УДК 004.032.26

Ю.Л. ИВАСЬКИВ, О.Л. ЛЕЩИНСКИЙ, В.В. ЛЕВЧЕНКО

ОЦЕНКА КАЧЕСТВА ОБУЧАЮЩИХ МНОЖЕСТВ ДЛЯ НЕЙРОННЫХ СЕТЕЙ В ЗАДАЧАХ СЖАТИЯ ДАННЫХ БЕЗ ПОТЕРЬ

Abstract: *The experimental estimation of quality of the training sets, which are generated of data of various types for neural networks, by criteria of repeatability and discrepancy is given. Recommendations at the choice of parameters of training set for neural models in problems of lossless data compression are formulated.*
Key words: *neural model, statistic modeling, training set.*

Анотація: *Дано експериментальну оцінку якості навчальних множин, сформованих з даних різних типів для нейронних мереж за критеріями повторюваності та суперечності. Сформульовані рекомендації з вибору параметрів навчальної множини для нейромережевої моделі в задачах стиснення даних без втрат.*
Ключові слова: *нейромережева модель, статистичне моделювання, навчальна множина.*

Аннотация: *Дана экспериментальная оценка качества обучающих множеств, сформированных из данных различных типов для нейронных сетей по критериям повторяемости и противоречивости. Сформулированы рекомендации по выбору параметров обучающего множества для нейросетевой модели в задачах сжатия данных без потерь.*
Ключевые слова: *нейросетевая модель, статистическое моделирование, обучающее множество.*

1. Введение

Благодаря наличию ряда уникальных особенностей, связанных с возможностью адаптации к входным данным, высокой потенциальной параллельностью и однородностью выполняемых операций, технологической простотой физической реализации основных архитектурных элементов, применение нейронных сетей (НС) рассматривается как альтернатива многим классическим методам, используемым при адаптивной обработке данных [1–3]. При адаптивной обработке одной из важных задач является статистическое моделирование источников данных в процессе осуществления сжатия без потерь энтропийными методами [4].

Известны исследования, направленные на решение задачи статистического моделирования, основанные на использовании нейросетевой модели, представляющей собой многослойную НС обратного распространения ошибки, которая обучается предсказыванию значений, генерируемых моделируемым источником данных [5, 6]. Обучающее множество для моделей такого типа формируется в соответствии с так называемым методом “погружения” данных, генерируемых источником, в лаговое пространство. Подобный метод обычно используется для прогнозирования временных рядов [7]. В соответствии с этим методом, для каждого вектора входных сигналов НС, представляющего собой N последовательных значений ряда, в качестве выходного значения рассматривается следующее, $(N+1)$-е значение. Величину N в задачах прогнозирования временных рядов называют глубиной погружения, а в задачах сжатия данных без потерь – длиной контекста [8].

В задачах прогнозирования временных рядов существует оптимальная глубина погружения для обучающего множества [9]. Соответственно для нейросетевых моделей возникает необходимость в определении оптимальной длины контекста.

2. Постановка задачи

В задачах сжатия данных длина обучающего контекста зависит от исходных данных и, в частности, от типа источника (данные текстовые, графические и т.п.), а также от его свойств (в текстах – от

выбора языка, в графических данных – от цветности, вида изображения и т.д.). В настоящей работе предлагается метод выбора контекста для нейросетевой модели, характеризуемого оптимальной длиной и отличающегося учетом типа данных, генерируемых источником. Определение оптимальной длины контекста для нейросетевой модели связывается с применением экспериментальных методов анализа обучающих множеств, основанных на использовании специальных алгоритмов и реализации их на основе имеющихся компьютерных средств, и предполагает анализ качества обучающих множеств с использованием критериев повторяемости и противоречивости [10].

3. Основные результаты

Выбор глубины погружения при решении задачи прогнозирования временных рядов производится на основе предварительной оценки свойств обучающего множества НС, сформированного из значений прогнозируемого ряда [10]. С целью выбора глубины погружения введены два критерия, характеризующие обучающее множество: повторяемость и противоречивость, а также сформулированы правила интерпретации значений этих критериев. Одной из особенностей определения численных значений введенных критериев является их ориентация на обучающие множества, составленные из категориальных значений [7]. В задачах прогнозирования временных рядов, состоящих в общем случае из номинальных значений, такие значения должны быть предварительно преобразованы из номинальных в категориальные. В задачах сжатия данные представляются в цифровом виде и их заведомо можно рассматривать как категориальные. Поэтому необходимость преобразования исходных данных в категориальные отсутствует. В результате уменьшается трудоемкость оценки численных значений критериев.

Для выполнения экспериментального анализа были выбраны несколько файлов из популярного набора Calgary Corpus Test, доступного в сети Internet и предназначенного для оценки эффективности программных методов сжатия данных без потерь. Параметры использованных файлов приведены в табл. 1.

Таблица 1. Структура набора Calgary Corpus Test

Название файла	Тип данных	Размер, байт
geo	Бинарные данные различной природы	102400
obj1		21504
obj2		246814
lena	Растровые изображения	211624
peppers		212340
tulips		311876
progc	Тексты программ на различных языках	39611
progl		71876
paper1	Тексты на английском языке	53161
paper2		82199
paper3		46526
paper4		13286
paper5		11954
paper6		38105

С целью экспериментальной оценки качества обучающих множеств в среде CBuilder 6.0

создана специальная программа (рис. 1), отличающаяся возможностью обработки цифровых данных произвольного типа. Эта программа использована для определения значений повторяемости и противоречивости для каждого из файлов в табл. 1 с учетом объема обучающего множества и длины контекста.

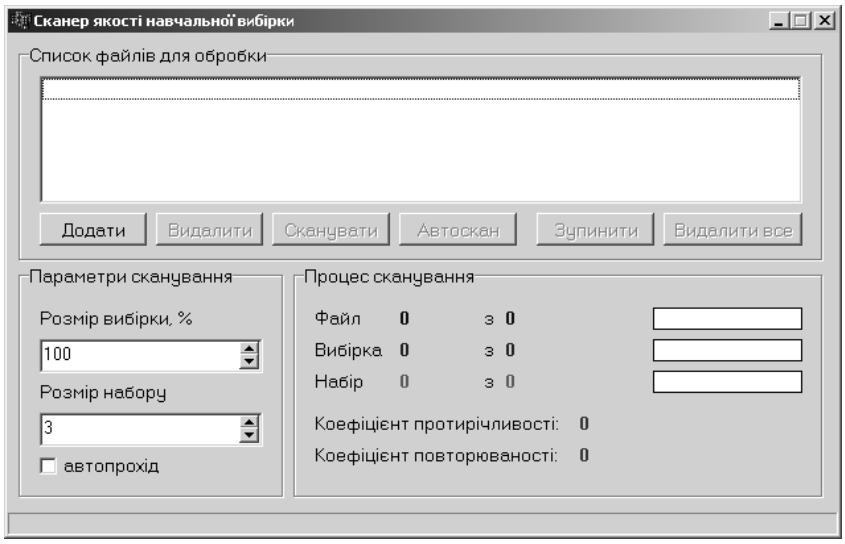


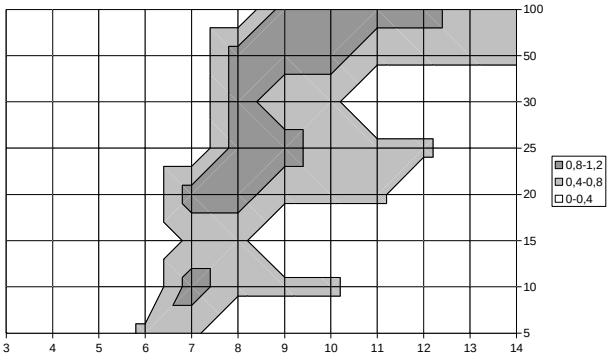
Рис. 1. Интерфейс программы

В результате оценки качества получены значения повторяемости и противоречивости, для визуализации которых была использована следующая функция качества обучающего множества [6]:

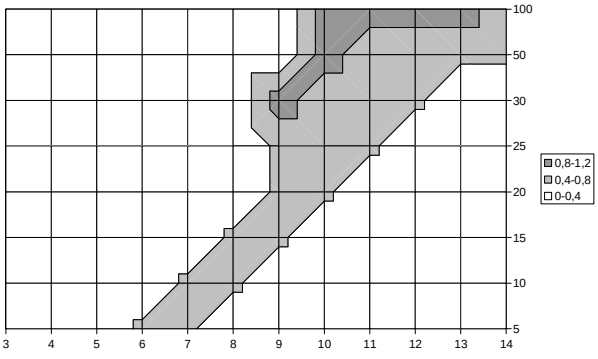
$$Q(\rho,\delta)=\begin{cases}1{,}0 & 0{,}7<\rho\le 1,\ \delta<0{,}2;\\ 0{,}5, & 0{,}4\le\rho<0{,}7,\ \delta<0{,}2;\\ 0, & \rho<0{,}4,\end{cases} \qquad (1)$$

где ρ – повторяемость, δ – противоречивость.

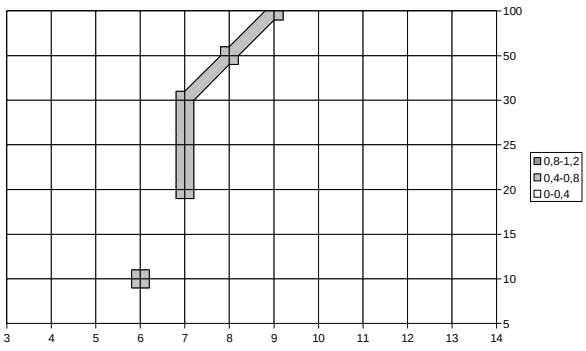
Значения 1,0 и 0,5 функции (1) характеризуются как идеальное и оптимальное сочетание критериев соответственно, а значение 0 свидетельствует о недостаточном для эффективного прогнозирования качестве обучающего множества [6]. Результаты оценки качества приведены на рис. 2.



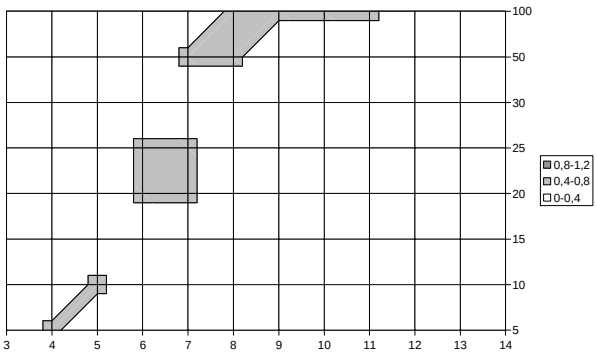
а)



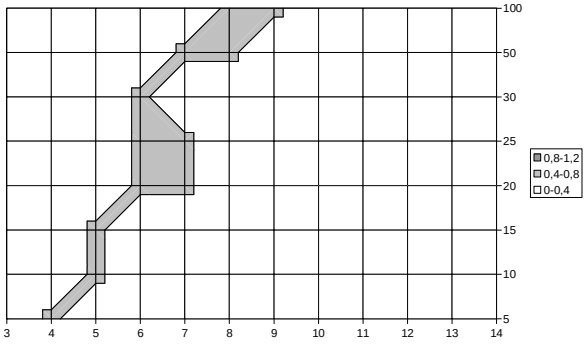
б)


0,8-1,2
0,4-0,8
0-0,4

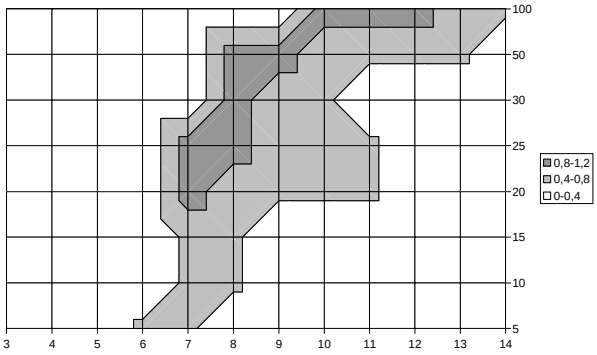
в)


0,8-1,2
0,4-0,8
0-0,4

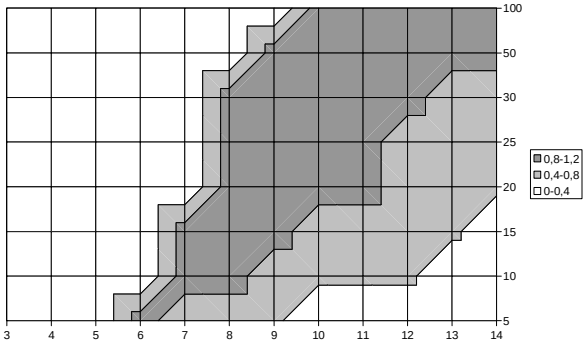
г)


0,8-1,2
0,4-0,8
0-0,4

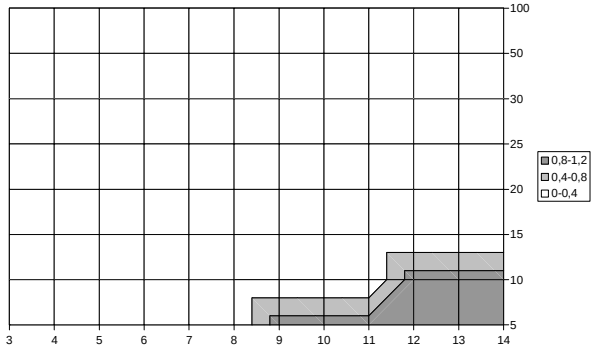
д)


0,8-1,2
0,4-0,8
0-0,4

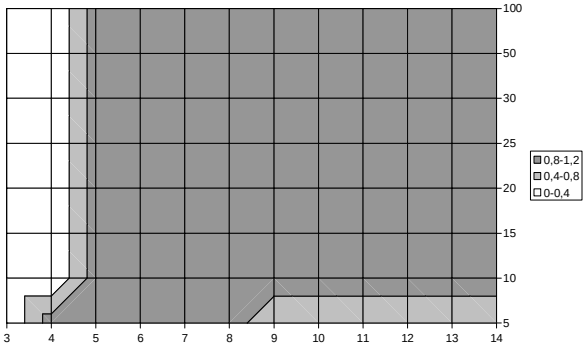
е)


0,8-1,2
0,4-0,8
0-0,4

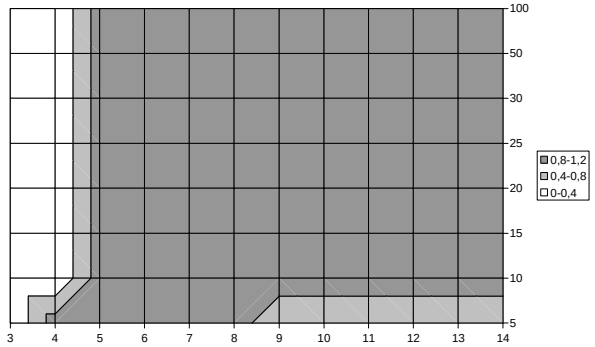
ж)


0,8-1,2
0,4-0,8
0-0,4

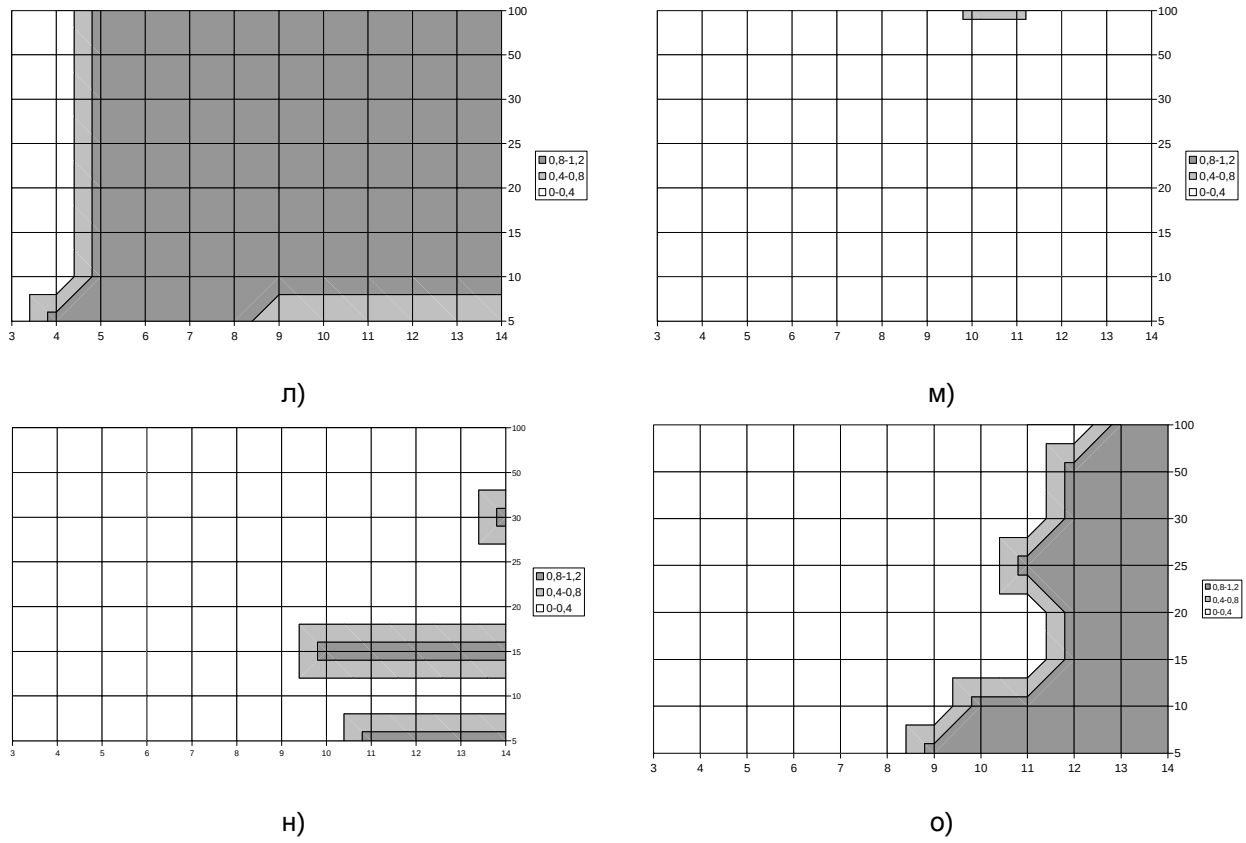
з)


0,8-1,2
0,4-0,8
0-0,4

и)


0,8-1,2
0,4-0,8
0-0,4

к)



л) м)

н) о)

Рис. 2. Результаты оценки качества обучающих множеств: а) paper1; б) paper2; в) paper3; г) paper4; д) paper5; е) paper6; ж) progc; з) progl; и) lena; к) peppers; л) tulips; м) geo; н) obj1; о) obj2

На рис. 2 (а – о) по горизонтали представлена длина контекста в байтах, а по вертикали – размер обучающего множества в процентах от размера файла. Показано, что качество обучающего множества, прежде всего, зависит от типа обрабатываемых данных и характеризуется следующими особенностями:

1) в текстовых файлах (рис. 2 а – з) наибольшее значение функция качества принимает в области средних значений длины контекста;

2) в текстовых и графических файлах существует зависимость минимальной длины контекста, при которой достигается наибольшее значение функции качества, от размера обучающего множества;

3) в бинарных файлах (рис. 2 м, н, о) общие закономерности поведения функции качества отсутствуют.

Полученные оценки могут быть использованы в качестве начальных при построении модели, представляющей собой многослойную НС, обучаемую в соответствии с алгоритмом обратного распространения ошибки [1]. Поскольку такая модель с N входами фактически учитывает как контекст длины N, так и все остальные контексты меньшей длины: $N-1$, $N-2,...,1$, то в процессе ее использования для некоторых отдельных наборов данных может быть получена уточненная оценка длины контекста. Однако для ее получения требуется разработка

специальных методов, учитывающих изменения оптимальной длины контекста в зависимости от его расположения в сжимаемом файле.

4. Выводы

В результате разработки и исследования метода выбора оптимальной длины контекста для нейросетевой модели было установлено, что оптимальная длина контекста зависит, прежде всего, от природы сжимаемых данных и характеризуется особым диапазоном значений, зависящим от типа данных, а также их свойств.

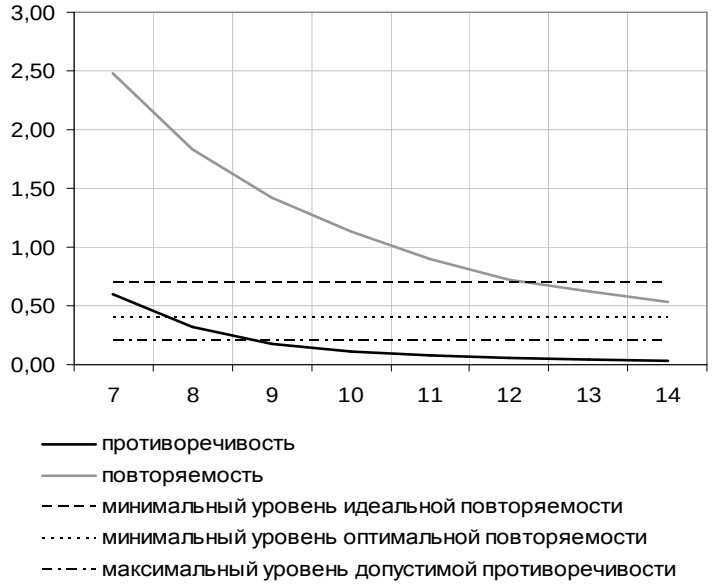


Рис. 3. Поведение повторяемости и противоречивости в текстовых данных

В текстовых данных появление участков оптимального сочетания критериев качества (рис. 3) связывается с одновременным снижением противоречивости и повторяемости обучающего множества при увеличении длины контекста. Для текстовых данных в байтовой кодировке оптимальная длина контекста составляет 7…11 байт и практически не зависит от природы текста за некоторыми исключениями, отражающими его структуру, семантику и т. п.

Для графических данных оптимальная длина контекста колеблется в широком диапазоне, что требует применения более тонких методов оценки при построении и функционировании нейросетевой модели, позволяющих сузить этот диапазон или динамически изменять его в процессе сжатия данных.

СПИСОК ЛИТЕРАТУРЫ

1. Нейроматематика: Учеб. пособие для вузов / А.Д. Агеев, А.В. Балухто и др.; Общая ред. А.И. Галушкина. – Кн. 6. – М.: ИПРЖР, 2002. – 448 с.
2. Ланнэ А.А. Нейронные цепи и синтез нелинейных операторов обработки сигналов // Труды 4-й Международной научно-технической конференции студентов, аспирантов и молодых специалистов стран СНГ. – Алматы, Казахстан, 2002. – С. 34–42.
3. Хрящев В.В. Нейросетевое восстановление амплитуды дискретного сигнала по его фазовому спектру / В.В. Хрящев, Е.А. Соколенко, А.Л. Приоров // 5-я Междунар. конф. “Цифровая обработка сигналов и ее применение”. – 2003. – № 2. – С. 622–624.
4. Jorma Rissanen, Glen G. Langdon Universal modelling and coding // IEEE transaction on information theory. – 1981. – Vol. 27. – P. 12–33.
5. Schmidhuber J. Sequential neural text compression // IEEE transaction on Neural Networks. – 1996. – Vol. 7. – P.

142–146.
6. Jiang J. Image compression with neural networks // A survey. Signal Processing: Image Commun. – 1999. – Vol. 14. – P. 737–760.
7. Ежов А.А., Шумский С.А. Нейрокомпьютинг и его применение в экономике и бизнесе. Серия “Учебники экономико-аналитического института МИФИ” / Под ред. проф. В.В. Харитонова. – М.: МИФИ, 1998. – 244 с.
8. Методы сжатия данных. Устройство архиваторов, сжатие изображений и видео / Д. Ватолин, А. Ратушняк, М. Смирнов и др. – М.: ДИАЛОГ-МИФИ, 2002. – 384 с.
9. Малинецкий Г.Г., Потапов А.Б. Современные проблемы нелинейной динамики. – М.: Эдиториал УРСС, 2000. – 336 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. – №. 1. – С. 25–28.

Стаття надійшла до редакції 14.11.2007