Московский государственный университет им. М.В.Ломоносова

Факультет вычислительной математики и кибернетики

На правах рукописи

Гужва Александр Георгиевич

Разработка методологии и программного комплекса для определения существенности входных признаков при нейросетевом анализе данных

05.13.18 – математическое моделирование, численные методы и комплексы программ

АВТОРЕФЕРАТ
диссертации на соискание ученой степени
кандидата физико-математических наук

Москва, 2011

Работа выполнена в Московском Государственном Университете им. М.В. Ломоносова на кафедре атомной физики, физики плазмы и микроэлектроники Физического факультета.

Научный руководитель – доктор физико-математических наук,
профессор Персианцев Игорь Георгиевич

Официальные оппоненты – доктор физико-математических наук,
профессор Дмитриев Владимир Иванович

доктор физико-математических наук,
профессор, Редько Владимир Георгиевич

Ведущая организация – Учреждение Российской академии наук
Вычислительный центр им. А. А. Дородницына РАН

Защита состоится «02» марта 2011 г. в 15 ч. 30 мин. на заседании Диссертационного Совета Д.501.001.43 при Московском Государственном Университете им. М.В. Ломоносова по адресу: 119991, ГСП-1, Москва, Ленинские Горы, МГУ, 2-ой учебный корпус, факультет ВМК, аудитория 685.

С диссертацией можно ознакомиться в библиотеке факультета ВМК Московского Государственного Университета им М.В. Ломоносова.

Автореферат разослан «__» _____ ____ г.

Ученый секретарь
диссертационного совета
доктор физико-математических наук,
профессор Захаров Е. В.

Актуальность темы

Современные задачи комплексного исследования различных научно-технических проблем могут оперировать многомерными *наборами данных* большого объема, включающими множественные измерения большого[1] количества различных характеристик (*входных признаков*) исследуемого объекта. В случаях невозможности построения адекватной содержательной модели объекта традиционными методами, в качестве математической модели могут применяться *искусственные нейронные сети (ИНС);* соответствующие математические модели называются *нейросетевыми моделями*. ИНС также применяются для решения некорректно поставленных задач.

К достоинствам ИНС относятся: возможность обучаться на примерах, устойчивость к шумам, возможность работы с неполными и противоречивыми данными, параллельность архитектуры. Поэтому ИНС часто привлекаются для решения плохо формализуемых задач.

При решении практических задач, в рамках которых проводится построение нейросетевых моделей, необходимо учитывать ряд ограничений, особенно существенных при работе с наборами данных, содержащих большое число входных признаков:

1. Проблема интерпретации входных признаков состоит в усложнении определения вклада различных входных признаков в построенную нейросетевую модель при повышении числа рассматриваемых признаков.
2. Проблема снижения качества нейросетевой модели заключается в ухудшении качества получаемой нейросетевой модели при увеличении числа рассматриваемых входных признаков.
3. Проблема множественных запусков алгоритма обучения связана с ситуациями попадания в локальный минимум итерационной процедуры

[1] Большого – не только в смысле большого числа, но и в том смысле, что реально определять поведение объекта может лишь небольшое количество измеряемых величин из полного множества измеряемых.

построения нейросетевой модели. Важность этой проблемы возрастает с увеличением числа имеющихся входных признаков.

Указанные трудности возможно преодолеть путем сокращения числа входных признаков с помощью выявления *наиболее существенных входных признаков* и последующего исключения малосущественных.

Таким образом, определение наиболее существенных входных признаков представляет собой весьма *актуальную самостоятельную проблему*.

Цель работы. Основные задачи

Целью диссертационной работы являлась разработка новой *методики построения ИНС* (многослойных персептронов) для решения задач нелинейной регрессии, понимаемых в смысле *моделирования* поведения некоторой неизвестной зависимости, на основе *отбора* существенных входных признаков. Круг рассматриваемых задач ограничен задачами нелинейной регрессии для многомерных зависимостей *одной скалярной* действительной величины. Методика применяется для решения плохо формализуемых задач, в т.ч. в случае невозможности их решения традиционными методами.

Под *методами анализа существенности входных признаков* (*методы АСВП*) будут пониматься различные алгоритмы, которые можно использовать для выявления существенных входных признаков. В результате применения методики строится нейросетевая модель, решающая задачу регрессии с использованием наиболее существенных входных признаков, выявляемых с помощью комбинации методов АСВП.

Методика применяется для достижения следующих целей:

1. Повышение точности решения задачи регрессии (с помощью ИНС).
2. Сокращение числа используемых входных признаков в рассматриваемом наборе данных.

Для достижения цели ставились следующие **основные задачи**:

1. Построить методику в виде алгоритма, содержащего порядок использования и условия применимости тех или иных методов АСВП.

a. Исследовать возможность и целесообразность включения в методику различных известных методов АСВП.
b. Сформулировать критерии для включения новых методов АСВП в методику.

2. Всесторонне исследовать построенную методику, применяя её для решения модельных задач и ряда прикладных задач, в том числе *эталонных* задач, опубликованных в Интернете и обычно используемых для тестирования новых методов анализа данных.
3. Создать программное обеспечение, реализующее все необходимые алгоритмы и методы.

В работе *не* ставилась цель сбора всевозможных методов АСВП. Также, задача всестороннего исследования методов АСВП не являлась основной.

Основные положения, выносимые на защиту

1. Разработанная методика построения ИНС (многослойных персептронов) для решения задач нелинейной регрессии на основе отбора существенных входных признаков, представленная в виде *алгоритма*.
2. Результаты решения 5 модельных задач, 40 эталонных задач, а также 3 прикладных задач с использованием разработанной методики.
3. Разработанный оригинальный алгоритм параллельного обучения большого числа многослойных персептронов с одним скрытым слоем. Алгоритм был реализован с использованием технологии CUDA [C1].
4. Созданный единый программный комплекс, реализующий все использованные при разработке методики методы АСВП, с возможностью производить вычисления на нескольких компьютерах в локальной сети под управлением ОС MS Windows с управлением вычислениями из единого центра.

Научная новизна

1. Разработанная методика построения ИНС для решения задач нелинейной регрессии на основе отбора существенных входных признаков является оригинальной и представляет собой усовершенствование традиционного метода построения нейронных сетей (многослойных персептронов) для решения задач нелинейной регрессии.
2. Проведена апробация разработанной методики на основе большого числа эталонных наборов данных. Получены количественные оценки эффективности применения разработанной методики для различных типов задач.
3. Разработанный и реализованный алгоритм параллельного обучения персептронов с одним скрытым слоем, показывающий существенное ускорение для графических процессоров компании NVIDIA по сравнению с реализациями алгоритма обучения методом обратного распространения ошибки для современных центральных процессоров, является оригинальным.
4. Получено решение следующих прикладных задач (задач реального мира) с использованием разработанной методики:
    a. Задача прогнозирования значения геомагнитного Dst-индекса на основании значений параметров солнечного ветра.
    b. Задача нелинейной регрессии из области электроразведки (магнитотеллурики) по восстановлению распределения электропроводности участка земной коры на основании измеренных на поверхности земли характеристик ЭМ полей (компонент тензора импеданса).
    c. Задача оценки токсичности химических соединений на основе дескрипторов молекул этих соединений.

Практическая значимость

1. Предложенная *методика построения ИНС* может быть использована при решении широкого круга задач нелинейной регрессии и прогнозирования. Применение разработанной методики позволяет в среднем снизить погрешность решения, а также сделать выводы о существенности различных входных признаков при построении модели.
2. Разработанный оригинальный алгоритм параллельного обучения персептронов с одним скрытым слоем, а также реализация алгоритма, использующая технологию CUDA, могут быть с высокой вычислительной эффективностью использованы для построения нейросетевых моделей. Разработанный алгоритм открывает перспективы для решения более масштабных задач на персональных компьютерах за меньшее время.
3. Разработанный в ходе выполнения диссертационной работы программный комплекс внедрен в Российском государственном геологоразведочном университете для проведения расчётов и обучения студентов.

Решения прикладных задач, полученные при разработке методики, а также информация о выделенных наборах существенных признаков, могут быть использованы в соответствующих предметных областях.

Апробация работы

Результаты, полученные в диссертационной работе, представлены в устных и стендовых докладах на 8 Всероссийских и международных конференциях:

1. VIII Всероссийская научно-техническая конференция «Нейроинформатика-2006», г. Москва, МИФИ, 24-27 января 2006 г.
2. IX Всероссийская научно-техническая конференция «Нейроинформатика-2007», г. Москва, МИФИ, 23-26 января 2007 г.

3. 8-я Международная конференция "Распознавание образов и анализ изображений: новые информационные технологии" (РОАИ-8-2007), г. Йошкар-Ола, 8-12 октября 2007 г.
4. X Всероссийская научно-техническая конференция «Нейроинформатика-2008», г. Москва, МИФИ, 22-25 января 2008 г.
5. 9-я Международная конференция " Распознавание образов и анализ изображений: новые информационные технологии " (РОАИ-9-2008), г. Нижний Новгород, 14-20 сентября 2008 г.
6. XI Всероссийская научно-техническая конференция «Нейроинформатика-2009», г. Москва, МИФИ, 27-30 января 2009 г.
7. 19th International Conference on Artificial Neural Networks (ICANN-2009), 14-17 September, Limassol, Cyprus.
8. XII Всероссийская научно-техническая конференция «Нейроинформатика-2010», г. Москва, МИФИ, 25-29 января 2010 г.

Публикации

Основные результаты диссертации опубликованы в 15 статьях, в том числе в 3 журнальных публикациях, материалах 5 Всероссийских и 3 международных конференций. 6 статей размещено в изданиях, рекомендованных ВАК.

Основное содержание работы

Во **введении** обосновывается актуальность работы, сформулированы цели работы, поставленные задачи, научная новизна и основные положения, выносимые на защиту.

В **главе 1** приведен обзор литературы из рассматриваемой области понижения размерности данных, разъясняются используемые термины, приведены ссылки на основополагающие принципы, наиболее часто используемые методы и алгоритмы, существующие подходы к рассмотрению задач, а также описание ряда актуальных проблем.

Приведена формальная постановка задачи, решаемой с помощью методики.

Имеется некоторая неизвестная зависимость $\boldsymbol{y} = \boldsymbol{y}^*(\boldsymbol{x})$, $\boldsymbol{y}^*: \mathbb{R}^M \to \mathbb{R}^P$, значения которой известны только на некотором конечном обучающем множестве векторов $X = \{\boldsymbol{x^1}, \dots, \boldsymbol{x^n}\}$, $\boldsymbol{x^i} \in \mathbb{R}^M$. Задано разбиение X на тренировочный набор данных X^{Trn} и тестовый набор данных X^{Tst}, $X = X^{Trn} \oplus X^{Tst}$. Задан вид семейства параметрических функций M переменных $\boldsymbol{y} = \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{\alpha})$, где $\boldsymbol{\alpha}$ – вектор параметров семейства функций, используемых для аппроксимации $\boldsymbol{y}^*$ по ее известным значениям. Задан вид функционала ошибки $E(\boldsymbol{\alpha}, Z) = \sum_{x \in Z} |y^*(\boldsymbol{x}) - \boldsymbol{f}(\boldsymbol{x}, \alpha)|^2$.

Требуется подобрать функцию $\boldsymbol{f}(\boldsymbol{x}, \boldsymbol{\alpha}^*)$, или – эквивалентно – решение $\boldsymbol{\alpha}^*$ системы нелинейных уравнений $\boldsymbol{y}^*(\boldsymbol{x^i}) = \boldsymbol{f}(\boldsymbol{x^i}, \boldsymbol{\alpha})$, $\boldsymbol{x^i} \in X^{Trn}$, при которых достигался бы минимум функционала ошибки на тестовом наборе данных $E(\boldsymbol{\alpha}^*, X^{Tst})$.

Традиционно решение поставленной задачи осуществляется с помощью решения системы методом наименьших квадратов путем минимизации функционала ошибки $E(\boldsymbol{\alpha}, X^{Trn})$ с последующей проверкой интерполяционных свойств найденной функции $\boldsymbol{f}(\boldsymbol{x}, \boldsymbol{\alpha}^*)$ на тестовом наборе данных.

Многослойный персептрон задает вид используемого семейства параметрических функций $\boldsymbol{f}(\boldsymbol{x}, \boldsymbol{\alpha})$. Часто используемый *многослойный персептрон с одним скрытым слоем* задает следующий вид $\boldsymbol{f}(\boldsymbol{x}, \boldsymbol{\alpha})$ (в компонентах), где $\boldsymbol{\alpha} = (\varphi, \psi, N, U, V)$:

$$f_p(\boldsymbol{x}, \boldsymbol{\alpha}) = f_p(\boldsymbol{x}, \varphi, \psi, N, U, V) = \psi\left(v_{p,0} + \sum_{i=1}^{N} v_{p,i} * \varphi\left(u_{i,0} + \sum_{j=1}^{M} u_{i,j} * x_j\right)\right),$$

где $p = 1, \dots, P$, $\varphi(\xi) \in \mathrm{C}_1$ и $\psi(\xi) \in \mathrm{C}_1$ – некоторые ограниченные монотонно возрастающие функции, N – натуральное число, характеризующее сложность данной аппроксимирующей конструкции, $U = \{u_{ij}\}$ и $V = \{v_{ij}\}$ – матрицы весовых коэффициентов. Многослойного персептрона с одним скрытым слоем достаточно для равномерной аппроксимации $\boldsymbol{y}^*$ при некотором выборе U, V и N [С2]. Размерность вектора $\boldsymbol{\alpha}$, которая зависит от числа

элементов матриц U, V, составляет $N * (M + 1) + P * (N + 1) \approx N * (P + M)$. За счет уменьшения числа входных признаков возможно ускорить процесс построения нейросетевой модели $\boldsymbol{f}(\boldsymbol{x}, \boldsymbol{\alpha})$.

Предложенная методика является усовершенствованием традиционного способа решения поставленной задачи в случае использования многослойных персептронов в качестве $\boldsymbol{f}(\boldsymbol{x}, \boldsymbol{\alpha})$. Во-первых, вместо $\boldsymbol{f}(\boldsymbol{x}, \boldsymbol{\alpha}) = \boldsymbol{f}(x_1, \dots, x_M, \boldsymbol{\alpha})$ могут рассматриваться функции $J \leq M$ числа переменных $\boldsymbol{f}\left(x_{k_1}, \dots, x_{k_J}, \boldsymbol{\alpha}\right)$, где $k_i = 1, \dots, M$, $k_i \neq k_j$ для $i \neq j$, получаемые путем сужения области допустимых значений $\boldsymbol{\alpha}$, а именно – фиксации некоторых компонент вектора $\boldsymbol{\alpha}$. Во-вторых, вводится дополнительная итерационная процедура поиска оптимальных значений J и $k_1, \dots, k_J$, на каждой итерации которой для найденных субоптимальных J^* и $k_1^*, \dots, k_J^*$ ищется промежуточное решение задачи $f^*\left(x_{k_1}^*, \dots, x_{k_J}^*, \alpha\right)$ с помощью традиционного способа. Под решением задачи будет пониматься наилучшее найденное промежуточное решение задачи на различных итерациях процедуры. За счет использования меньшего числа $J \leq M$ входных признаков по сравнению с исходным числом M, а также сужения области допустимых значений $\boldsymbol{\alpha}$, возможно получение лучшего решения задачи за счет улучшения обобщающих свойств. Поиск коэффициентов J^* и $k_1^*, \dots, k_J^*$ осуществляется с помощью различных методов АСВП.

В разделе 1.5 рассмотрено 13 известных алгоритмов анализа данных, с точки зрения достоинств и недостатков (в т.ч. вычислительных и временных), а также особенностей их использования в качестве *методов анализа существенности входных признаков (методы АСВП)*. Предложен способ сравнения различных методов АСВП.

Глава 2 посвящена изучению модельных и эталонных задач, на основании которого была построена методика. Преследовались следующие цели:

1. Определить, какие из имеющихся методов АСВП могут быть использованы для построения методики в ее «первичном» виде.
2. Выявить в практических исследованиях особенности различных методов АСВП, связанные с совместным использованием методов.

В **разделе 2.1** анализируются результаты решения четырех модельных задач построения регрессии, в которых исследовался ряд зависимостей, соответствующих некоторым типичным ситуациям, возникающим при анализе данных.

В **разделе 2.2** приведены подробные результаты решения задачи Фридмана – модельной задачи, используемой для тестирования различных алгоритмов анализа данных. Вид использованной формулы следующий:

$$Y = 10\sin(\pi x_1 x_2) + 20(x_3 - 0.5)^2 + 10x_4 + 5x_5 + E$$

Здесь E представляет собой гауссовый шум с нормальным распределением, с нулевым средним и единичной дисперсией.

Для изучения этой задачи использовался комплект из 80 синтетических наборов данных, взятых из базы данных WEKA [С3]. Наборы данных различались по следующим характеристикам:

1. Число входных признаков, не связанных с признаками $x_1 \dots x_5$, и содержащих случайно сгенерированные значения.
2. Число примеров.
3. Степень взаимосвязи входных признаков.

При исследовании задачи получены убедительные свидетельства целесообразности объединения ряда методов АСВП в некоторую последовательность. Такое объединение позволяло исключать из рассмотрения признаки, содержащие случайно сгенерированные значения, тем самым повышая качество решения задачи регрессии.

В **разделах 2.3** и **2.4** приведены результаты построения нейросетевых моделей для ряда наборов данных, взятых соответственно из комплектов Numeric и Regression-datasets базы данных WEKA [С3].

На основании изучения задач главы 2 был выделен ряд алгоритмов, перспективных с точки зрения их совместного использования в качестве методов АСВП.

Глава 3 посвящена созданной методике построения ИНС.

Как было упомянуто выше, методика представляет собой усовершенствование традиционного способа решения задач нелинейной регрессии. Применительно к многослойным персептронам, методика предполагает использование некоторого алгоритма выделения оптимального набора входных признаков из полного набора имеющихся. Алгоритм получен на основании изучения множества модельных и прикладных задач.

В общем случае, при наличии N входных признаков для поиска наилучшего решения задачи регрессии необходимо перебрать $2^N - 1$ возможных наборов входных признаков, для каждого из которых традиционным способом решить задачу регрессии. При больших значениях N это потребует существенных вычислительных затрат.

Алгоритм состоит из ряда этапов, выстроенных в порядке возрастания подробности исследования. На каждом этапе производится прореживание набора входных признаков с помощью определенных методов АСВП, и на основании выделенного набора строится промежуточное решение. Если на каком-то этапе полученного решения достаточно, то следующие этапы можно не выполнять («достаточность» полученного решения определяется целями исследования или заказчиком исследования). Решением задачи, полученным с помощью методики, считается лучшее из промежуточных решений, оптимальным набором входных признаков – набор, соответствующий лучшему промежуточному решению.

На нулевом этапе задача регрессии решается на полном наборе имеющихся входных признаков. С получившимся *исходным решением* будут сравниваться все последующие найденные решения.

На 1 этапе для отбора используются алгоритмы корреляционного анализа и кросс-энтропийного анализа. Цель этапа – исключение малосущественных входных признаков.

На 2 этапе используются алгоритмы семейства анализа весов нейронной сети (АВНС). Цель этапа – исключение малосущественных входных признаков.

На 3 этапе используется алгоритм последовательного добавления входных признаков (пошаговая регрессия). Цель этапа – выявление наиболее существенных входных признаков, а также исключение взаимозависимых входных признаков.

Выбор последовательности алгоритмов также обусловлен вычислительными затратами. Были получены следующие оценки.

Теорема: *Пусть P – число примеров в тренировочном наборе данных, Q_i – число входных признаков, используемых на i-ом этапе методики, N_h – число нейронов первого скрытого слоя в используемых многослойных персептронах. Тогда:*

1. *Вычисления 0-ого этапа требуют $O(\gamma P Q_0 N_h)$ операций, где γ – число эпох обучения, $\gamma \sim 10^2 - 10^3$.*
2. *Вычисления 1-ого этапа требуют $O(PQ_1)$ операций, в случае распараллеливания вычислений по входных признакам – $O(P)$ операций.*
3. *Вычисления 2-ого этапа требуют $O(\gamma P Q_2 N_h)$ операций, где γ – число эпох обучения, $\gamma \sim 10^2 - 10^3$.*
4. *Вычисления 3-его этапа требуют $O(\gamma P Q_3^3 N_h)$ операций, в случае распараллеливания вычислений по входным признакам – $O(\gamma P Q_3^2 N_h)$ операций, где γ – число эпох обучения, $\gamma \sim 10^2 - 10^3$.*

Для многослойного персептрона с 1 скрытым слоем верхним пределом значения N_h является значение P [С4].

Чем «глубже» этап (начиная с 1-ого), тем большие вычислительные затраты необходимы. Последовательное сокращение числа входных признаков

от этапа к этапу, когда $Q_0 > Q_1 > Q_2 > Q_3$, позволяет существенно сократить число операций, требуемых для более поздних этапов.

Информация о наиболее существенных входных признаках, полученная с помощью предложенной методики для какой-либо задачи математического моделирования, может использоваться в качестве вспомогательной для других исследований.

Также предложены критерии для включения новых методов в методику и рассмотрен специальный случай применения методики для 4-х ступенчатого алгоритма анализа многомерных временных рядов [С5].

Раздел 3.3 посвящен разработанному алгоритму параллельного обучения множества многослойных персептронов (МСП) с одним скрытым слоем, основанному на стандартном алгоритме обучения методом обратного распространения ошибки (с использованием момента обучения). Разработанный алгоритм может использоваться для ускорения нейросетевых вычислений на 0-м, 2-ом и 3-ем этапах методики.

В алгоритме использовались следующие допущения: 1) все обучаемые МСП должны обучаться на одном и том же наборе входных данных, выходные наборы данных могут различаться, 2) должны быть идентичны архитектуры всех обучаемых нейронных сетей (МСП с 1 скрытым слоем). Такие допущения позволяют существенно ускорить обучение за счет эффективного распараллеливания итерации (эпохи) обучения для всех МСП.

Алгоритм предполагает одновременное обучение не более чем некоторого числа МСП. При этом, как только завершается обучение одного из МСП, сразу начинается обучение следующего МСП из «очереди на обучение». Значения параметров обучения могут быть индивидуальны для каждого МСП.

Рассмотрены особенности реализации предложенного алгоритма, предназначенной для работы с графическими процессорами (GPU) компании NVIDIA с использованием технологии CUDA. В реализации алгоритма используются вычисления одинарной точности.

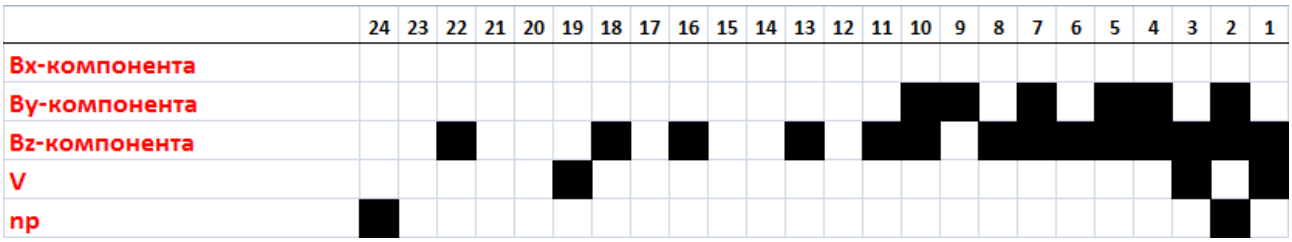
В практических исследованиях (в частности, при решении задачи из раздела 4.2) было опробовано одновременное обучение до 256 МСП.

Для проведения ряда вычислений для задачи из раздела 4.2 с помощью реализации алгоритма для графических процессоров компании NVIDIA потребовались (оценочно) примерно в 100 раз меньшие временные затраты, по сравнению с рядом коммерческих нейросетевых пакетов, реализующих стандартный алгоритм обратного распространения ошибки.

Глава 4 посвящена решению прикладных задач (задач реального мира).

В **разделе 4.1** рассмотрена задача прогнозирования среднечасовых значений геомагнитного Dst-индекса [С6] на 1 час вперёд. Исследовалась зависимость значения Dst-индекса от значений трех компонент межпланетного магнитного поля B_x, B_y, B_z, скорости V и плотности протонов n_p солнечного ветра, измеряемых в течение последних 24 часов относительно прогнозируемого значения индекса, – т.о. имелось 120 входных признаков. В качестве исходного набора данных были взяты почасовые временные ряды за 1999-2003 годы, содержащие усредненные почасовые значения указанных величин – всего 35000 точек.

С помощью методики выделены следующие временные диапазоны (т.е. значения задержек) относительно прогнозируемого значения Dst-индекса, в пределах которых значения различных величин оказывают наибольшее влияние на поведение прогнозируемого Dst-индекса.



Ниже приведены значения линейного коэффициента корреляции r и коэффициента множественной детерминации R^2 для решений задачи регрессии, получаемых на разных этапах методики:

Этап методики	Вх. признаков	r	R^2
Исходная нейронная сеть	120	0.97	0.93
Первый этап методики	74	0.97	0.94
Второй этап методики	33	0.96	0.93
Третий этап методики	25	0.96	0.93

Наиболее важные выводы, полученные при решении данной задачи:

1. С помощью разработанной методики выделены наиболее существенные признаки, соответствующие важным значениям задержки при прогнозировании Dst-индекса:
   a. B_x компонента ММП – не используется,
   b. B_y компонента ММП с задержками от 2 до 10 часов,
   c. B_z компонента ММП с задержками до 22 часов,
   d. Скорость солнечного ветра V с задержками 19, 3 и 1 час,
   e. плотность протонов в солнечном ветре n_p с задержками 24 и 2 часа.
2. Наиболее важный интервал значений задержек – от 3 до 5 часов.
3. Удалось сократить количество используемых признаков до 5 раз с незначительной потерей качества прогноза.

В **разделе 4.2** рассмотрено решение задачи из области электроразведки (магнитотеллурики, МТ) по восстановлению распределения электропроводности двумерного геоэлектрического разреза (участка земной коры в вертикальной плоскости) на основании измеренных значений ЭМ поля на поверхности земли. Процесс решения заключался в построении нелинейной регрессионной нейросетевой модели, описывающей поведение вектора искомых геофизических параметров разреза на основе наблюдаемых на поверхности Земли значений электромагнитного поля. Эти параметры представляют собой распределение электропроводности в разных точках исследуемой области, геометрические размеры отдельных подобластей (геологических структур) и т.п.

Приведенные в разделе результаты применения разработанной методики понимаются как усовершенствование и сравниваются с результатами

предложенного в работе [С7] метода построения нейросетевых моделей для решения ряда задач регрессии из области МТЗ (магнитотеллурического зондирования). Соответствующие наборы эталонных данных для численных экспериментов предоставлены авторами работы [С8].

Математическая постановка рассматриваемой задачи состоит в следующем. Обратная задача МТЗ *в конечно-параметрическом k-ом классе сред* сводится к системе нелинейных уравнений относительно вектора макропараметров среды $\gamma = (\gamma_1, \dots, \gamma_{N_k})$ вида [С8]:

$$A_k\gamma = \beta, \quad \gamma \in \Gamma_k \subset R^{N_k}, \quad \beta \in C^{M_k} \qquad (*)$$

где A_k – заданный дискретный (конечно-разностный) оператор прямой задачи, $\beta = (\beta_1, \beta_2, \dots, \beta_{M_k})$ – вектор данных, представляющий собой набор значений импедансов, или иных характеристик МТ поля, измеренных на поверхности Земли в точках r_i, $i = 1, \dots, N_r$, на частотах ω_j, $j = 1, \dots, N_\omega$, и упорядоченных определенным образом, R^{N_k}, C^{M_k} – эвклидовы пространства вещественных векторов размерности N_k, M_k соответственно, Γ_k – область допустимого изменения макропараметров.

Нейросетевой подход к решению коэффициентных задач указанного типа реализует аппроксимационную парадигму, согласно которой приближенное решение обратной задачи *в заданном конечно-параметрическом классе сред* ищется в виде заданной функции от входных данных β_1,…,β_{M_k} [С8] и неопределенных коэффициентов $a_1, \dots, a_J$:

$$\gamma^* \approx \Psi_{\mathrm{k}}^{\mathrm{app}}(a_1, \dots, a_J, \beta_1, \dots, \beta_{M_k})$$

В качестве семейства функций-аппроксиматоров рассматривались многослойные персептроны. Способ построения аппроксимационной конструкции рассмотрен в главе 1. Построенную нейросетевую модель $\Psi_{\mathrm{k}}^{\mathrm{app}}$ можно рассматривать как приближенный обратный НС оператор для системы (*), который позволяет получать её решения для любого предложенного вектора входных данных. Таким образом, получаемые решения, являющиеся

результатом построения нелинейной регрессии, являются интерполяционными решениями обратной задачи МТЗ [С8].

Входные данные представляли собой значения комплексной амплитуды и фазы компонент тензора импеданса двумерной среды, наблюдаемые на разных частотах электромагнитного поля в разных точках на поверхности земли, а выходные данные – значения электропроводности в разных точках исследуемой области.

Исходя из примененной параметризации разрезов, размерность входных данных составляла $D_I = 6552$ (модули и фазы диагональных компонент тензора импеданса, пространственная сетка из 126 точек, частотная сетка из 13 точек), а размерность выходных данных – $D_O \approx 3 * 10^2$.

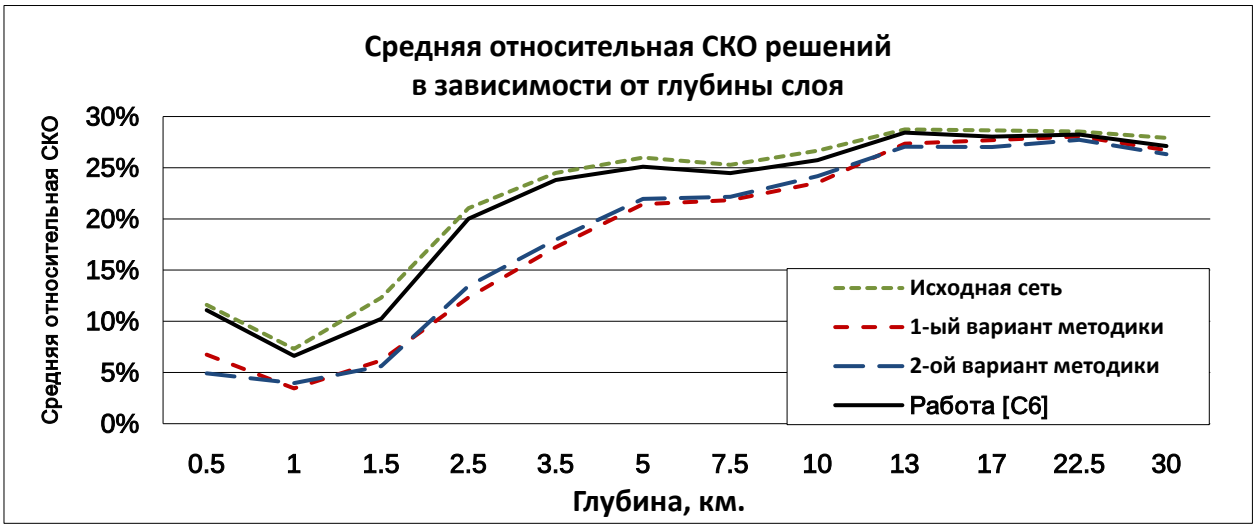
Исходная полная задача нелинейной регрессии была поделена на множество более простых. Рассматривалось множество задач с тем же набором входных признаков, но каждая с одним выходным признаком, соответствовавшим электропроводности в одной из точек исследуемой области.

Для решения рассматриваемой задачи были применены два варианта разработанной методики, различавшиеся используемым набором параметров методики и, как следствие, степенью подробности исследования. В первом варианте методики не использовались методы 1-го этапа, но дважды применялся метод 2-го этапа АВНС, являющийся вычислительно существенно более дорогим, чем методы 1-го этапа. Во втором варианте использовались как методы 1-го этапа, так и метод 2-го этапа. В обоих случаях метод 3-его этапа не использовался из-за высокой вычислительной стоимости соответствующих расчётов.

В первом варианте методики был применен разработанный алгоритм параллельного обучения множества МСП (см. раздел 3.3).

На диаграмме ниже в качестве примера приведены относительные значения среднеквадратичной ошибки (усредненные по слою для слоев, находящихся на различной глубине) на независимом наборе данных для обоих

вариантов методики, в зависимости от глубины залегания ячейки для одной из использованных схем параметризации неоднородной двумерной среды («модель 5.0»).



В таблице приведено число отобранных входных признаков (усредненное по слою для слоев, находящихся на различной глубине) для обоих вариантов методики, в зависимости от глубины залегания ячейки для той же схемы параметризации «модель 5.0».

НС\глубина, км.	**0.5**	**1**	**1.5**	**2.5**	**3.5**	**5**	**7.5**	**10**	**13**	**17**	**22.5**	**30**
Исходная НС	6552	6552	6552	6552	6552	6552	6552	6552	6552	6552	6552	6552
1 вариант мет.	33	47	30	18	42	42	44	45	48	65	43	47
2 вариант мет.	59	68	55	51	74	73	82	79	78	97	88	149

Также в диссертации приведены результаты для ряда других моделей параметризации разреза.

Основной вывод, полученный при решении данной задачи: предложенная методика показала себя эффективной. Продемонстрировано заметное увеличение точности решения задачи нелинейной регрессии по сравнению как с исходным решением, так и с решением, предложенным в работе [С7], при значительном (примерно на два порядка) сокращении количества входных признаков. Анализ адаптивно отбираемого списка признаков показывает, что этот список хорошо согласуется с априорными физическими соображениями о сравнительной существенности признаков.

В **разделе 4.3** описывается пример использования разработанной методики как части решения задачи типа «структура-свойство» из области токсикологии. Задача связана с изучением токсичности различных химических соединений для бактерии *Tetrahymena pyriformis*.

В качестве количественной величины, характеризующей токсичность относительно *T. pyriformis*, рассматривался индекс $\log(IGC50^{-1})$ [C9], стандартный показатель токсичности химических соединений.

Каждое из химических соединений описывалось совокупностью из 2223 характеристик (дескрипторов) молекул данных соединений [C10], рассчитанных с помощью стандартных пакетов программ физико-химического моделирования. Дескрипторы включали различные квантовомеханические, химико-топологические и прочие характеристики молекул.

В результате исследования предложенных наборов данных с помощью предложенной методики была выделена группа из 12 наиболее существенных дескрипторов. На основании этих дескрипторов был предложен нейросетевой алгоритм для окончательного решения рассматриваемой задачи.

В **главе 5** приведено описание созданного программного комплекса (ПК), реализующего все рассмотренные алгоритмы, необходимые для применения методики при изучении незнакомых задач. ПК реализован как приложение, имеющее графический интерфейс пользователя и работающее под операционной системой Microsoft Windows. Для ряда алгоритмов ПК позволяет осуществление параллельных вычислений при наличии локальной сети с машинами, на которых установлена соответствующая программная часть ПК. Реализация алгоритма параллельного обучения многослойных персептронов, упомянутая в разделе 4.3, существует в виде отдельного автономного модуля.

В **заключении** кратко сформулированы основные результаты, полученные в диссертационной работе:

1. Создана, исследована и опробована для решения ряда задач методика построения искусственных нейронных сетей (многослойных

персептронов) для решения задач нелинейной регрессии и прогнозирования на основе отбора существенных входных признаков. Намечены дальнейшие пути развития предложенной методики. Предложены пути включения в имеющуюся методику новых алгоритмов.

2. Создан единый комплекс программ, реализующий все использованные алгоритмы для исследования данных с помощью предложенной методики.
3. Показано, что применение разработанной методики при использовании нейронных сетей позволяет получать решения задач, в среднем имеющие более низкую погрешность и/или использующие меньшее количество входных переменных.
4. При решении задачи по прогнозированию значений Dst-индекса (раздел 4.1) с использованием разработанной методики удалось сократить количество используемых признаков до 5 раз с небольшой потерей качества прогноза. Были выделены наиболее существенные входные признаки; сделанные на основании полученного набора выводы согласуются с современными представлениями о физике рассматриваемого в задаче процесса.
5. Применение разработанной методики для решения задачи по восстановлению распределения электропроводности участка земной коры (раздел 4.2) позволило существенно (на два порядка) сократить число рассматриваемых входных признаков при улучшении качества решения.
6. Разработан алгоритм параллельного обучения множества многослойных персептронов с одним скрытым слоем. Создана реализация данного алгоритма с использованием технологии CUDA компании NVIDIA. Продемонстрировано, что скорость работы программы с использованием графического процессора (GPU) на полтора-два порядка выше, чем у программ (в том числе коммерческих), использующих центральный процессор (CPU), что открывает новые возможности в исследовании

наборов данных большой размерности с помощью нейросетевых алгоритмов.

Список публикаций автора по теме диссертации

1. А.Г.Гужва, С.А.Доленко, Е.А.Оборнев, И.Г.Персианцев, М.И.Шимелевич, Ю.С.Шугай. Использование адаптивных алгоритмов отбора существенных признаков при нейросетевых решении обратной задачи электроразведки. *Нейрокомпьютеры: разработка, применение*, 2010, № 3, стр. 46-54.
2. А.Г.Гужва, С.А.Доленко, И.Г.Персианцев, Ю.С.Шугай. Многоступенчатый алгоритм на основе комитета нейронных сетей для анализа многомерных временных рядов. *Нейрокомпьютеры: разработка, применение*, 2010, № 3, стр. 4-13.
3. А.Г.Гужва, С.А.Доленко, И.Г.Персианцев. Методика отбора существенных входных признаков при нейросетевом решении задач регрессии. *Нейрокомпьютеры: разработка, применение*, 2010, № 3, стр.20-32.
4. А.Г.Гужва, С.А.Доленко, И.Г.Персианцев, Ю.С.Шугай. Сравнительный анализ методов важности входных переменных при нейросетевом прогнозировании. *VIII Всероссийская научно-техническая конференция «Нейроинформатика-2006»*, сборник научных трудов, часть 1, стр.31-39, М., МИФИ, 2006.
5. А.Г.Гужва, С.А.Доленко, И.Г.Персианцев, Ю.С.Шугай, В.Г.Еленский. Отбор существенных переменных при нейросетевом прогнозировании: сравнительный анализ методов. *IX Всероссийская научно-техническая конференция «Нейроинформатика-2007»*, сборник научных трудов, часть 2, стр.251-258, М., МИФИ, 2007.
6. А.Г.Гужва, С.А.Доленко, И.Г.Персианцев, Ю.С.Шугай. Сравнительный анализ методов определения существенности входных переменных при нейросетевом моделировании: методика сравнения и ее применение к известным задачам реального мира. *X Всероссийская научно-техническая конференция «Нейроинформатика-2008»*, сборник научных трудов, часть 2, стр.216-225, М., МИФИ, 2008.
7. А.Г.Гужва, С.А.Доленко, И.Г.Персианцев, Ю.С.Шугай. Многоступенчатый алгоритм на основе комитета нейронных сетей для прогнозирования и поиска предвестников в многомерных временных рядах. *XI Всероссийская научно-техническая конференция «Нейроинформатика-2009»*, сборник научных трудов, часть 2, стр.116-125, М., МИФИ, 2009.
8. А.Г.Гужва, С.А.Доленко, И.Г.Персианцев. Многократное ускорение нейросетевых вычислений с использованием видеоадаптера. *XI Всероссийская научно-техническая конференция «Нейроинформатика-2009»*, сборник научных трудов, часть 2, стр.126-133, М., МИФИ, 2009.
9. А.Г.Гужва, С.А.Доленко, Е.А.Оборнев, И.Г.Персианцев, М.И.Шимелевич. Нейросетевой метод решения обратной задачи геоэлектрического мониторинга параметров в трехмерных средах. *XII Всероссийская научно-*

техническая конференция «Нейроинформатика-2010», сборник научных трудов, часть 2, стр. 111-121, М., МИФИ, 2010.

10. A.G.Guzhva, S.A.Dolenko, I.G.Persiantsev, J.S.Shugai. Comparative Analysis of Methods for Determination of Significance of Input Variables in Neural Network Modeling: Procedure of Comparison and its Application to Model Problems. *8th International Conference "Pattern Recognition and Image Analysis: New Information Technologies" (PRIA-8-2007)*: Conference Proceedings, V.2, pp.29-32. Yoshkar-Ola, 2007.
11. J.S.Shugai, A.G.Guzhva, S.A.Dolenko, I.G.Persiantsev. An Algorithm for Construction of a Hierarchical Neural Network Complex for Time Series Analysis and its Application for Studying Sun-Earth Relations. *8th International Conference "Pattern Recognition and Image Analysis: New Information Technologies" (PRIA-8-2007)*: Conference Proceedings, V.2, pp.355-358. Yoshkar-Ola, 2007.
12. Guzhva, A., Dolenko, S., Persiantsev, I. Multifold Acceleration Of Neural Networks Computations Using GPU. C.Alippi et al (Eds.): ICANN 2009, Part I. (Lecture Notes in Computer Science, 2009, V.5768, pp.373-380.) Springer-Verlag Berlin Heidelberg 2009.
13. A.G.Guzhva, S.A.Dolenko, E.A.Obornev, I.G.Persiantsev, M.I.Shimelevich, J.S.Shugai. Use of Significant Feature Selection Adaptive Algorithms in Neural Network Based Solution of the Inverse Problem of Electrical Prospecting. *9th International Conference "Pattern Recognition and Image Analysis: New Information Technologies" (PRIA-9-2008)*: Conference Proceedings. Vol.1, pp.215-218. Nizhni Novgorod, 2008.
14. Dolenko, S., Guzhva, A., Persiantsev, I., Shugai, J. Multi-stage Algorithm Based on Neural Network Committee for Prediction and Search for Precursors in Multi-dimensional Time Series. C.Alippi et al (Eds.): ICANN 2009, Part II. (Lecture Notes in Computer Science, 2009, V.5769, pp.295-304.) Springer-Verlag Berlin Heidelberg 2009.
15. Dolenko, S., Guzhva, A., Obornev, E., Persiantsev, I., Shimelevich, M. Comparison of Adaptive Algorithms for Significant Feature Selection in Neural Networks Based Solution of the Inverse Problem of Electrical Prospecting. C.Alippi et al (Eds.): ICANN 2009, Part II. (Lecture Notes in Computer Science, 2009, V.5769, pp.397-405.) Springer-Verlag Berlin Heidelberg 2009.

Цитированная литература

C1. CUDA. URL: http://www.nvidia.com

C2. G.Cybenko. Approximation by superpositions of sigmoidal function. Mathematics of Control, Signals, and Systems, 1989, v.2, pp.304-314.

C3. База данных WEKA, URL: http://www.cs.waikato.ac.nz/ml/weka/

C4. G. Huang, H. A. Babri. Upper Bounds of the Number of Hidden Neurons in Feedforward Networks with Arbitrary Bounded Nonlinear Activation Functions. // IEEE Transactions on Neural Networks, vol. 9, no. 1, January 1998.

C5. S.A. Dolenko, Yu.V. Orlov, I.G. Persiantsev, Yu.S. Shugai. Neural Network Algorithms for Analyzing Multidimensional Time Series for Predicting Events and Their Application to Study of Sun-Earth Relations. // Pattern Recognition and Image Analysis, 2007, Vol.17, No.4, pp. 584-591.

C6. Ю. С. Шугай. Разработка нейросетевых алгоритмов анализа многомерных временных рядов и их применение при исследовании солнечно-земных связей. Дисс. канд. физ.-мат. наук, Москва 2006.

C7. Е. А. Оборнев. Инверсия двумерных магнитотеллурических данных на основе нейросетевой аппроксимации. Дисс. канд.физ.-мат. наук, Москва 2007.

C8. Шимелевич М.И., Оборнев Е.А. Аппроксимационный метод решения обратной задачи МТЗ с использованием нейронных сетей. // Физика Земли, 2009, 12. С.22-38

C9. Schultz, T.W. 1997. TETRATOX: Tetrahymena pyriformis population growth impairment endpoint-A surrogate for fish lethality. // Toxicol. Methods 7: 289-309. URL: http://www.vet.utk.edu/TETRATOX/

C10. Environmental Toxicity Prediction Challenge. URL: http://www.cadaster.eu/node/65