: с падением числа данных, участвующих в расчёте критерия, растёт вес каждой отдельной точки выборки этих данных. Таким образом, происходит некоторая переоценка значимости (именно при указанных соотношениях) отдельно взятой точки для обучения сети, или для оценки качества её работы. Это – объяснение второго и третьего факта. Первый факт, с учётом предыдущего вывода, говорит, что с ростом объёма обучающей выборки растёт эффективность функционирования сети на обоих выборках. Однако незначительный рост качества свидетельствует о существовании «периода насыщения», когда рост объёма обучающей выборки даёт прирост показателей качества, которыми можно пренебречь. Шестой факт является отчасти следствием ограничения на максимальное число итераций при обучении сети – сеть с большим числом входов требует большего времени на обучение,– а отчасти констатацией аналогичного факта существования «периода насыщения» для количества входов сети, так как отмечается незначительный прирост качества. Объяснение четвёртого факта уже приводилось в подразделе 1.2: RMSE является чувствительным к масштабу исследуемых величин, в то время, когда MAPE является практически нечувствительным. Пятый факт – соответствие поведений обоих рассматриваемых критериев на одних и тех же выборках – свидетельствует об адекватности выбора критерия обучения сети для конкретных данных и об отсутствии необходимости в его модификации. Главный вывод, с учётом всех вышеуказанных рассуждений и выводов предыдущих глав – сеть достаточно точно распознаёт поведение рынка и с приемлемой точностью угадывает котировки валют.

Выводы и перспективы дальнейших исследований

Прежде всего, необходимо отметить, что приложение сети к реальным котировкам валютных пар показало эффективность сети применительно к задаче прогноза значений валютных котировок на основании их предыстории. Данное заключение было сделано на основании полученных значений критериев качества, которые приведены в работе. Однако сразу же следует отметить, что необходимы дальнейшие исследования, направленные на усовершенствования нейросетевой архитектуры и методики её применения. Это следует из того, что лучшее значение критерия MAPE было равно где-то 0,003, в то время, когда практически полным распознаванием ситуации является его значение меньше чем 0,0001.

Это улучшение может быть получено путём предварительного анализа и преобразования данных, которые поступают на вход сети. Также значения параметра σ в функциях принадлежности системы ([2], [3]) были для каждого отдельного входа сети фиксированными. Однако возможно устранение этого недостатка путём дообучения сети путём оптимизации градиентным методом для каждого нечёткого правила сети этих параметров. Оптимизация должна происходить с целью максимизации значений итоговых весов правил на точках соответствующего подмножества обучающей выборки. Кроме проведения модификации среды использования сети необходимо также произвести сравнительный анализ результатов полученных в данной работе с результатами применения к вышеуказанной задаче прочих нейросетевых архитектур, в особенности, архитектур, реализующих нечёткий вывод Такаги-Сугено (TSK, радиальные базисные нейронные сети, сети на нео-фази нейронах и пр.). Всё это будет предметом дальнейших исследований. Ещё необходимо отметить, что выбор критериев RMSE и MAPE для анализа сети в данной работе был продиктован целью исследования того, насколько точно сеть распознаёт значение конкретной котировки. То есть, ошибкой считаются и значения большие и значения меньшие реального значения котировки. Однако необходимо учитывать тот факт, для чего производится анализ – с целью дальнейшей покупки либо продажи валюты. В этом случае понятие «ошибка» значительно меняет свой смысл.

В заключение следует ещё раз повториться, что, как показывают эксперименты, сеть при предложенных модификациях способна относительно точно угадывать котировки валют и вполне применима для практического использования.

Благодарности

Статья частично финансирована из проекта ITHEA XXI Института Информационных теорий и Приложений FOI ITHEA и консорциума FOI Bulgaria (www.ithea.org, www.foibg.com)

Литература

1. Дневные котировки валют за период с 25.03.2009 по 24.03.2010 взяты со странички http://www.finam.ru/analysis/export/default.asp
2. Зайченко Ю.П. Основи проектування інтелектуальних систем. Навчальний посібник. – К.: Видавничий Дім «Слово», 2004. – С. 352
3. Зайченко Ю.П. Нечёткие модели и методы в интеллектуальных системах. Учебное пособие для студентов высших учебных заведений. – К.: «Издательский Дом «Слово»», 2008. – С. 344
4 Зайченко Ю.П. Дослідження операцій. Підручник. Сьоме видання, перероблене та доповнене. – К.: Видавничий Дім «Слово», 2006. – С. 816

Информация про автора

Мурга Николай Алексеевич *– аспирант Национального технического университета Украины «КПИ», адрес электронной почты: murga.nicholas@gmail.com*

Основные сферы научных исследований автора: *применение теории нечёткой логики и теории детерминированного хаоса к анализу финансовых рынков.*