

ИСПОЛЬЗОВАНИЕ БАЙЕСОВСКОЙ РЕГУЛЯРИЗАЦИИ МОДЕЛИ АНАЛИЗА УСЛОВИЙ И ОБРАЗА ЖИЗНИ ОБУЧАЮЩЕЙСЯ МОЛОДЕЖИ

Горбатков С.А.¹, Горбаткова Е. Ю.²

¹Уфимский филиал Финансового университета при правительстве Российской Федерации, г. Уфа, Россия, E-mail: gorbatkovaeu@mail.ru;

²ФГБОУ ВПО «Башкирский государственный педагогический университет им. М. Акмуллы», Уфа, Республика Башкортостан

В статье изложен алгоритм обработки данных комплексной оценки состояния здоровья студентов высших учебных заведений. Определена актуальность изучения условий и образа жизни студенческой молодежи, так как здоровье подрастающего поколения составляет основу экономического и военного потенциала страны. Применительно к исследованиям, связанным с изучением образа жизни и состоянием здоровья обучающейся молодежи, разработана концепция повышения устойчивости нейросетевой модели, которая восстанавливает многомерные нелинейные зависимости, скрытые в данных анкетных опросов и тестовых заданий студентов. Описано также применение метода байесовской регуляризации при оценке условий жизни обучающейся молодежи.

Ключевые слова: здоровьесбережение обучающейся молодежи, студенты, нейросеть, байесовская регуляризация, здоровый образ жизни, анкетирование, здоровье, заболеваемость

THE USE BAYESIAN REGULARIZATION FOR ASSESSING CONDITIONS AND LIFESTYLES STUDENTS

Gorbatkov S.A.¹, Gorbatkova E.J.²

¹Federal State-Funded Educational Institution of Higher Professional Education "Financial University under the Government of the Russian Federation", E-mail: gorbatkovaeu@mail.ru;

²«Bashkir state pedagogical university n. a. M. Akmulla», E-mail: gorbatkovaeu@mail.ru

The article describes the process of data processing integrated assessment of the health status of students in higher education. Relevance of the study of conditions and way of life of students was determined, as the health of the younger generation is the basis of economic and military potential of the country. The concept of increasing of stability neural network model, which restores the multidimensional nonlinear relationships hidden in questionnaires data and test answers of students was developed as applied to research related with the study of lifestyle and health status of students. Also there was described the use of Bayesian regularization method in the evaluation of the living conditions of students.

Keywords: zdoorovesberezheniya educate young people, students, neural network, a Bayesian regularization, healthy lifestyle, questioning, health, morbidity.

Результаты ряда научных исследований [3, 4, 9] свидетельствуют о том, что сложившиеся в предшествующие годы тенденции ухудшения здоровья студентов принимают устойчивый характер. Наблюдаются неблагоприятная динамика основных показателей здоровья молодежи, учащение перехода острых форм заболеваний в хронические и т.д. Объясняется это не только экономической ситуацией, социальным напряжением и ухудшением экологической обстановки в России. Важнейшим фактором является отношение к здоровью как к индивидуальной и социальной ценности. К сожалению, далеко не все студенты соблюдают нормы и правила здорового образа жизни [7]. Особенностью современных молодых людей является повышение распространенности среди них аутоагрессивных видов поведения, таких как курение, употребление алкоголя, наркотиков,

связанное с риском половое поведение, что существенно усугубляет негативные тенденции состояния здоровья, наблюдаемые в настоящий период.

Очевидно, что здоровье молодежи является важнейшей проблемой государства, так как составляет основу кадрового, экономического потенциала страны, определяет ее обороноспособность. Репродуктивное здоровье молодежи имеет решающее значение для воспроизводства населения и здоровья будущих поколений.

Следует отметить, что высшие учебные заведения могут организовать эффективную оздоровительную и реабилитационную работу среди студенческой молодежи. Одним из результативных вариантов решения рассматриваемой проблемы является внедрение в деятельность образовательных организаций современных комплексных медико-педагогических технологий, что является целью нашего исследования [2].

Учитывая актуальность вышеизложенного, мы проводим комплексную оценку состояния здоровья студентов 1–4-х курсов высших учебных заведений г. Уфа (БГУ, БГПУ, УГАТУ и др.). В рамках нашего исследования проводится изучение заболеваемости студентов по данным обращаемости, с временной утратой трудоспособности в случаях и днях госпитализированной заболеваемости [4]. Особое внимание было уделено анализу репродуктивного здоровья студенток. Производилась также оценка некоторых показателей психического здоровья студентов посредством тестирований.

Рассматриваемая в статье медико-социальная система характеризуется высоким уровнем априорной неопределенности.

Известны три группы методов, способных эффективно работать в отмеченных сложных условиях при сильных зашумленных данных:

- 1) нейросетевые методы (НСМ) [11];
- 2) методы нечеткого описания модели и правил логического вывода [6];
- 3) нейро-нечеткие методы [13].

В данной статье использовали адаптивные нейросетевые методы. Однако, несмотря на их эффективность в условиях сильного зашумления данных, здесь возникли две проблемы:

- неустойчивость модели к вариации входных данных при выходе за пределы некоторого критического уровня шума при оценке анкетных данных студентов и результатов психологических тестирований;
- оценка и обеспечение адекватности нейросетевой модели в условиях нарушения практически всех предпосылок классического регрессионного анализа [1].

Для решения этих проблем в статье разработан байесовский подход (концепция) нейросетевого моделирования, подробно изложенная ниже.

1. Постановка задачи моделирования

Пусть в качестве инструмента моделирования при обработке анкетных данных используется нейросетевой эмулятор (программный продукт, например Neuro Solutions 5.0), реализующий нейросеть типа многослойного перцептрона (Multi Layer Perceptron (MLP)) с алгоритмом обучения типа обратного распространения ошибки (Back Propagation Error (BP)). MLP-BP-нейросеть является универсальным аппроксиматором [11]:

$$\hat{Y}_m = F(\bar{X}, W); m = \overline{1, M},$$

где \hat{Y}_m – m -я моделируемая (эндогенная) расчетная величина, т.е. показатель обследования, характеризующий определенный уровень здоровья;

$\bar{X} = (X_1, X_2, \dots, X_j, \dots, X_n)$ – вектор экзогенных (объясняющих) переменных (далее коротко «факторов»), которые могут быть как количественными, так и качественными; W – матрица синаптических весов, т.е. весовых коэффициентов сигналов $X_{p,q}$, поступающих от p -го нейрона к следующему за ним нейрону сети в направлении прямого распространения сигналов от входа к выходу; $F(\cdot)$ – оператор нейросетевого отображения сигналов, представляющий собой композицию двух операторов F : $F_1 \circ F_2$, который в каждом нейроне (элементарном процессоре) нейросети реализует две последовательные операции:

1. $F_2(\cdot)$ – весового суммирования сигналов поступающих сети на вход p -го нейрона в промежуточном (скрытом) слое; $S_p = \sum_{q=1}^n W_{p,q} X_{p,q}$;

2. $F_2(\cdot)$ – нелинейного преобразования результата суммирования, например с помощью функции гиперболического тангенса $F_1(S_p) = \int(S_p) = \text{th}(S_p)$.

Функцию $\int(S_p)$ называют активационной, или передаточной. Обычно в выходном слое нейросети используется линейная функция.

Замечание 1. Задача аппроксимации данных в математическом аспекте является обратной задачей, т.е. требуется найти причины явления по заданному следствию – наблюдаемым характеристикам системы. Решение таких задач гораздо сложнее, чем прямых задач прямого расчета. С содержательной точки зрения задачи аппроксимации данных суть задачи восстановления многомерных нелинейных закономерностей, скрытых в данных. Физики их называют также задачами интерпретации. Как известно [11], задачи этого класса являются некорректно поставленными по Адамару и поэтому нуждаются в специальных мерах регуляризации.

Замечание 2. Принципиальным отличием нейросетевых методов от классических методов регрессионного анализа [6], где вектор факторов \bar{X} считается детерминированным, а

выходные величины $\{Y_m\}$ — случайными величинами, является то, что в НСМ вектор \vec{X} тоже считается случайным. Этот подход к построению модели значительно увеличивает прогностическую силу НСМ, приближая его к реальным условиям моделирования, однако исключает применимость традиционных методов оценки адекватности модели.

Замечание 3. В НСМ относительно просто решается вопрос учета многомерности вектора выходных величин $\vec{Y}=(Y_1, Y_2, \dots, Y_m, \dots, Y_M)$. Нужно лишь в выходном слое сети взять не один, а несколько нейронов по числу компонент вектора \vec{Y} .

Можно также выходной слой реализовать в виде агрегатора (свертки), т.е. НСМ будет давать на выходе некоторый обобщенный критерий $\Phi(\vec{Y}(\vec{X}))$ уровня здоровья обучающейся молодежи.

2. Идея разработки концепции регуляризации нейросетевой модели

Идея 1 предлагаемой концепции байесовского подхода к регуляризации НСМ вполне понятна на интуитивном уровне. Восстанавливаемая функциональная зависимость $Y(\vec{X})$ аппроксимируется не одной моделью, а ансамблем адаптивных параметрических моделей с априорным выбором их вида, причем все выходные характеристики моделирования получаются путем осреднения (прямого или весового) на отфильтрованном по критерию качества НСМ ансамбле. При этом «эксплуатируется» широко известная закономерность теории вероятностей: среднее значение случайной величины \bar{Y} колеблется в наблюдениях гораздо меньше, чем сама случайная величина Y , причем тем меньше, чем больше число наблюдений N .

Идея 2, основная для разработки предлагаемой концепции, состоит в том, что регуляризация НСМ на байесовском ансамбле априорно выбираемых нейросетевых моделей-гипотез $\{h_q|H\}$, $q=\overline{1, Q}$, $h_q \in H$, где H – общий для всех гипотез класс моделей (мета-гипотеза) и затем фильтруемых по качеству, укладывается в общую парадигму теории регуляризации обратных задач А.Н. Тихонова [8]. Суть этой парадигмы состоит в том, что регуляризация и соответственно повышение устойчивости модели достигается «сужением класса искомых решений обратной задачи» с использованием априорной информации о решении.

Пути реализации этой парадигмы при разработке концепций могут быть разными: в теории А.Н. Тихонова [8] априорная информация используется при конструировании специальных стабилизаторов Тихонова, а в нашей концепции — при задании архитектуры нейросетей $\{h_q\}$ в байесовском ансамбле и выборе активационных функций.

3. Алгоритм метода регуляризации НСМ

В известных работах по байесовской регуляризации нейросетевых моделей [10], как правило, задача решается аналитически на основе вычисления функции правдоподобия. Это сопряжено с введением допущения об априорном задании закона распределения плотности вероятности шумов в данных, которое на практике часто не выполняется.

Мы отказались от этого весьма жесткого допущения, заменив его вычислением вероятностного частотного критерия апостериорной фильтрации НСМ ансамбля по качеству аппроксимации данных. Детализируем алгоритм предлагаемого квазيبайесовского метода регуляризации (КБМР).

1. Опишем алгоритм метода по шагам, т.е. будем полагать, что отсутствуют какие-либо предпочтения между несколькими методами моделирования данных (мета-моделями) ($P(H) = \text{const}$), и мы можем выбрать мета-гипотезы на основе экспертных оценок и предварительных вычислительных экспериментов.

2. Построим байесовский ансамбль НСМ, который показан на рисунке 1, где введены обозначения: D – база данных; M – блоки масштабирования (нормировки); Cp – блок усреднения расчетных характеристик на отфильтрованном байесовском ансамбле; ДМ – блоки демасштабирования; S, C – множества выходных характеристик моделируемой системы; R2 – индекс множественной детерминации; RBF, MLP – типы нейросетей, входящих в байесовский ансамбль.

3. Принадлежность всех $\{h_q\}$ одному классу H есть ограничение на нейросетевые модели – гипотезы, что и является механизмом регуляризации согласно основной парадигме регуляризации, используемой в монографии.

4. После обучения нейросетевых моделей-гипотез $\{h_q\}$ на экспериментальных примерах $\langle \bar{x}_i, y_i \rangle, i = \overline{1, N}$ должна быть разработана числовая мера оценки апостериорной вероятности гипотез $P(h_q|D|H)$ о порождении данных. Эта мера служит критерием отбора для процедуры (фильтрации) гипотез-нейросетей $\{h_q\}$, сужающим априорный ансамбль. В редуцированном ансамбле отсеиваются нейросети-гипотезы, у которых качество сети оказывается неудовлетворительным по выбранной числовой мере.

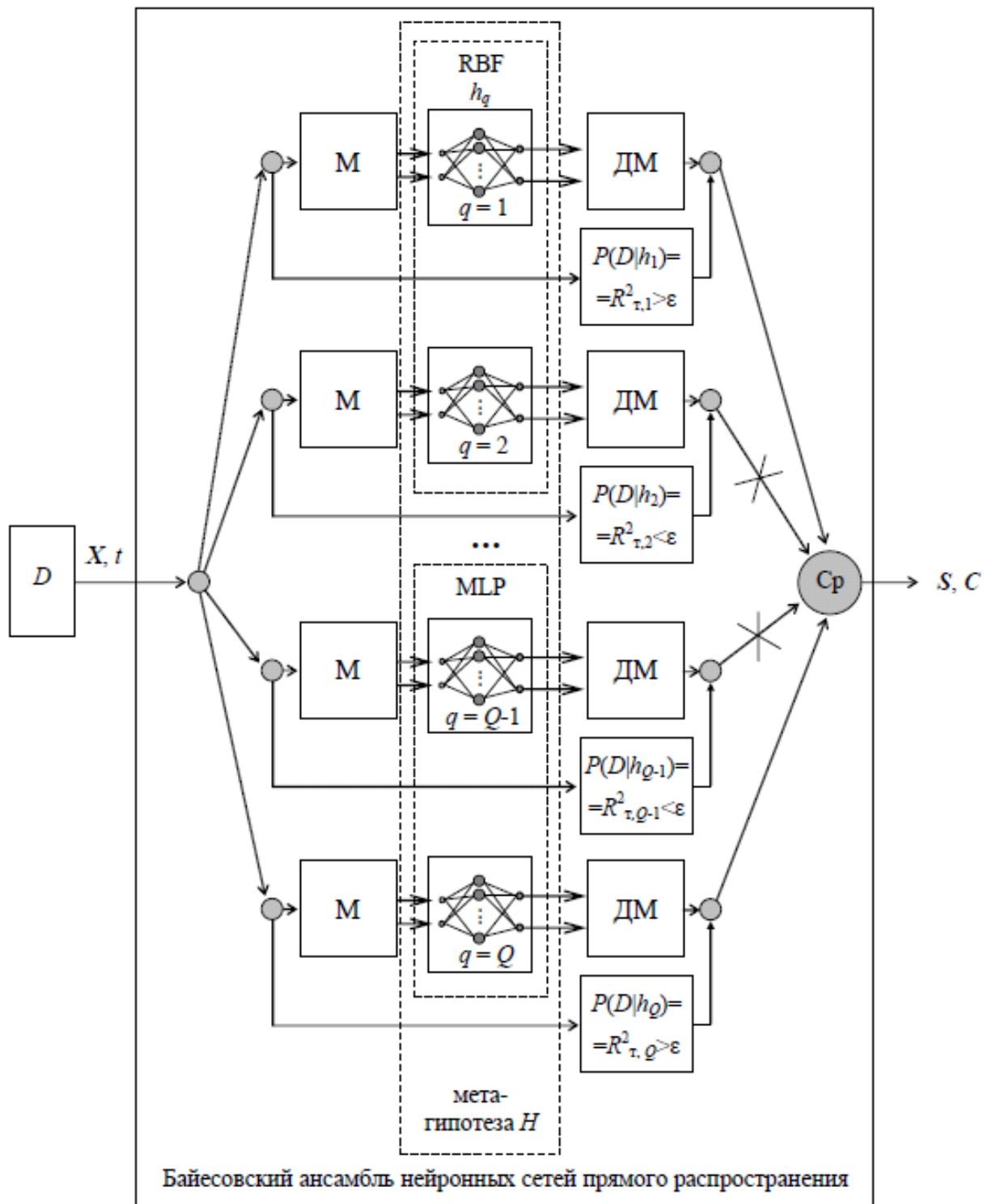


Рис. 1. Структура байесовского ансамбля нейросетевых моделей

Мы опирались на интуитивно понятное замечание, высказанное в [10]: «если функцию ошибок $E(\cdot)$ задать в виде числовой меры по модулю, то из максимизации $P(D|h_q, w)$ будет вытекать минимизация функции ошибок»

$$\max_{h_q} P(D | h_q, W) \Rightarrow \min_{h_q} \delta_{q,t} = \min_{h_q} \left| \frac{y_i - h_{qi}(\vec{x}, W)}{y_i} \right|. \quad (6)$$

Будем также учитывать, что в самом общем случае оценка максимального правдоподобия согласно [9] основывается на том, что «разные генеральные совокупности (population) генерируют различные данные, при этом происхождение любого заданного примера более правдоподобно для некоторой определенной совокупности, чем для остальных совокупностей».

Если под генеральными совокупностями понимать совокупности данных, порождаемых в соответствии с конкретной гипотезой-нейросетью $\{D | h_q(\bar{x}, W)\}$, то для оценки апостериорных вероятностей гипотез – нейросетевых моделей о порождении данных $P(h_q(w, \bar{x}) | D | H)$ можно предложить числовую меру:

$$P(h_q(W, \bar{x}) | D | H) \rightarrow P_q^* = N_q^* / N, q = \overline{1, Q} \in \Omega^{test}; \quad (7)$$

$$N_q^* = \{i_q^*\}; \delta_{q, i_q^*} \leq \xi; q = \overline{1, Q}; i = \overline{1, N}, \quad (8)$$

где: N_q^* – число «хороших» точек, в которых остатки $\{\delta_{q,i}\}$ имеют приемлемый уровень ξ для q -ой гипотезы – нейросети $\{h_q(\bar{X}, W)\}$; N – общее число примеров данных, одинаковое для всех гипотез-нейросетей; Ω^{test} – тестовое множество примеров; стрелка « \rightarrow » означает «оценку»; i_q^* – номер «хорошей» точки; ξ – экспертно задаваемый параметр, например $\xi=0,05\dots 0,1$.

Поскольку в предложенном методе не используется прямое вычисление функции правдоподобия $P(D|h,N)$ в формуле Байеса, мы назвали его квазيبайесовским методом регуляризации (КБМР).

По мнению авторов, в пользу предложенного критерия (7) – (8) говорят следующие соображения:

1) он очень прост в реализации и не требует никаких ограничивающих предположений о законах распределения шумов;

2) критерий является финишным и носит характер интегральной оценки качества объяснения сетью $h_q(W, \bar{x})$ данных: он «вбирает» в себя оценки всех процедур предобработки данных, а также алгоритмов выбора мета-гипотезы H , гипотез-нейросетей $\{h_q\}$ и качества их обучения;

3) в критерии (7)–(8) назначаемый уровень ξ играет роль параметра регуляризации, который согласует между собой два противоречивых требования – минимизацию критерия ошибок аппроксимации (6) и устойчивость НСМ, получаемые осреднением на байесовском ансамбле гипотез-нейросетей $\{h_q\}$.

При таком подходе можно оптимизировать этот параметр с учетом показателя устойчивости S . Оптимальное значение параметра ξ^* находится в результате оптимального итерационного процесса фильтрации байесовского ансамбля:

$$\xi^* = \arg \max_{\xi} [\bar{P}^*(\xi)]^{(k)}, k = 1, 2, \dots \quad \xi \leq \xi_0, \quad \xi_0 > 0, \quad (9)$$

где: черточка над \bar{P}^* означает усреднение по отфильтрованному ансамблю;

ξ_0 – заданное число (предельно допустимый уровень ошибок).

Фильтрация априорных гипотез

Теперь мы можем сузить пространство априорных гипотез $\{h_q(\bar{x}, W)\}$, т.е. провести их фильтрацию, отбраковав гипотезы с малой апостериорной вероятностью порождения данных D . Вычислим $\{P_q^*\}$ по (7)–(8) и оставим в ансамбле только те гипотезы, которые имеют достаточную апостериорную вероятность:

$$q^* : P_q [h_q(\bar{x}, W, \xi^*)] \geq \omega, \omega < 1, \quad (10)$$

где ω – экспертно задаваемый уровень апостериорной вероятности объяснения данных в байесовском ансамбле гипотез; ξ^* – параметр регуляризации по (9).

Здесь ω – второй параметр регуляризации.

Замечание. Для оптимизации пары параметров (ξ, ω) можно применить схему алгоритма EM из [10]; а именно: поочередно на каждом шаге итераций k фиксируется $\omega^{(k-1)}$ и оптимизируется $\xi^{(k)}$ по критерию (9), затем на следующем шаге $(k+1)$ фиксируется $\xi^{(k)}$ и оптимизируется ω :

$$\omega^{*(k+1)} = \arg \max_{\omega} [\bar{P}^*(\xi, \omega, k)], \quad \omega \geq \omega_0, \quad \omega_0 > 0. \quad (11)$$

где: ω_0 – предельно допустимый уровень снижения качества объяснения данных, определяемый контекстом задачи.

Результаты моделирования получаются путем усреднения на отфильтрованном ансамбле гипотез:

$$\bar{P}^* = \frac{\sum_{q^*=1}^{Q^*} P_{q^*}}{Q^*}, Q^* < Q. \quad (12)$$

где: \bar{P}^* – средняя по отфильтрованному байесовскому ансамблю гипотез-нейросетей апостериорная вероятность порождения наблюдаемых данных; значком «*» отмечены величины, относящиеся к отфильтрованному ансамблю; P_{q^*} – апостериорная вероятность порождения данных D в соответствии с гипотезой $h_{q^*}(\bar{x}, W, \xi^*)$ при оптимальном значении параметра регуляризации ξ^* согласно (9).

Замечание. От известного байесовского метода обучения нейросети [10] предложенный квазибайесовский метод регуляризации отличается следующими признаками.

I. Снято ограничение на необходимость аналитического задания какого-либо закона распределения шумов в данных.

II. Введена приближенная операция оценки апостериорных вероятностей $P_{q^*} [h_{q^*}(\bar{x}, W, \xi^*) | D] \rightarrow (N_{q^*}^* / N)$.

III. Введена оптимизационная процедура для оценки параметров регуляризации ξ и ω .

IV. Введен алгоритм (9) апостериорной фильтрации байесовских гипотез – нейросетей $\{h_{q^*}(\bar{x}, W, \xi^*) | H\}$ в ансамбле.

V. Отметим также, что контролируемая длина описания нейросетевым байесовским ансамблем данных задается при его формировании в виде априорных гипотез $\{P(h_q | H)\}$, причем структурная оптимизация модели и выбор активационных функций $f(\bullet)$ в промежуточных слоях сети реализуются через механизм апостериорной фильтрации нейросетей $\{h_q\}_{q=1, \overline{Q}}$ по (10).

Направление дальнейших исследований

1. На реальных данных опросных анкет и тестов, направленных на изучение условий и образа жизни обучающейся молодежи, исследовать вопросы оценки адекватности нейросетевой модели, границ уровня зашумленности данных, в рамках которых эффективна предложенная концепция.

2. Использовать нейросетевую модель для комплексной оценки состояния здоровья студентов высших учебных заведений (изучить заболеваемость, показатели антропометрии, функциональных проб и пр.).

5. Выводы

1. На основе парадигмы регуляризации алгоритмов решения обратных задач восстановления многомерных нелинейных зависимостей, скрытых в данных, заключающейся в сужения класса искомых решений с использованием априорной информации об этих решениях, разработана концепция регуляризации нейросетевой модели.

2. Концепция разработана в деталях и реализована в квазибайесовском методе.

3. Предложенный квазибайесовский метод ослабляет остроту проблемы обеспечения адекватности НСМ, т.е. сводит ее к последовательному повышению достоверности и точности результатов моделирования на этапах априорного задания архитектуры нейросетей в байесовском ансамбле апостериорной (после обучения НСМ) фильтрации НСМ в ансамбле — осреднении результатов моделирования на отфильтрованном ансамбле.

Список литературы

1. Айвазян С.А., Мхитарян В.С. Прикладная статистика и основы эконометрики: Учебник. — М.: ЮНИТИ, 1998 – 1023 с.
2. Горбаткова Е.Ю., Мануйлова Г.Р., Ланговой В.Е. Некоторые подходы к оценке результативности профилактической деятельности, направленной на сохранение и укрепление здоровья обучающихся // Современные проблемы науки и образования. – 2015. – № 1.
3. Горбаткова Е.Ю., Мануйлова Г.Р. Социально-педагогические условия повышения стрессоустойчивости участников образовательного процесса / Наркомания: проблемы и пути их решения в свете Стратегии государственной антинаркотической политики // Всероссийская науч.-практ. конф. – Уфа, 2011. – С. 223–225.
4. Горбаткова Е.Ю., Мануйлова Г.Р. Влияние спортивных занятий на показатели заболеваемости школьников / Медико-биологические аспекты физической культуры // Всероссийская науч.-практ. конф. с международным участием. – Казань, 2013. – С. 80–83.
5. Гублер Е.В Информатика в патологии, клинической медицине и педиатрии. — Л.: Медицина, 1990. — 176 с.
6. Осовский С. Нейронные сети для обработки информации: Монография / Пер.с польского. —М.: Финансы и статистика, 2002. — 344 с.
7. Поздеева Т.В. Научное обоснование концепции и организационной модели формирования здоровьесберегающего поведения студенческой молодежи// Дисс. на соиск. степени доктора мед. Наук-специальность 14.00.33 — Общественное здоровье и здравоохранение. – Нижний Новгород: Нижегородская государственная медицинская академия, 2008. — 256 с.
8. Тихонова А.Н., Арсенин В.Я. Методы решения некорректных задач. Монография. –М.: Наука. Физмалит,1979. — 286 с.
9. Нарсеева О.В., Горбаткова Е.Ю. Актуальные проблемы профилактики девиантного поведения подростков / Партнерство социальных институтов воспитания в интересах детства // Международная науч.-практ. конф. – Прага. С. 112–115.
10. Шумский С.А. Байесова регуляризация обучения //Лекции школы-семинара «Современные проблемы нейроинформатики» (23–25января 2002 г., Москва). — М.:МИФИ, 2002, С. 61–94.
11. Хайкин С. Нейронные сети: полный курс: Монография / Пер. с англ., 2-е издание.-М.: Издательский дом «Вильямс», 2006. - 1104 с.

12. Хуснутдинова З.А., Мануйлова Г.Р., Горбаткова Е.Ю., Титова Т.А. Основные вопросы реализации магистерской программы «Профилактика социальных отклонений (превентология)» / Молодой ученый. – №3 (62). – СПб, 2014. – с. 1053-1056.
13. Ярушкина Н.Г. Основы теории нечетких и гибридных систем: Учеб.пособие.-М.: Финансы и статистика, 2004 – 320 с.
14. Fisher R.A. Theory of statistical estimation. Proceedings of the Cambridge Philisophical Society,1925, Vol. 22, pp. 700-725.

Рецензенты:

Лобанов С.А., д.м.н., профессор кафедры охраны здоровья и безопасности жизнедеятельности, ФГБОУ ВПО «Башкирский государственный педагогический университет им. М. Акмуллы, г. Уфа;

Горбушина С.Н., д.п.н., профессор кафедры стандартизации и сертификации Уфимского государственного авиационного технического университета, г. Уфа.