ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ И МАШИННОЕ ОБУЧЕНИЕ

УДК 004.81

МОДЕЛЬ ОЦЕНКИ СТЕПЕНИ УНИКАЛЬНОСТИ И ВОССТАНОВЛЕНИЯ СЛАБО-ОПРЕДЕЛЕННЫХ ДАННЫХ НА ОСНОВЕ МОДИФИКАЦИИ НЕЙРОННОЙ СЕТИ АРТ-2

Гатин Р.Р., Новикова С.В.
Казанский национальный исследовательский технический университет имени А.Н. Туполева – КАИ, г. Казань

Поступила в редакцию 10.02.2024, после переработки 12.03.2024.

В статье рассматривается задача анализа и восстановления данных в малых выборках со слабо изученными взаимосвязями, названными авторами слабо-определенными данными. Предложен метод, основанный на известной нейросетевой модели классификации АРТ-2, способный как производить непосредственно классификацию, так и определять степень уникальности входного вектора по отношению к имеющейся выборке с учетом особенностей слабо-определенных данных. Также разработана модификация предложенного метода, позволяющая восстанавливать пропущенные атрибуты в векторах слабо-определенных данных в случае наличия векторов с полными данными в соответствующем классе. Проведены численные эксперименты для слабо-определенных данных о содержании металлов в крови детей в возрасте от 1 до 14 лет, проживающих на территории г. Казани. Эксперименты продемонстрировали эффективность разработанных методов.

Ключевые слова: редкие данные, слабо изученные взаимосвязи, нейронная сеть АРТ-2, уникальные данные, пропущенные атрибуты, восстановление атрибутов.

Вестник ТвГУ. Серия: Прикладная математика. 2024. № 2. С. 39–59.
https://doi.org/10.26456/vtpmk709

Введение

На практике есть класс задач с очень ограниченным набором экспериментальных данных. Такие данные редки, их получение затруднено физическими, материальными, юридическими, морально-этическими и др. причинами. Например, данные проб грунта на других планетах; персональные данные, особенно относящиеся к несовершеннолетним гражданам; данные медицинских обследований орфанных заболеваний и др. Зачастую построение моделей процессов для подобных данных сопряжено с дополнительными трудностями – неизвестными или слабо

изученными физическими законами, описывающими взаимосвязи в таких системах [1], неполнотой сведений о влияющих внешних факторах [2], недостаточной точностью и периодичностью измерений [3] и пр. Назовем подобные данные слабо-определенными.

В подобных случаях для построения моделей невозможно применить стандартные методы аналитического моделирования. Методы машинного обучения (ML-Machine Learning) также слабо применимы из-за требовательности последних к репрезентативности данных в обучающей выборке. Однако на сегодняшний день существует ряд специальных методов ML, с тем или иным успехом применяемым для обучения моделей на малых объемах данных.

Для методов ML, как правило, расширение выборки считается желательным и полезным для повышения точности и адекватности модели [4]. Однако в случае с редкими или практически уникальными данными добавление в набор новой информации должно быть тщательно взвешено, так как даже единичное измерение, значительно отличающееся от прочих, способно привести к разбалансировке модели. Это может иметь как положительные, так и отрицательные последствия для ее адекватности.

Поэтому при изучении малых выборок слабо-определенных данных при каждом возникновении нового набора следует сначала оценить степень его уникальности по отношению к уже существующим данным. Решение о том, следует ли включать в набор данные, значительно отличающиеся от всех прочих, может приниматься дополнительным экспертным оцениванием, либо при помощи дополнительной процедуры оценки релевантности набора [5].

1. Постановка задачи

Задачей настоящего исследования является разработка метода, способного:

- оценить степень уникальности нового вектора по отношению к имеющейся выборке слабо-определенных данных;
- в случае не-уникальных данных, классифицировать новый вектор, отнеся его к одной из уже известных групп слабо-определенных данных;
- в случае пропущенных атрибутов в векторе, восстановить значения пропущенного атрибута.

В качестве набора слабо-определенных данных выступает датасет, содержащий результаты анализов крови 240 детей в возрасте от 1 до 17 лет, проживающих в г. Казани, на содержание металлов, в частности, цинка. Каждый кортеж датасета, наряду с данными анализов биосубстратов, содержит информацию по физиологическим особенностям (рост, вес) и содержанию металла в питьевой воде, отобранной в местах проживания обследуемых. Примерно 15% имеющихся кортежей имеют пропущенные атрибуты.

2. Методология

Для решения задачи предлагается использовать идею адаптивного резонанса, предложенную Карпентером и Гроссбергом [6]. В основе адаптивно-резонансной

теории (АРТ) лежит внутренний детектор новизны, суть которого заключается в сравнении входного образа с содержимым памяти модели. Резонанс возникает в случае, если входной вектор в «достаточной» степени «похож» на сохраненные в памяти АРТ шаблоны. Если похожего шаблона не найдено, происходит адаптация. Различные варианты моделей АРТ различаются способом вычисления «степени похожести» и критерием, используемым в качестве оценки «достаточности». Как правило, похожесть определяется как расстояние между объектами или группами объектов по какой-либо метрике, а достаточность задается как значение порога разности таких расстояний.

В классической реализации модели АРТ строятся в виде нейронных сетей с двумя слоями нейронов – слоем сравнения, или входным слоем, и слоем распознавания, или слоем эталонов. Сеть решает задачу классификации. Количество нейронов в слое сравнения равно размерности классифицируемых векторов, а число нейронов в слое распознавания равно количеству классов, где каждый нейрон представляет класс [7-8].

Работу сети на основе АРТ можно разделить на три этапа: на первом этапе новый входной вектор сравнивается с нейронами слоя распознавания, из которых выбирается один, называемый нейроном-победителем, наиболее «похожий» на вектор на входе. На втором этапе происходит расчет степени «похожести» входного вектора и нейрона победителя. На третьем этапе полученное значение степени похожести сравнивается с предварительно заданным порогом. Если порог преодолен, то говорят, что «возник резонанс» входного вектора с нейроном-победителем, и вектор относится к классу этого нейрона. В противном случае резонанс не возникает, и входной вектор становится новым классом-эталоном, количество нейронов в слое распознавания увеличивается. Говорят, «сеть адаптируется».

Существует несколько вариантов АРТ-сетей. Первым вариантом архитектуры, реализующей адаптивно-резонансную теорию, считается архитектура АРТ-1, созданная для классификации двоичных образов [9-10]. Следующим шагом стала разработка архитектуры АРТ-2 для обработки непрерывных сигналов [11]. Известны также нейро-нечеткая модификация Fuzzy-ART [12], модификация с учителем ARTMAP [13] и др. [14-15].

Идея настоящего исследования базируется на реализации АРТ-2. В общем виде схему АРТ-2 можно представить следующим образом (Рис. 1).

Условно схему АРТ-2 можно поделить на четыре основных блока: блок предварительной обработки, блок классификации, блок управления, и блок трансформации. Алгоритм функционирования сети можно описать следующим образом.

1. Вектор $X = (x_1, x_2, \ldots, x_m)$ подается на вход сети для последующей классификации. В зависимости от конкретной реализации, затем он поступает в блок **I** предварительной обработки, и может подвергаться нормализации и/или очистке от случайного шума.

2. Результирующий вектор $U = (u_1, u_2, \ldots, u_m)$ поступает в блок классификации **II**, на слой сравнения, где поэлементно умножается на синаптические коэффициенты восходящих связей w_{ij}, активируя тем самым распознающие нейроны $Y = (y_1, y_2, \ldots, y_k)$. Каждый j-тый нейрон слоя распознавания представляет класс, а восходящие веса к j-му нейрону представляют собой средневзвешенные оценки элементов j-го класса, k – количество запомненных в

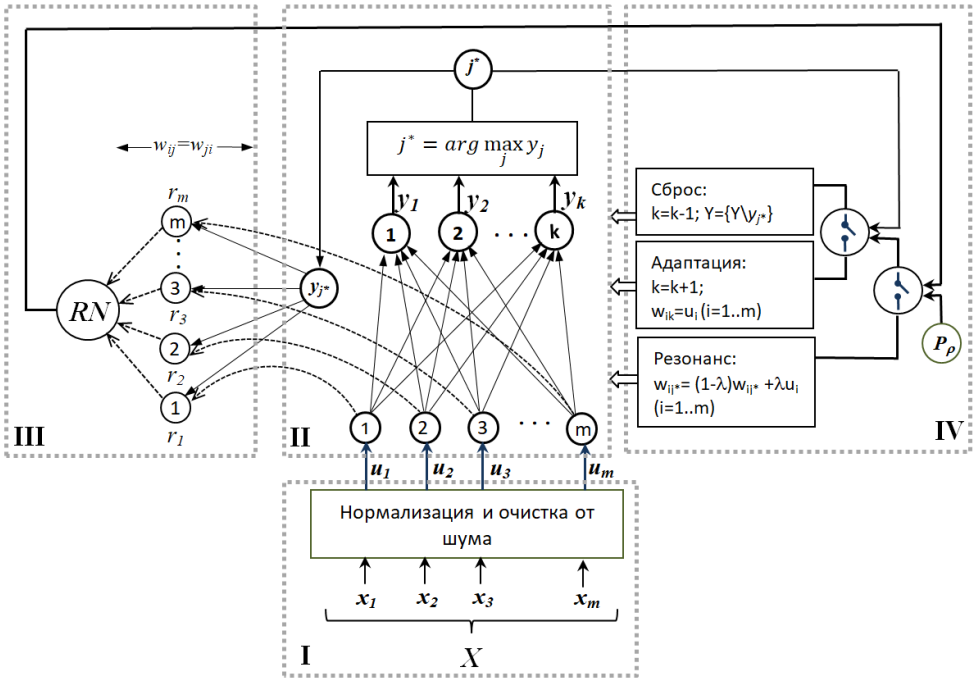

Рис. 1: *Архитектура нейронной сети APT-2 для классификации непрерывных сигналов: I – блок предварительной обработки; II-блок классификации; III – блок управления; IV – блок трансформации*

модели классов. Выходом y_j каждого нейрона слоя распознавания является мера сходства поданного на вход вектора X и средневзвешенного запомненного в j-м нейроне класса, где класс представлен набором восходящих весов: $y_j = \sum_{i=0}^{m} w_{ij}u_i$.

3. Далее выбирается нейрон-победитель (также, как в нейронных сетях Кохонена [16]), по принципу максимально сходства: $j^* = \underset{j}{\operatorname{argmax}}\, y_j$. Ситуацию можно охарактеризовать как «из всех нейронов-классов входной вектор больше всего похож на нейрон-класс j^*».

4. Номер класса-победителя j^* вместе со значениями $U = (u_1, u_2, \ldots, u_m)$ со слоя сравнения подаются в блок управления **III**. Там рассчитывается расстояние между классом-победителем и вектором U с использованием нисходящих связей w_{j^*i}, т.е. в расчетах принимает участие только нейрон (класс)-победитель. Результатом работы блока является величина RN – количественное выражение расстояния между победившим нейроном j и входным вектором X. В классической реализации нисходящие синаптические веса равны восходящим, а степень сходства рассчитывается как Евклидово расстояние между обработанным входным вектором и нейроном-победителем.

5. Полученное значение сравнивается с предварительно заданным порогом резонанса (сходства) P_ρ. Результат сравнения вместе с параметром нейрона-победителя j^* и элементами обработанного входного вектора U передается в блок трансформации **IV**. В данном блоке предусмотрено три типа трансформации:

   (a) **Резонанс.** Возникает, если $RN < P_\rho$. Ситуацию можно охарактеризовать как «входной вектор в достаточной степени похож на нейрон-класс j^*». В этом случае входной вектор считается классифицированным, его класс – j^*. Для внесения информации о новом векторе, относящемся к данному классу, восходящие (а в классической постановке, и нисходящие) веса нейрона j^* корректируются, сдвигаясь в сторону нового элемента класса: $w_{ij^*} = w_{ij^*} = (1-\lambda)w_{ij^*} + \lambda u_i$, $i = \overline{1,m}$. Здесь λ – параметр алгоритма.

   (b) $RN \geq P_\rho$, то есть «входной вектор недостаточно похож на нейрон-класс j*». При этом может выполняться два типа трансформации:

   – Подавление, или сброс, нейрона-победителя j^*. Нейрон временно удаляется из сети вместе со всеми своими восходящими и нисходящими связями, и поиск нейрона-победителя производится заново, среди оставшихся нейронов - возврат к шагу 3. Ситуация «возможно, среди оставшихся нейронов-классов найдется тот, который в достаточной степени похож на входной вектор». Для нового нейрона-победителя вновь рассчитывается мера расстояния RN, и вновь сравнивается с порогом. Цикл повторяется до тех пор, пока либо для какого-то нейрона слоя распознавания не возникнет резонанс, либо будут сброшены все нейроны слоя распознавания.

   – **Адаптация**. Если все нейроны слоя распознавания были последовательно сброшены, а резонанса ни для одного из них так и не наступило, возникает ситуация «входной вектор не похож ни на один ранее известный нейрон-класс». Тогда входной вектор X порождает новый класс, то есть в сети АРТ-2 порождается новый k+1-ый нейрон распознающего слоя, восходящие и нисходящие веса которого устанавливаются равными соответствующим элементам обработанного входного вектора U: $\mathrm{w_{i}}_{k+1} = \mathrm{u_i}$, $\mathrm{i} = \overline{1,m}$.

Два возможных варианта трансформации – резонанс и адаптация – могут рассматриваться как потенциальные детекторы новизны вектора слабо-определенных данных по отношению к имеющейся выборке. В этом случае условие возникновения адаптации можно интерпретировать как наличие новизны во входном векторе, а условие резонанса - как отсутствие новизны.

3. Модификация модели АРТ-2 для задачи оценки уникальности данных

Для модификации классической архитектуры АРТ-2 под решение задачи оценки уникальности слабо-определенных данных необходимо учитывать специфику поставленной задачи, в частности:

1. Так как слабо-определенные данные характеризуются малым количеством доступных наблюдений, логично стремиться сохранить в модели максимальное количество информации о данных, в идеале – модель должна уметь сохранять информацию о каждом предъявленном ей векторе-образе.

2. Из-за слабой изученности и апериодичности сбора слабо-определенных данных, при применении механизма классификации следует ожидать появления большого числа небольших по объему классов с неоднородным количеством содержащихся в каждом классе данных.

3. В случае пропущенных атрибутов в векторах слабо-определенных данных обычные практики, как, например, исключение векторов с пропущенными атрибутами из рассмотрения, неприемлемы. Так как подобные данные встречаются очень редко, и полученная информация, даже неполная, обладает большой ценностью, необходимо разработать специальные приемы для замещения/восполнения пропущенных атрибутов в данных.

Для учета указанных требований, классическая модель APT-2 была преобразована в модифицированную модель, названную нами APT-WD (weakly-defined).

Блок I – предварительная обработка данных. Содержит стандартные процедуры нормализации и очистки от шума. В предлагаемой реализации не рассматривается возможность обработки векторов с пропущенными атрибутами. Авторский способ восстановления пропущенных атрибутов будет описан далее, в отдельном разделе данной статьи.

Блок II – классификация вектора. Блок не подвергся структурным изменениям, однако уточнен способ определения степени соответствия входного вектора имеющимся классам, а также способ вычисления восходящих весов для каждого j-го нейрона-прототипа класса j:

– нейрон-победитель выбирается из условия минимума Евклидова расстояния между вектором U, подаваемым на вход для классификации, и вектором весов восходящих связей к нейрону j по формуле:

$$j^* = \underset{j}{\operatorname{argmin}}\, y_j; \quad y_j = d\left(U, W_j\right) = \sqrt{\sum_{i=1}^{m} \left(u_i - w_{ij}\right)^2}.$$

– восходящие веса w_{ij} рассчитываются как вектор-центроид класса Y_j, содержащего в себе все вектора $X^l = (x_1^l, x_2^l, \ldots, x_m^l)$, относящиеся к классу j: $w_{ij} = \frac{\sum_{l=1}^{S} x_i^l}{S}$; $X^l \in Y_j$, S – количество векторов в классе j.

Связи Блока II, таким образом, несут агрегированную, не детализированную информацию о данных.

Блок III – управление. Структура блока модифицирована следующим образом:

– информация о каждом векторе слабо-определенных данных сохраняется в нисходящих весах нейронов Y_j распознающего слоя. Каждый нейрон Y_j содержит информацию лишь о тех векторах обработанных данных

$U^l = (u_1^l, u_2^l, \ldots, u_m^l)$, $l=\overline{1,S}$, которые относятся к соответствующему классу: $w_{ji}^l = u_i^l$. Здесь S – количество отнесенных ранее к классу j векторов. Таким образом, общее число нисходящих связей нейрона Y_j составляет $(m \times S)$.

– для расчета количественной степени соответствия входного обработанного вектора U и класса $j*$, соответствующего нейрону-победителю распознающего слоя, применяется мера близости «взвешенное попарное среднее». Данная мера позволяет, с одной стороны, учесть единичное влияние каждого распознанного ранее вектора класса $j*$, а с другой – усреднить результат по объему данных, нивелировав тем самым неравномерное распределение векторов по классам. Попарные расстояния определяются в слое нейронов r, нормирование производится в выходном нейроне блока управления RN:

$$r_l = \rho\left(U, U^l\right),\ l=\overline{1,S}; \quad RN = \frac{\sum\limits_{l=1}^{S} r_l}{S}.$$

Графически модифицированный блок управления представлен на Рис. 2.

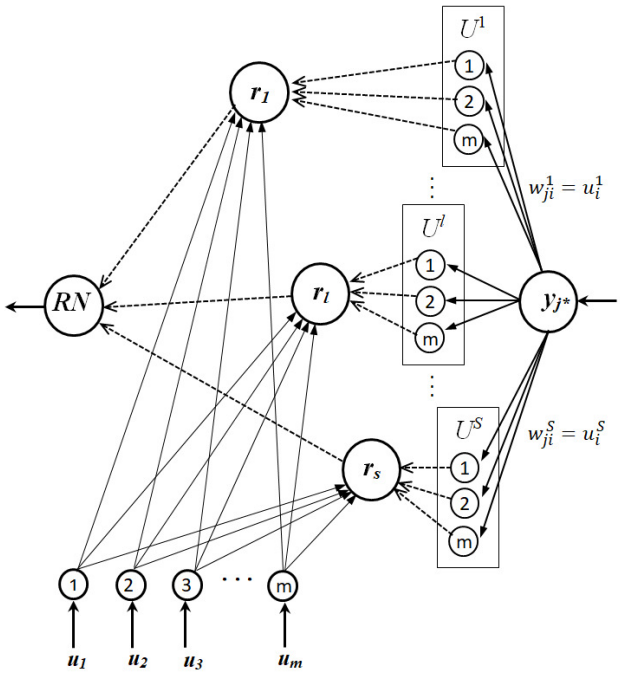

Рис. 2: *Архитектура блока III – управление для модифицированной модели APT-WD*

Блок IV – трансформация модели. В данном блоке модификации подверглись все три процедуры: сброса, адаптации и резонанса, а также в блок добавлен новый нейрон, принимающий логическое решение о новизне поданного на вход вектора:

– **Резонанс**. Восходящие веса w_{ij^*} нейрона-победителя j^* в распознающем слое пересчитываются согласно смещению центра масс соответствующего класса Y_{j^*} после включения в него входного обработанного вектора U: $w_{ij^*} = \frac{Sw_{ij^*}+U_i}{S+1}$, $i = \overline{1,m}$, S – количество элементов класса Y_{j^*} до включения в него вектора U. Нисходящие веса $w_{j^*i}^{S+1} = u_i$, $i = \overline{1,m}$ добавляются к нейрону j^* как результат включения U в Y_{j^*}. Количество элементов класса Y_{j^*} увеличивается на единицу $S = S+1$.

– **Адаптация**. Аналогично классическому методу, в слой распознавания добавляется новый $k+1$-й нейрон Y_{k+1}. Класс Y_{k+1} нейрона содержит единственный вектор U. Добавлены восходящие веса: $w_{ik+1} = u_i$, $i = \overline{1,m}$. Добавлены нисходящие веса: $w_{k+1i}^1 = u_i$, $i = \overline{1,m}, S = 1$. Количество нейронов распознающего слоя $k = k+1$.

– **Сброс** нейрона-победителя в классическом варианте APT-2 производится в случае, если количественная мера близости вектора и соответствующего класса не удовлетворяет условию резонанса, но при этом существуют другие нейроны распознающего слоя, потенциально способные выступить в качестве нейрона-победителя с достаточной степенью соответствия. Из-за большого количества малых классов-кластеров в распознающем слое, для случая слабо-определенных данных процедура сброса требует большого количества вычислений для последовательного перебора всех потенциальных классов-победителей. Для сокращения вычислений в качестве критерия сброса предложено дополнительно использовать свойство мер расстояния между кластерами: расстояние от вектора до центра тяжести кластера не превышает попарного среднего расстояния между вектором и всеми элементами кластера, то есть для любого входного вектора U справедливо $y_{j^*} \leq RN$. Тогда, в случае превышения величиной y_{j^*} заданного порога резонанса P_ρ, условие отсутствия для нейрона $j*$ резонанса $RN \geq P_\rho$ выполнится автоматически, и серия последующих сбросов не приведет к появлению резонирующего нейрона. В качестве модификации, уменьшающей количество вычислений, стандартное условие сброса дополнено условием: $y_{j^*} < P_\rho$, то есть сброс целесообразно выполнять, если нейрон-победитель не резонирует ($RN > P_\rho$), не все распознающие нейроны сброшены ($k>1$), и при этом выход нейрона-победителя меньше порога резонанса ($y_{j^*} < P_\rho$).

– Добавлен логический нейрон принятия решения ***D*** о новизне (уникальности) поданного на вход вектора X. На входы нейрона подаются два логических параметра – результаты условий ($RN < P_\rho$) и ($k>1$). Первое условие соответствует ситуации резонанса, второе – наличию в распознающем слое вакантных несброшенных нейронов, выступающих в роли потенциального источника резонанса. Нейрон содержит единственное логическое правило вида: ЕСЛИ ($RN > P_\rho$)&($k = 1$) ИЛИ ($RN > P_\rho$)&($y_{j^*} < P_\rho$) ТО «Вектор X является уникальным» ИНАЧЕ «Вектор X не является уникальным».

Архитектура блока после модификации представлена на Рис. 3.

Таким образом, после произведенных модификаций, разработанная нейросетевая модель APT- WD отвечает особенностям обработки слабо-определенных дан-

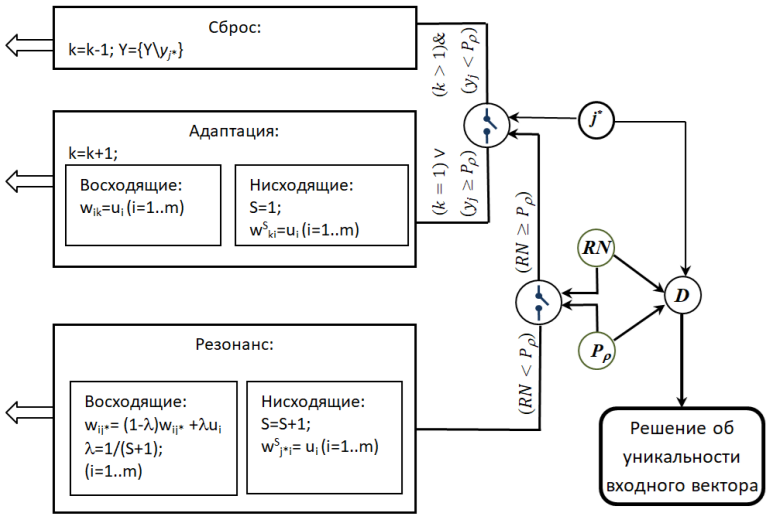

Рис. 3: *Блок IV -трансформация модели для модифицированной модели APT-WD с дополнительным нейроном принятия решения*

ных в части сохранения в ней полной информации о классифицированных векторах с учетом специфики выделяемых классов.

Однако вопрос восстановления пропущенных атрибутов в векторах данных остается актуальным. Для решения этой проблемы разработан метод обработки векторов с пропущенными атрибутами и с последующим их восстановлением.

4. Способ обработки и восстановления векторов данных с пропущенными атрибутами

Идея способа восстановления пропущенного атрибута в векторе данных базируется на замене пропуска средневзвешенным значением соответствующего атрибута среди всех слабо-определенных полных векторов того же класса, что и восстанавливаемый вектор.

Идея реализуема, если класс вектора с пропущенными атрибутами известен, и в этом классе присутствует хотя бы один полный вектор. Таким образом, встает задача первоначальной классификации вектора с пропущенными атрибутами. Для ее решения на основе APT-WD была разработана динамическая модель APT-DWD (Dynamic Weakly-Defined) со специальным блоком предварительной обработки входного вектора, а также с модификацией процедуры резонанса. В случае, если на вход модели APT- DWD подан полный вектор, функционирование модели аналогично функционированию APT-WD.

Блок I: предварительная обработка данных. Включен модуль проверки вектора на наличие пропущенных атрибутов. Для адекватной оценки возможности восстановления информации в пропущенных позициях введен новый параметр модели P_a , названный «порог пригодности». Под данным термином понимается максимальное число недостающих атрибутов в векторе, при котором восстановление

считается возможным. Может выражаться как в абсолютном числе атрибутов, так и в процентах от длины вектора. Конкретное значение может быть получено из данных экспериментов.

Алгоритм функционирования блока включает следующие шаги:

1. На вход сети поступает необработанный вектор $X = (x_1,\ x_2,\ \ldots x_m)$ с q пропущенными атрибутами. Обозначим множество пропущенных атрибутов вектора X как I_X. В случае, если количество пропущенных атрибутов больше порога пригодности $q > P_a$, классификация и восстановление считаются невозможными, вектор отклоняется как непригодный. Если порог пригодности не преодолен, происходит редукция вектора X путем удаления позиций неопределенных атрибутов. Размерность вектора сокращается: $X^-_{(1\times z)} = \left\{X_{(1\times m)} \backslash x_g : g \in I_X\right\},\ z = m - q.$
2. На основе нейросетевой модели архитектуры APT-WD динамически создается временная редуцированная упрощенная модель APT-WD$^{(-)}$:

Блок II: распознавание. Нейроны слоя U, соответствующие номерам пропущенных в X атрибутов, удаляются из сети: $U^{(-)}_{(1\times z)} = \left\{U_{(1\times m)} \backslash u_g : g \in I_X\right\},\ z = m - q$ вместе с восходящими к распознающему слою связями $w_{gj};\ g \in I_X,\ j = \overline{1,k}$.

Блок III: управление. Для всех распознающих нейронов Y_j удаляются нисходящие синаптические связи $w^l_{jg},\ l{=}\overline{1,S}$, соответствующие номерам пропущенных атрибутов $g \in I_X$.

На вход редуцированной модели подается вектор X^- и обрабатывается способом, описанным в предыдущем разделе, до расчета меры близости RN включительно. В блок трансформации передается найденное значение RN, а также номер j^* нейрона-победителя $Y^{(-)}_{j^*}$ распознающего слоя $Y^{(-)}$ временной редуцированной модели.

Блок IV: трансформация.

- **Сброс.** Изменений не предусмотрено.
- **Адаптация.** В модели отсутствует процедура адаптации, так как создание нового класса из входного вектора, а также вывод об его новизне возможен лишь для полных векторов данных. В случае отсутствующих атрибутов вектор, выделенный как потенциально уникальный, признается непригодным для анализа.
- **Резонанс.** В модели отсутствует процедура резонанса. В случае возникновения условий, приводящих к резонансу в модели APT-WD, в модели APT-DWD запускается процедура восстановления вектора. В данном случае вектор X^- определен как принадлежащий классу j^*. Производится восстановление пропущенных атрибутов вектора X на основе значений данных атрибутов в полных векторах класса j^*:

$$x_g = \frac{\sum\limits_{l=1}^{S} x^l_g}{S},$$

здесь S – количество элементов класса j^*, x^l_g – g-тые атрибуты полных векторов X^l, принадлежащих классу j^*, $g \in I_X$.

1. После восстановления временная редуцированная модель уничтожается. Восстановленный вектор X резонирует с нейроном j^* полной модели обычным образом, описанном в предыдущем разделе.

Структура разработанной модели APT-DWD с динамической и статической частями представлена на Рис. 4.

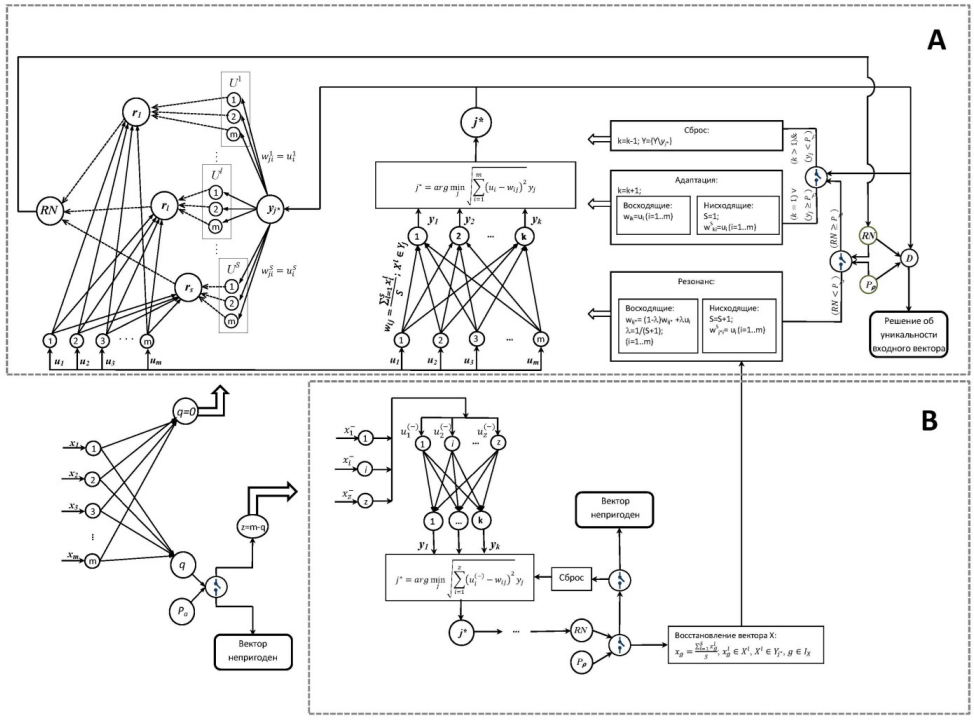

Рис. 4: *Архитектура динамической нейросетевой модели APT-DWD: A – статическая часть модели; B- динамическая часть модели*

5. Вычислительные эксперименты

5.1. Данные для экспериментов

Работоспособность предложенной модели была опробована на задаче классификации и частичного восстановления информации о содержании в крови металла (цинк) 240 детей в возрасте от 1 до 14 лет, обследованных в период 2021-2022 года в лаборатории Института проблем экологии и недропользования академии наук Республики Татарстан, г. Казань, (Россия) [17]. Одновременно с анализом крови испытуемого, поводились инструментальные исследования на содержания металла в питьевой воде в квартире пациента. В качестве физиологических особенностей выступают данные о весе и росте обследуемого ребенка. Детское население было

выбрано в качестве целевой группы из-за отсутствия сопутствующих путей привнесения металла в организм, таких как употребление в пищу неспецифических продуктов, вредные условия производства, вредные привычки и т.д. Каждому обследуемому присвоен идентификационный номер (ID), все данные обезличены. Классифицирующим признаком является район проживания, ассоциированный с точкой водозабора для питьевого водоснабжения.

Таким образом, каждый вектор данных содержит 4 признака для классификации: вес (кг), рост (см), уровень металла в питьевой воде (мг/мл), уровень металла в крови (мг/мл). Класс кодируется номером района проживания. Фрагмент используемого для экспериментов набора данных приведен в Таблице 1.

Таблица 1: *Набор данных обследования детского населения на содержание цинка в крови (фрагмент)*

ID	Вес	Рост	Zn в крови	Zn в воде	N района
1	33,6	143	0,7042	1,7	4
2	33,5	143	0,6245	1,65	4
3	42	166	1,9003	4,4	5
4	34	142	0,475	1,6	4
5	42	156	0,83317	1,82222	1

Для математического описания данных приняты следующие обозначения: x_1 – Вес; x_2 – Рост; x_3 – Zn в крови; x_4 – Zn в воде.

5.2. План проведения экспериментов

Вычислительные эксперименты проводились по следующему сценарию:

1. Из генеральной выборки случайным образом выделяется подвыборка для экспериментов заданного объема.
2. 85% процентов экспериментальной выборки считается обучающим набором, 15% - тестовым набором. Выбор тестовых векторов данных осуществляется случайным образом.
3. Строится модель APT-WD с применением обучающего набора.
4. Анализируется точность модели APT-WD на обучающем наборе.

5.3. Тестирование модели APT-WD на способность оценки степени уникальности вектора

1. В тестовый набор атрибутов добавляется 30% искусственно сгенерированных векторов, заведомо не относящихся ни к одному из имеющихся классов. В результате получаем расширенный тестовый набор.

2. Элементы расширенного тестового набора подаются поочередно на вход обученной модели APT-WD, анализируется точность модели на тестовом наборе.

5.4. Тестирование модели APT-DWD на точность классификации и восстановления пропущенных атрибутов

1. Из векторов тестового набора данных удаляется заданный процент атрибутов.
2. На основе обученной модели APT-WD строится динамическая модель APT-DWD.
3. Модель APT-DWD применяется для классификации и восстановления векторов тестового набора.
4. Исследуется точность модели APT-DWD на тестовом наборе.

5.5. Способ оценки точности модели

Так как предложенная модель выполняет не только задачу классификации, но и восстановления пропущенных атрибутов, точность оценивалась по двум критериям:

Критерий 1: Процент правильной классификации (тип оценки – accuracy)

Критерий 2: Точность восстановления атрибутов (тип оценки - среднеквадратическое отклонение)

Точность модели APT-WD оценивалась только по Критерию 1. Точность модели APT-DWD оценивалась как конъюнкция двух критериев с применением порога отклонения для Критерия 2: пример считается правильно распознанным, если он отнесен моделью к правильному классу И квадрат отклонения рассчитанного значения пропущенного атрибута от реального не превосходит заданный порог. Допустимый порог отклонения принят равным 0,5 (20%).

5.6. Эксперименты на сверхмалой подвыборке

Условия эксперимента:

– Объем экспериментальной подвыборки -50

– Объем обучающего множества – 43

– Объем тестового множества – 7

– Количество классов в экспериментальной подвыборке – 5

– Процент пропущенных атрибутов -25% (один из четырех)

– Пропущенный (скрытый) атрибут – «Цинк в питьевой воде» - x_4.

Результаты эксперимента.

– Точность обучения модели APT-WD -100%

– Точность модели APT-WD по оценке уникальности входного вектора – 100%

– Тестовая точность классификации модели APT-DWD -86%

– Тестовая точность восстановления атрибута модели APT-DWD - 90 %.

– Совокупная средняя тестовая точность модели APT-DWD - 71,5%.

Подробно результаты эксперимента представлены в Таблице 2. Графически результаты представлены на Рис. 5.

Таблица 2: *Результаты классификации векторов с пропущенным атрибутом и расчетов значения атрибута*

ID	x_1	x_2	x_3	x_4 (скрытый атрибут)	Класс (проверочный)	Класс (рассчитанный)	x_4 (рассчитанный атрибут)	Квадр. Откл.	Ошибка
9	45	156	0,58	2,2	2	2	1,7775	0,42	19%
11	43	156	0,652	1,7	2	2	1,7775	0,08	5%
7	49	163	0,5738	3,4	1	1	2,8932	0,51	15%
46	35	141	0,637	2,29286	0	0	1,7175	0,58	25%
2	33,5833	143	0,6245	1,65	3	0	1,7175	0,07	4%
14	42	156	0,804	1,8	2	2	1,7775	0,02	1%
5	42	156	0,83317	1,82222	2	2	1,7772	0,05	2%
							Средн.	0,25	10%

Аналогичные результаты на сверхмалой подвыборке получены и для остальных трех атрибутов.

Также была проведена серия экспериментов на малой (100 строк), средней (150 строк), крупной (200 строк) и полной (240 строк) подвыборках. Результаты точности восстановления каждого из четырех атрибутов (в процентах) отражает Таблица 3.

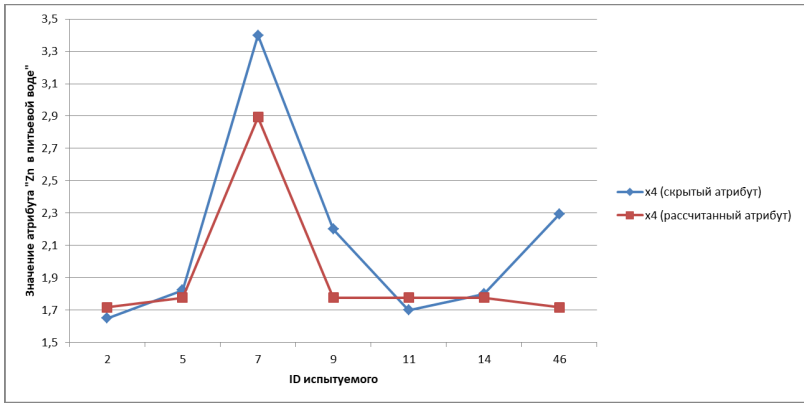

Рис. 5: *Сравнение реальных и восстановленных значений атрибута «Цинк в питьевой воде»*

Таблица 3: *Влияние объема выборки на точность расчета пропущенного атрибута*

Объем выборки Восст. атрибут	50	100	150	200	240
Цинк в Крови	100	100	90,9	83,3	81,1
Цинк в воде	100	93,3	90,9	90	83
Вес	100	93,3	90,9	100	100
Рост	100	100	90,9	100	100

Видно, что для наиболее вариабельных атрибутов кортежей слабо-определенных данных, таких как уровень металла в питьевой воде и крови, точность восстановления снижается с ростом объема выборки, и составляет менее 90%. Данный факт подтверждает гипотезу об ограничении использования модели на малых и сверхмалых выборках

Заключение

Разработано две модификации классической нейросетевой модели для классификации АРТ-2, способные распознавать, классифицировать и оценивать степень уникальности слабо-определенных данных (статическая модель APT-WD), а также восстанавливать пропущенные атрибуты в векторах данных (динамическая модель APT-DWD). Под слабо-определенными данными в данном случае понимаются малые выборки экспериментальных данных с неизвестными аналитическими взаимосвязями, характеризующиеся низкой частотой и апериодичностью измерений, с пропусками отдельных параметров.

Основными отличиями предложенных модификаций являются:

– В сети APT-2 веса восходящих и нисходящих связей каждого распознающего $Y-$нейрона одинаковы, и сохраняют единственный вектор, полученный в результате усреднения свойств всех обучающих векторов, отнесенных к его классу. Индивидуальные признаки, присущие отдельным векторам данных, в памяти сети не хранятся. В модификациях APT-WD и APT-DWD хранится полная информация о каждом векторе данных в нисходящих весах распознающих $Y-$нейронов. Восходящие веса по-прежнему представляют усредненные значения векторов соответствующего класса в виде его центра тяжести.

– Процедура сброса нейрона-победителя в модификациях APT-WD и APT-DWD требует меньшего количества вычислений по сравнению с APT-2 за счет сокращения перебора потенциальных нейронов-кандидатов на резонанс. Сокращение достигается введением дополнительного условия-сравнения меры близости нейрона-победителя с входным вектором по двум критериям: близости центра тяжести, и близости попарного среднего.

– Модификации APT-WD и APT-DWD позволяют производить как классификацию данных по принципу работы APT-2, так и делать вывод о степени уникальности входного вектора за счет включения в блок трансформации дополнительного узла принятия решений. При этом адаптация сети под новый вектор данных говорит об его уникальности, а резонанс – об отсутствии уникальности.

– Модификация APT-DWD дополнительно способна восстанавливать пропущенные атрибуты во входных данных за счет специфической процедуры классификации неполного вектора и расчета его недостающих атрибутов на основе известных векторов распознанного класса. Классификация неполного вектора производится в динамически порождаемой временной упрощенной сети APT-WD с редуцированными слоями сравнения и распознавания, а также с отсутствующими блоками адаптации и редукции. Для осуществления адаптации и редукции восстановленный вектор передается в соответствующие участки статической сети APT-WD. Там же происходит принятие решения о степени уникальности входного неполного вектора.

– Модель APT-DWD для неполного вектора способна, кроме состояний «вектор уникален» и «вектор не уникален», определять третье состояние «вектор не пригоден» за счет дополнительных условий на этапе обработки данных и этапе трансформации и принятия решений. На этапе обработки данных вектор признается непригодным, если количество пропущенных атрибутов больше заданного порога. На этапе трансформации вектор признается непригодным, если он не резонирует ни с одним существующим классом, и восстановление его атрибутов невозможно.

Список литературы

[1] Евдокимов Р.А. Обзор докладов московских международных симпозиумов по исследованиям солнечной системы 2019 и 2020 гг. (ЮМ-S3 И ИМ-S3). Часть 2. Луна и Меркурий, Венера // Космическая техника и технологии. 2022. № 2 (37).

[2] Кравчук Ж.П., Румянцева О.А. Орфанные заболевания: определение, проблемы, перспективы // Проблемы здоровья и экологии. 2013. № 4 (38). С. 7–11.

[3] Magalhaes J.P., Wang J. The fog of genetics: what is known, unknown and unknowable in the genetics of complex traits and diseases. EMBO rep (2019) 20: e48054. 2019. http://doi.org/10.15252/embr.201948054

[4] Su H., Zhu X., Gong S. Deep Learning Logo Detection with Data Expansion by Synthesising Contex. arXiv:1612.09322.

[5] Ильин В.А., Кирюшов Н.П. Метод проверки тренажерных моделей на адекватность // Программные продукты и системы. 2021. № 1. С. 61–66.

[6] Carpenter G.A., Grossberg S. Adaptive Resonance Theory // Encyclopedia of Machine Learning and Data Mining. Eds. by Sammut C., Webb G. . Boston, MA: Springer. Pp. 1–17. https://doi.org/10.1007/978-1-4899-7502-7_6-1

[7] Андреев А.А., Шатилова Е.В. Реализация искусственной нейронной сети на основе адаптивной резонансной теории // Вестник российских университетов. Математика. 2008. № 1. С. 78–80.

[8] Bartfai G. An Adaptive Resonance Theory-based Neural Network for Autonomous Learning via Iterative Knowledge Redescription // 2021 International Joint Conference on Neural Networks (IJCNN) (Shenzhen, China). 2021. Pp. 1–8. https://doi.org/10.1109/IJCNN52387.2021.9533351

[9] Дмитриенко В.Д., Поворознюк О.А. Новые алгоритмы обучения одно и многомодульных дискретных нейронных сетей АРТ // Вестник Национального технического университета Харьковский политехнический институт. Серия: Информатика и моделирование. 2008. № 24. С. 51–64.

[10] Georgiopoulos M., Heileman G.L., Huang J. The N-N-N conjecture in ART1 // Neural Networks. 1992. Vol. 5, № 5. Pp. 745–753. https://doi.org/10.1016/S0893-6080(05)80135-2

[11] Karthikeyan B., Gopal S., Venkatesh S. ART 2 - an unsupervised neural network for PD pattern recognition and classification // Expert Systems with Applications. 2006. Vol. 31, № 2. Pp. 345–350. https://doi.org/10.1016/j.eswa.2005.09.029

[12] Carpenter G.A., Grossberg S., Rosen D.B. Fuzzy ART: an adaptive resonance algorithm for rapid, stable classification of analog patterns // IJCNN-91-Seattle International Joint Conference on Neural Networks. 1991. Pp. 411–416. https://doi.org/10.1109/IJCNN.1991.155368

[13] Каширина И.Л., Федутинов К.А. Построение решающих правил с помощью нейронной сети ARTMAP // Моделирование, оптимизация и информационные технологии. 2019. Т. 7, № 3. ID 634. https://doi.org/10.26102/2310-6018/2019.26.3.029

[14] Буханов Д.Г., Поляков В.М. Сеть адаптивно-резонансной теории с многоуровневой памятью // Экономика. Информатика. 2018. № 4. С. 709–717.

[15] Леонов С.Ю. К-значная нейронная сеть арт для анализа работоспособности вычислительных устройств // Вестник Национального технического университета Харьковский политехнический институт. Серия: Информатика и моделирование. 2013. № 39 (1012). С. 115–128.

[16] Кохонен Т. Самоорганизующиеся карты. М.: БИНОМ. Лаборатория знаний, 2013. 660 с.

[17] Tunakova Yu., Novikova S., Shagidullin A., Valiev V., Novikova K. Neural network assessment of metal retention in the body of adolescent children, depending on the air and water-food routes of intake // AIP Conference Proceedings. Vol. 2948. 2023. ID 020019. https://doi.org/10.1063/5.0166276

Образец цитирования

Гатин Р.Р., Новикова С.В. Модель оценки степени уникальности и восстановления слабо-определенных данных на основе модификации нейронной сети АРТ-2 // Вестник ТвГУ. Серия: Прикладная математика. 2024. № 2. С. 39–59. https://doi.org/10.26456/vtpmk709

Сведения об авторах

1. **Гатин Руслан Ришатович**
аспирант Казанского национального исследовательского технического университета им А.Н. Туполева – КАИ.

Россия, 420015, г. Казань, ул. Большая Красная, д. 55, КНИТУ-КАИ.
E-mail: RRGatin@kai.ru

2. **Новикова Светлана Владимировна**
профессор кафедры Прикладной математики и информатики Казанского национального исследовательского технического университета им А.Н. Туполева – КАИ.

Россия, 420015, г. Казань, ул. Большая Красная, д. 55, КНИТУ-КАИ.
E-mail: SVNovikova@kai.ru

MODEL FOR UNIQUENESS ASSESSING DEGREE AND FOR RESTORATION OF WEAKLY DEFINED DATA BASED ON ART-2 NEURAL NETWORK MODIFICATION

Gatin R.R., Novikova S.V.
Kazan National Research Technical University – KAI, Kazan

Received 10.02.2024, revised 12.03.2024.

The article examines the problem of analyzing and recovering data in small samples with poorly studied relationships, called weakly defined data, by the authors. A method is proposed based on the well-known neural network classification model ART-2, capable of both direct classification and determining the degree of uniqueness of the input vector about the existing sample, taking into account the characteristics of weakly defined data. A modification of the proposed method has also been developed that makes it possible to restore missing attributes in vectors of weakly defined data in the case of the presence of vectors with complete data in the corresponding class. Numerical experiments were carried out for weakly defined data on the content of metals in the blood of children aged 1 to 14 years living in Kazan. Experiments demonstrated the effectiveness of the developed methods.

Keywords: rare data, poorly studied relationships, ART-2 neural network, unique data, missing attributes, attribute restoration.

Citation

Gatin R.R., Novikova S.V., "Model for uniqueness assessing degree and for restoration of weakly defined data based on ART-2 neural network modification", *Vestnik TvGU. Seriya: Prikladnaya Matematika [Herald of Tver State University. Series: Applied Mathematics]*, 2024, № 2, 39–59 (in Russian). https://doi.org/10.26456/vtpmk709

References

[1] Evdokimov R.A., "Review of the reports of the Moscow International Symposia on Solar System Research in 2019 and 2020 (YUM-S3 and IM-S3). Part 2. The Moon and Mercury, Venus", *Kosmicheskaya tekhnika i tekhnologii [Space engineering and technology]*, 2022, № 2 (37) (in Russian), 141 pp.

[2] Kravchuk Zh.P., Rumyantseva O.A., "Orphan diseases: definition, problems, prospects", *Problemy zdorovya i ekologii [Health and environmental issues]*, 2013, № 4 (38), 7–11 (in Russian).

[3] Magalhaes J.P., Wang J., *The fog of genetics: what is known, unknown and unknowable in the genetics of complex traits and diseases*, EMBO rep (2019) 20: e48054, 2019, http://doi.org/10.15252/embr.201948054.

[4] Su H., Zhu X., Gong S., *Deep Learning Logo Detection with Data Expansion by Synthesising Contex*, arXiv:1612.09322.

[5] Ilin V.A., Kiryushov N.P., "The method of checking the fitness models for adequacy", *Software and Systems*, 2021, № 1, 61–66 (in Russian).

[6] Carpenter G.A., Grossberg S., "Adaptive Resonance Theory", *Encyclopedia of Machine Learning and Data Mining*, eds. Sammut C., Webb G., Springer, Boston, MA, 1–17, https://doi.org/10.1007/978-1-4899-7502-7_6-1.

[7] Andreev A.A., Shatilova E.V., "Implementation of an artificial neural network based on adaptive resonance theory", *Vestnik rossijskikh universitetov. Matematika [Russian Universities Reports. Mathematics]*, 2008, № 1, 78–80 (in Russian).

[8] Bartfai G., "An Adaptive Resonance Theory-based Neural Network for Autonomous Learning via Iterative Knowledge Redescription", *2021 International Joint Conference on Neural Networks (IJCNN)* (Shenzhen, China), 2021, 1–8, https://doi.org/10.1109/IJCNN52387.2021.9533351.

[9] Dmitrienko V.D., Povoroznyuk O.A., "New learning algorithms for single and multi-module discrete neural networks ART", *Vestnik Natsionalnogo tekhnicheskogo universiteta Kharkovskij politekhnicheskij institut. Seriya: Informatika i modelirovanie [Bulletin of the National Technical University Kharkiv Polytechnic Institute. Series: Computer Science and Modeling]*, 2008, № 24, 51–64 (in Russian).

[10] Georgiopoulos M., Heileman G.L., Huang J., "The N-N-N conjecture in ART1", *Neural Networks*, **5**:5 (1992), 745–753, https://doi.org/10.1016/S0893-6080(05)80135-2.

[11] Karthikeyan B., Gopal S., Venkatesh S., "ART 2 - an unsupervised neural network for PD pattern recognition and classification", *Expert Systems with Applications*, **31**:2 (2006), 345–350, https://doi.org/10.1016/j.eswa.2005.09.029.

[12] Carpenter G.A., Grossberg S., Rosen D.B., "Fuzzy ART: an adaptive resonance algorithm for rapid, stable classification of analog patterns", *IJCNN-91-Seattle International Joint Conference on Neural Networks*, 1991, 411–416, https://doi.org/10.1109/IJCNN.1991.155368.

[13] Kashirina I.L., Fedutinov K.A., "Building decision rules using the ARTMAP neural network", *Modelirovanie, optimizatsiya i informatsionnye tekhnologii [Modeling, optimization and information technology]*, **7**:3 (2019), 634 (in Russian), https://doi.org/10.26102/2310-6018/2019.26.3.029.

[14] Bukhanov D.G., Polyakov V.M., "Adaptive Resonance Theory network with multilevel memory", *Ekonomika. Informatika [Economy. Computer science]*, 2018, № 4, 709–717 (in Russian).

[15] Leonov S.Yu., "K-digit neural network art for analyzing the performance of computing devices", *Vestnik Natsionalnogo tekhnicheskogo universiteta Kharkovskij*

politekhnicheskij institut. Seriya: Informatika i modelirovanie [Bulletin of the National Technical University Kharkiv Polytechnic Institute. Series: Computer Science and Modeling], 2013, № 39 (1012), 115–128 (in Russian).

[16] Kokhonen T., *Samoorganizuyushchiesya karty [Self-organizing maps]*, BINOM. Laboratoriya znanij, Moscow, 2013 (in Russian), 660 pp.

[17] Tunakova Yu., Novikova S., Shagidullin A., Valiev V., Novikova K., "Neural network assessment of metal retention in the body of adolescent children, depending on the air and water-food routes of intake", *AIP Conference Proceedings*. V. 2948, 2023, https://doi.org/10.1063/5.0166276.

Author Info

1. **Gatin Ruslan Rishatovich**
PhD student, Kazan National Research Technical University named after A.N. Tupolev – KAI.

Russia, 420015, Kazan, 55 Bolshaya Krasnaya str., KNRTU-KAI.
E-mail: RRGatin@kai.ru

2. **Novikova Svetlana Vladimirovna**
Professor at the Department of Applied Mathematics and Computer Science, Kazan National Research Technical University named after A.N. Tupolev – KAI.

Russia, 420015, Kazan, 55 Bolshaya Krasnaya str., KNRTU-KAI.
E-mail: SVNovikova@kai.ru