Глава 7.

Скрытые параметры и транспонированная регрессия

А.Н.Кирдин, А.Ю.Новоходько, В. Г.Царегородцев

Вычислительный центр СО РАН в г. Красноярске[1]

Решается классическая проблема восстановления недостающих данных в следующей постановке: найти для каждого объекта наилучшую формулу, выражающую его признаки через признаки других объектов (которых должно быть по возможности меньше). Эта формула должна быть инвариантна относительно смены шкал измерения. Инвариантность достигается тем, что решение представляется в виде суперпозиции однородных дробно - линейных функций.

Строится отношение "объект - опорная группа объектов". Опорная группа выделена тем, что по признакам ее элементов наилучшим образом восстанавливаются признаки исходного объекта. Решение дается с помощью нейронной сети специальной архитектуры. Предлагается способ минимизации опорной группы, использующий преимущества нейросетевого подхода.

Метод транспонированной регрессии применяется к задаче интерполяции свойств химических элементов. Исследуется точность интерполяции потенциалов ионизации химических элементов при помощи транспонированной линейной регрессии. Достигнутая точность позволяет предсказать отсутствующие в справочной литературе значения высших (с 5-го по 10-й) потенциалов ионизации для элементов с атомными номерами от 59-го до 77-го и рекомендовать метод для интерполяции иных физических и химических свойств элементов и соединений.

1. Гипотеза о скрытых параметрах

Пусть задано некоторое множество объектов и совокупность («номенклатура») признаков этих объектов. Для каждого объекта может быть

[1] 660036, Красноярск-36, ВЦК СО РАН. E-mail: amse@cc.krascience.rssi.ru

определен вектор значений его признаков - полностью или частично. Если эти значения для каких-либо объектов определены не полностью, то возникает классическая проблема восстановления пробелов в таблицах данных [1].

Наиболее распространенный путь ее решения - построение регрессионных зависимостей. Предполагается, что одни свойства каждого из объектов могут быть с достаточной степенью точности описаны как функции других свойств. **Эти функции одинаковы для различных объектов**. Последнее предположение выполняется далеко не всегда.

Что делать, если не удается построить регрессионной зависимости, общей для всех объектов? В этом случае естественно предположить, что существуют неописанные и неизмеренные свойства объектов - и именно в них и заключаются скрытые различия, не дающие построить искомые зависимости. Эти неучтенные и неизмеренные свойства; от которых зависят наблюдаемые параметры, называют «скрытыми параметрами», а предположение о том, что все дело в них - **гипотезой о скрытых параметрах.**

Проблема скрытых параметров стала знаменитой, благодаря квантовой механике. Многолетние попытки свести квантовые неопределенности к различию в значениях скрытых параметров и поиск этих самых параметров не увенчались успехом. В этом случае проблема отсутствия однозначных связей между характеристиками объектов оказалась глубже, а квантовые неопределенности признаны несводимыми к различию в значениях неизмеренных, но в принципе доступных измерению величин - для квантовых объектов микромира скрытых параметров не нашли.

За пределами миров квантовой механики различия между объектами всегда объяснимы наличием скрытых параметров. В нашем обычном макроскопическом мире проблема состоит не в существовании скрытых параметров, а в эффективной процедуре их поиска и учета, а также в разделении ситуаций на те, для которых разумно искать скрытые параметры, и те, для которых больше

подходит представления о неустранимых (в данном контексте) случайных различиях.

Одна из простейших форм предположения о скрытых параметрах - **гипотеза о качественной неоднородности выборки**. Она означает, что скрытые параметры принимают сравнительно небольшое конечное число значений и всю выборку можно разбить на классы, внутри которых скрытые параметры, существенные для решения интересующей нас задачи регрессии, постоянны. Каждой такой выборке будет соответствовать «хорошая» регрессионная зависимость.

Построить классификацию (без учителя), соответствующую данной гипотезе можно только на основе предположении о форме искомой регрессионной зависимости наблюдаемых параметров от наблюдаемых же параметров внутри классов (**задача о мозаичной регрессии**). Если предполагается линейная зависимость, то эта задача классификации решается методом динамических ядер, только место точек - центров тяжести классов (как в сетях Кохонена) - занимают линейные многообразия, каждое из которых соответствует линейному регрессионному закону своего класса [2].

Регрессионные зависимости, которые строятся с помощью нейронных сетей, также образуют вполне определенный класс и для них тоже возможна соответствующая классификация без учителя. Изящный способов решения проблемы скрытых параметров для нейросетевых уравнений регрессии реализован в пакете «MultiNeuron» [2,3]. Достаточно большая нейронная сеть может освоить любую непротиворечивую обучающую выборку, однако, как показывает опыт, если малая нейронная сеть не может обучиться, то из этого можно извлечь полезную информацию. Если не удается построить удовлетворительную регрессионную зависимость при заданном (небольшом) числе нейронов и фиксированной характеристике («крутизне» функции активации) каждого нейрона, то из обучающей выборки исключаются наиболее сложные примеры до тех пор, пока сеть не обучится. Так получается класс,

который предположительно соответствует одному значению скрытых параметров. Далее обучение можно продолжить на отброшенных примерах и т.д.

Пример. В одном из проводимых исследований [3] нейросеть обучали ставить диагноз вторичного иммунодефицита (недостаточности иммунной системы) по иммунологическим и метаболическим параметрам лимфоцитов. В реальной ситуации по сдвигам таких параметров иногда бывает трудно сделать верное заключение (и это хорошо известная в иммунологии проблема соотношения клинической картины и биохимических проявлений иммунодефицитов). Были обследованы здоровые и больные люди, параметры которых использовались для обучения. Однако нейросеть не обучалась, причем хорошо распознавала все до единого примеры здоровых людей, а часть примеров больных путала со здоровыми. Тогда был сделан следующий шаг: каждый раз, когда сеть останавливала работу, из обучающей выборки убирался пример, на данный момент самый трудный для распознавания, и после этого вновь запускался процесс обучения. Постепенно из обучающей выборки были исключена примерно одна треть больных (при этом ни одного здорового!), и только тогда сеть обучилась полностью. Так как ни один здоровый человек не был исключен из обучения, группа здоровых не изменилась, а группа больных оказалась разделена на 2 подгруппы - оставшиеся и исключенные примеры больных. После проведения статистического анализа выяснилось, что группа здоровых и исходная группа больных практически не отличаются друг от друга по показателям метаболизма лимфоцитов. Однако получившиеся 2 подгруппы больных статистически достоверно отличаются от здоровых людей и друг от друга по нескольким показателям внутриклеточного метаболизма лимфоцитов. Причем в одной подгруппе наблюдалось увеличение активности большинства лимфоцитарных ферментов по сравнению со здоровыми, а в другой подгруппе - депрессия (снижение активности).

В научном фольклоре проблема скрытых параметров описывается как задача отделения комаров от мух: на столе сидят вперемежку комары и мухи, требуется провести разделяющую поверхность, отделяющую комаров от мух.

Данные здесь - место на плоскости, скрытый параметр - видовая принадлежность, и он через данные не выражается.

2. Теорема о скрытых параметрах

Ряд алгоритмов решения проблемы скрытых параметров можно построить на основе следующей теоремы. Пусть n - число свойств, N - количество объектов, $\{x^i\}_{i=1}^N$ - множество векторов значений признаков. Скажем, что в данной группе объектов выполняется уравнения регрессии ранга r, если все векторы $\{x^i\}_{i=1}^N$ принадлежат n-r-мерному линейному многообразию. Как правило, в реальных задачах выполняется условие $N>n$. Если же $n \geq N$, то векторы $\{x^i\}_{i=1}^N$ принадлежат N-1-мерному линейному многообразию и нетривиальные регрессионные связи возникают лишь при ранге $r>n-N+1$. Ранг регрессии r измеряет, сколько независимых линейных связей допускают исследуемые свойства объектов. Число r является коразмерностью того линейного подпространства в пространстве векторов признаков, которому принадлежат наблюдаемы векторы признаков объектов. Разумеется, при обработке реальных экспериментальных данных необходимо всюду добавлять «с заданной точностью», однако пока будем вести речь о точных связях.

Следующая теорема о скрытых параметрах позволяет превращать вопрос о связях между различными свойствами одного объекта (одной и той же для разных объектов) в вопрос о связи между одним и тем же свойством различных объектов (одинаковой связи для различных свойств) - транспонировать задачу регрессии. При этом вопрос о качественной неоднородности выборки «транспонируется» в задачу поиска для каждого объекта такой группы объектов (опорной группы), через свойства которых различные свойства данного объекта выражаются одинаково и наилучшим образом.

Теорема. Пусть для некоторого $r>0$ существует такое разбиение $\{x^i\}_{i=1}^N$ на группы $\{x^i\}_{i=1}^N = \bigcup_{j=1}^k Y_j$, что $r>n-N_j+1$ (где N_j - число элементов в Y_j), и для

каждого класса Y_j выполняются уравнения регрессии ранга r. Тогда для каждого объекта x^i из $\{x^i\}_{i=1}^N$ найдется такое множество W_i (опорная группа объекта x^i) из k объектов, что $n\text{-}r+1 \geq k$ и для некоторого набора коэффициентов λ_y

$$x^i = \sum_{y \in W_i} \lambda_y y, \ \sum_{y \in W_i} \lambda_y = 1. \qquad (1)$$

Последнее означает, что значение *каждого* признака объекта x^i является линейной функцией от значений этого признака для объектов опорной группы. Эта линейная функция *одна и та же* для всех признаков.

Линейная зависимость (1) отличается тем, что она инвариантна к изменениям единиц измерения свойств и сдвигам начала отсчета. Действительно, пусть координаты всех векторов признаков подвергнуты неоднородным линейным преобразованиям: $x_j \mapsto a_j x_j + b_j$, где j - номер координаты. Нетрудно убедиться, что при этом линейная связь (1) сохранится. Инвариантность относительно преобразования масштаба обеспечивается линейностью и однородностью связи, а инвариантность относительно сдвига начала отсчета - еще и тем, что сумма коэффициентов λ_y равна 1.

Сформулированная теорема позволяет переходить от обычной задачи регрессии (поиска зависимостей значения признака от значений других признаков того же объекта) к транспонированной задаче регрессии - поиску линейной зависимости признаков объекта от признаков других объектов и отысканию опорных групп, для которых эта зависимость является наилучшей.

Доказательство основано на том, что на каждом k-мерном линейном многообразии для любого набора из q точек $y_1, y_2, ..., y_q$ при $q>k+1$ выполнено соотношение

$$\sum_{j=1}^{q} \lambda_j y_j = 0 \text{ для некоторого набора } \lambda_j, \ \sum_{j=1}^{q} \lambda_j = 0 \text{ и некоторые } \lambda_j \neq 0.$$

С математической точки зрения теорема о скрытых параметрах представляет собой вариант утверждения о равенстве ранга матрицы, вычисляемого по строкам, рангу, вычисляемому по столбцам.

3. Транспонированная задача линейной регрессии

Изложение в этом разделе следует работам [2,5,6]. Постановка обычной задачи регрессии (или мозаичной регрессии) исходит из гипотезы о том, что одни характеристики объектов могут быть функциями других и эти функции одни и те же для всех объектов (или соответственно классов объектов).

Транспонируем таблицу данных (поменяем местами слова "объект" и "признак"). Рассмотрим гипотезу от том, что значения признака одного объекта могут быть функциями значений того же признака других объектов и эти функции одни и те же для всех признаков (или классов признаков). Получаем формально те же задачи регрессии (транспонированные задачи регрессии). Есть, однако, два содержательных отличия транспонированных задач от исходных:

1) инвариантность к смене шкал измерения - кажется маловероятным, чтобы существенные связи между признаками различных объектов зависели от шкалы измерения, поэтому необходимо, чтобы уравнения транспонированной регрессии были инвариантны относительно смены шкалы измерения любого признака (обычно - линейного неоднородного преобразования $x'=ax+b$ однородная часть которого описывает смену единицы измерения, а свободный член - сдвиг начала отсчета);

2) в традиционных задачах регрессии предполагается, что объектов достаточно много (N), по сравнению с числом признаков n, иначе (при $N<n$) точные линейные соотношения возникнут просто из-за малого числа объектов, так как через N точек всегда можно провести линейное многообразие размерности N-1. В противовес этому "транспонированное" предположение о достаточно большом числе признаков ($n>N$) кажется нереалистичным.

Требование инвариантности к смене шкал приводит к специальным ограничениям на вид функций регрессии, а недостаточность количества признаков (в сравнении с числом объектов) для построения транспонированной регрессии вынуждает нас для каждого объекта искать небольшую группу, по свойствам которых можно восстановить характеристики данного.

Задача построения таких групп объектов была чрезвычайно популярна в химии перед открытием Менделеевым периодического закона (1871 г.). С 1817 г. (Деберейнер) были опубликованы десятки работ на эту тему [7]. Именно они поставили исходный материал для систематизации элементов. Деберейнер обнаружил триады, в которых свойства среднего элемента могут быть оценены как средние значения этих свойств для крайних членов триады. Его труды продолжили Гмелин, Гладстон, Дюма и другие. Вот некоторые из таких триад: K-Na-Li, Ba-Sr-Ca, Cl-Br-J, S-Se-Te, P-As-Sb, W-V-Mo, ...

Один из наиболее полных списков триад был опубликован Ленсеном (1857). Он же заметил, что для большей точности иногда полезно брать "эннеады" - девятки, составленные из трех триад.

Менделеев писал: "...между всеми... учеными, которые раньше меня занимались сравнением величин атомных весов элементов, я считаю, что обязан преимущественно двум: Ленсену и Дюма. Я изучил их исследования и они меня побудили искать действительный закон" (цит. по [7], с. 220-222).

Более общим образом задача ставится так: найти для каждого объекта наилучшую линейную формулу, выражающую его вектор признаков через векторы признаков других объектов (которых должно быть по возможности меньше). Эта формула должна быть инвариантна относительно смены шкал.

Итак, требуется построить отношение, связывающее объекты с группами объектов, по которым для него строятся интерполяционные формулы. Проделав эту работу "в лоб" (по базам данных и без обращения к интуиции химиков) для большого числа элементов (объектов) и потенциалов ионизации (признаков), мы получили хорошее согласие с экспериментом и предсказали ряд неизвестных ранее высших потенциалов ионизации. Результаты будут описаны в следующем разделе.

Предположим, что некоторый большой набор свойств - внешних, эмпирических данных об объекте (явление) является сюръекцией небольшого набора внутренних, теоретических переменных (сущности). Эта идея позволяет сделать предположение о том, что размер опорной группы объектов, по которой

наилучшим образом восстанавливаются свойства данного объекта, не только не должен превосходить размер набора свойств (иначе заведомо возникнут точные линейные соотношения), но и быть малым настолько, насколько это позволяет заданная точность [2-5].

Если предположить, что для некоторого множества объектов зависимость между теоретическим и эмпирическим линейна, и векторы теоретических параметров объектов данного множества лежат в линейном многообразии размерности q, то размер опорной группы не будет превосходить q+1.

Другое условие, налагаемое на искомую формулу, требует инвариантности к смене шкал измерений. Разумно считать, что глубинные связи не зависят от единиц, в которых выражены значения свойств объектов:

$$f(a\mathbf{y}^1+b,\ldots,a\mathbf{y}^q+b) = a\,f(\mathbf{y}^1,\ldots,\mathbf{y}^q)+b$$

Если в качестве искомой формулы рассматривать линейную комбинацию векторов опорной группы, то требуемой инвариантности можно достичь, наложив некоторое условие на коэффициенты разложения. Таковым условием является равенство суммы коэффициентов единице:

$$\tilde{\mathbf{y}} = \sum_i \alpha_i \mathbf{y}^i, \quad \sum_i \alpha_i = 1.$$

Для нелинейной регрессии естественно использовать однородные рациональные функции [2].

Рассматривались два вида решения. Первый:

$$\tilde{\mathbf{y}} = \mathbf{m}_\mathbf{y} + \sum_{i=1}^{q} \beta_i \left(\mathbf{y}^i - \mathbf{m}_\mathbf{y}\right),\ \alpha_i = \beta_i + \frac{1}{q} - \frac{1}{q}\sum_{k=1}^{q}\beta_k \qquad (2)$$

где $\tilde{\mathbf{y}}$ - восстановленный вектор свойств, $\mathbf{y}^i$ - вектор свойств i-го объекта опорной группы, q - мощность опорной группы, $\mathbf{m}_\mathbf{y} = \frac{1}{q}\sum_{i=1}^{q}\mathbf{y}^i$, - среднее значение

Во втором случае в качестве $\mathbf{m}_\mathbf{y}$ выбирался один из векторов опорной группы.

$$\tilde{\mathbf{y}} = \mathbf{y}^t + \sum_{i=1}^{q} \beta_i \left(\mathbf{y}^i - \mathbf{y}^t\right),\ \alpha_i = \beta_i\ ,\ \alpha_t = 1 - \sum_{\substack{k=1 \\ k\neq t}}^{q}\beta_k \qquad (3)$$

Заметим, что легко построить нейронную сеть, вычисляющую такие формулы [5,6].

Из-за предположения о малости опорной группы объектов в качестве одного из путей решения предлагается перебор всех наборов заданного размера. Было предложено искать минимум одного из двух критериев:

а) $\|\mathbf{y}-\tilde{\mathbf{y}}\|^2+\varepsilon^2\|\boldsymbol{\alpha}\|^2\to\min$, б) $\|\mathbf{y}-\tilde{\mathbf{y}}\|+\varepsilon\|\boldsymbol{\alpha}\|\to\min$.

В случае а) точное решение находится из системы линейных уравнений. Введем обозначения:

Y - матрица векторов опорной группы, *n* строк, *q* столбцов. *n* - число известных компонент восстанавливаемого вектора **y**.

$\hat{\mathbf{Y}}=(\mathbf{y}^i-\mathbf{m}_\mathbf{y})$ - матрица **Y** в которой из каждого столбца вычтен вектор $\mathbf{m}_\mathbf{y}$ ($\mathbf{y}^t$ в случае 2).

M - матрица, все элементы которой равны 1,

m - вектор, все компоненты которого равны 1,

E - единичная матрица,

$\boldsymbol{\alpha},\boldsymbol{\beta}$ - вектора размерностью *q*.

Для выражения (2)

$$\tilde{\mathbf{y}}=\mathbf{m}_\mathbf{y}+\hat{\mathbf{Y}}\boldsymbol{\beta},\ \boldsymbol{\alpha}=\frac{1}{q}\mathbf{m}+\left(\mathbf{E}-\frac{1}{q}\mathbf{M}\right)\boldsymbol{\beta}.$$

Дифференцируя выражение а) и приравнивая нулю, получаем:

$$\left(\hat{\mathbf{Y}}^\mathsf{T}\hat{\mathbf{Y}}+\varepsilon\left(\mathbf{E}-\frac{1}{q}\mathbf{M}\right)\right)\boldsymbol{\beta}=\hat{\mathbf{Y}}^\mathsf{T}\left(\mathbf{y}-\mathbf{m}_\mathbf{y}\right).$$

Для выражения (3),

$\mathbf{e}^t$ - вектор, *t*-ая компонента которого равна 1, остальные 0.

$\mathbf{L}_t=(\mathbf{e}^t)$ - матрица, столбцы которой равны вектору $\mathbf{e}^t$.

$$\mathbf{L}_t^\mathsf{T}\mathbf{L}_t=\mathbf{M},\ \mathbf{L}_t^\mathsf{T}\mathbf{e}^t=\mathbf{m}$$

Имеем

$$\tilde{\mathbf{y}}=\mathbf{y}^t+\hat{\mathbf{Y}}\boldsymbol{\beta},\ \boldsymbol{\alpha}=\mathbf{e}^t+\left(\mathbf{E}-\mathbf{L}_t\right)\boldsymbol{\beta}$$

$$\left(\hat{\mathbf{Y}}^\mathsf{T}\hat{\mathbf{Y}}+\varepsilon(\mathbf{E}+\mathbf{M})\right)\boldsymbol{\beta}=\hat{\mathbf{Y}}^\mathsf{T}\left(\mathbf{y}-\mathbf{y}^t\right)+\varepsilon\mathbf{m}$$

Система уравнений решается для известных значений компонент вектора **y**, полученное решение используется для предсказания неизвестных значений.

В случае критерия б) в качестве начального приближения для каждого испытуемого набора рассматривались β минимизирующие невязку $\Delta=\|\mathbf{y}-\tilde{\mathbf{y}}\|$. Минимум критерия находился BFGS-методом [8].

Нами рассмотрен вариант нахождения оптимальной опорной группы фиксированного размера в задаче транспонированной линейной регрессии, когда оптимальная опорная группа отбиралась в ходе полного перебора всех возможных опорных групп. Другой предложенный вариант (оптимизационный) предполагает первоначальное задание избыточного числа объектов в опорной группе и последующее сокращение ее размера в результате отбрасывания наименее значимых параметров.

Программная реализация и переборного, и оптимизационного вариантов решения транспонированной задачи линейной регрессии выполнялась в среде MS DOS с использованием транслятора Borland C++. Текст программы соответствует ANSI–стандарту языка C++, что делает возможным перенос программы на другие аппаратные платформы (что и делалось большие базы медицинских данных обрабатывалась на компьютере Alpha Station корпорации DEC). При этом зависимые от операционной системы фрагменты программы подключаются при помощи условных директив препроцессора языка. Так, для обеспечения работы с большими файлами данных в среде MS DOS используется обращение к интерфейсу DPMI (предоставляется DPMI–расширителями и операционными системами OS/2, Windows 3.xx, Windows 95, Windows NT) для переключения в защищенный режим и обхода ограничения в 640K памяти.

Программа позволяет пользователю определять файл данных, обрабатываемые строки (объекты) и столбцы (свойства объектов), выбирать между вариантами решения и видами функции критерия, задавать значения иных параметров метода. Для обработки порядковых признаков возможна спецификация некоторых столбцов, как содержащих значения не из непрерывного, а из дискретного множества значений. Прогнозные значения

отсутствующих данных в этом случае будут приводиться к ближайшему значению из дискретного множества значений.

Результатом работы программы является файл отчета. Для каждого обрабатываемого объекта (строки базы данных) в файле отчета содержится информация об оптимальным образом приближающей объект опорной группе (номера объектов, входящих в опорную группу, и коэффициенты разложения), значение функции критерия, ошибки интерполяции известных свойств объекта и прогнозные значения для неизвестных свойств. В конце файла отчета выводятся максимальные и средние ошибки аппроксимации известных данных для всех обрабатываемых столбцов базы данных (свойств объектов).

Тестирование предлагаемого метода проводилось на модельных данных. При построении модельных данных задаются размерность теоретической проекции (число скрытых переменных), размерность эмпирической проекции (число свойств объекта) , число различных классов, вектор среднего и разброса для генерируемых данных в каждом классе. Для каждого класса случайным образом порождается линейный оператор, отображающий пространство скрытых переменных в пространство свойств объектов. Для каждого объекта случайным образом выбираются значения скрытых переменных и рассчитываются значения свойств. Тестирование проводилось в скользящем режиме по всему задачнику. Полученные результаты (Табл.1) позволяют заключить, что предложенный метод весьма эффективен, критерий вида б) с большей эффективностью определяет опорную группу при избыточном и недостаточном наборах объектов (лучше, чем МНК а)), а решение вида (2) дает лучшие по сравнению с (3) результаты при избыточном наборе объектов.

Таблица 1.

Качество восстановления по модельным данным с теоретической размерностью 3

ε	критерий	вид	средняя относительная ошибка, % при размере опорной группы			
			3	4	5	18
0.01	а	1	5	0	15	66
	а	2	5	0	15	66
	б	1	5	0	13	40
	б	2	5	0	13	66
0.1	а	1	10	16	30	72
	а	2	10	16	30	72
	б	1	6	10	14	40
	б	2	6	10	14	66

При решении задачи заполнения пробелов в таблицах данных для любой таблицы общей рекомендацией является проведение серии пробных прогнозов для определения оптимального сочетания параметров.

4. Интерполяция свойств химических элементов

Идея интерполяции свойств элементов возникла в химии еще до создания периодической системы [7]. В триадах Деберейнера (1817г.) характеристики среднего элемента триады находились как средние арифметические значений характеристик крайних элементов. Были попытки работать с тетрадами, “эннеадами” (составленными из трех триад) и т.п. Периодическая таблица Менделеева позволяет по–разному определять группу ближайших соседей для интерполяции: от двух вертикальных соседей по ряду таблицы до окружения из восьми элементов (два из того же ряда и по три из соседних рядов). Однако интерполяция свойств путем взятия среднего арифметического по ближайшим элементам таблицы не всегда (не для всех свойств и элементов) дает приемлемые результаты – требуется либо иной выбор соседей, либо другая процедура интерполяции.

Более общим образом задачу интерполяции можно поставить так: найти для каждого элемента наилучшую формулу, выражающую его вектор свойств через векторы свойств других элементов. Эту задачу и решает метод транспонированной регрессии.

В работах [9,10] исследовался полуэмпирический метод, близкий по идее к методу транспонированной регрессии. Единственное и главное отличие заключалось в том, что среди параметров сразу фиксировался набор «теоретических» и строились зависимости остальных свойств от них (в частности, зависимости потенциалов ионизации от атомного номера).

Используем метод транспонированной линейной регрессии для интерполяции и прогноза высших потенциалов ионизации (ПИ). Напомним, что *n*–й потенциал ионизации А – энергия, которую необходимо затратить, чтобы оторвать *n*–й электрон от иона $A^{(n-1)+}$ (n–1 раз ионизированного атома А). Зависимость ПИ от атомного номера (рис.1) нелинейна и сложна.

Следуя формальному смыслу, *n*–й ПИ атома А следует относить все к тому же атому. Однако структура энергетических уровней иона определяется зарядом ядра и числом электронов. Для атома оба этих числа совпадают с атомным номером, но для ионов уже различны. Как и в работах [9,10], *n*–й потенциал ионизации атома с атомным номером *m* будем искать как функцию от *m*–*n*+1. Объектами будут служить, строго говоря, не атомы с атомным номером *m*, а *m*-электронные системы. Таким образом, второй ПИ гелия (атомный номер 2), третий ПИ лития (атомный номер 3) и т.д. относятся к одноэлектронной системе при различных зарядах ядра. Осуществляется привязка потенциала ионизации уже ионизированного атома не к этому же атому, а к *m*-электронной системе с *m*, равным имеющемуся числу электронов в ионе.

Рассмотрим результаты пробного прогноза высших потенциалов ионизации. Приведем результаты, полученные при использовании в функции критерия нормы в виде суммы абсолютных значений компонент вектора и значения ε=0.1, поскольку такое сочетание при тестировании показало себя наилучшим образом. Для того, чтобы невязки по каждому свойству равномерно входили в левую часть функции критерия, выполнялось нормирование каждого свойства (приведение к нулевому математическому ожиданию и единичному среднеквадратическому уклонению).

На рис.2 показаны ошибки прогноза ПИ (с 3-го по 10-й) при разных размерах опорных групп (2, 3 и 4 элемента в опорной группе). При этом для каждого ПИ опорные группы строились по предыдущим ПИ. Величины максимальной и средней ошибок показаны в процентах от диапазона изменения величин соответствующего ПИ. На основе приведенных графиков можно рекомендовать использование как можно большего набора однородных свойств для достижения оптимального прогноза.

Для попытки прогноза отсутствующих в справочной литературе [11,12] значений высших ПИ (с 5-го по 10-й ПИ для элементов с атомными номерами от 59-го до 77-го) изучим влияние размера опорной группы на точность прогноза при построении опорной группы по первым четырем ПИ (Рис.3). Удовлетворительная точность достигается при трех и четырех элементах в опорной группе.

Дальнейшее увеличение числа элементов в опорной группе себя не оправдывает. Увеличению точности прогноза мешают и погрешности при экспериментальном определении ПИ, особенно высших. В таблице 1 приводится прогноз отсутствующих значений ПИ.

Литература

1. Загоруйко Н.Г., Ёлкина В.Н., Тимеркаев В.С. Алгоритм заполнения пропусков в эмпирических таблицах (алгоритм “ZET”) // Вычислительные системы. — Новосибирск, 1975. — Вып. 61. Эмпирическое предсказание и распознавание образов. — С. 3-27.
2. Горбань А.Н., Россиев Д.А. Нейронные сети на персональном компьютере. Новосибирск: Наука (Сиб. отделение), 1996. 276 с.
3. Rossiev D.A., Savchenko A.A., Borisov A.G., Kochenov D.A. The employment of neural-network classifier for diagnostics of different phases of immunodeficiency // Modelling, Measurement & Control.- 1994.- V.42.- N.2. P.55-63.

4. Горбань А.Н. Проблема скрытых параметров и задачи транспонированной регрессии // Нейроинформатика и ее приложения. Тезисы докладов V Всероссийского семинара. Красноярск: изд. КГТУ, 1997.
5. Горбань А.Н., Новоходько А.Ю., Царегородцев В.Г. Нейросетевая реализация транспонированной задачи линейной регрессии // Нероинформатика и ее приложения. Тезисы докладов IV Всероссийского семинара, 5-7 октября 1996 г. Красноярск: изд. КГТУ, 1996. С. 37-39.
6. A.N. Gorban and A.Yu.Novokhodko. Neural Networks In Transposed Regression Problem, Proc. of the World Congress on Neural Networks, Sept. 15-18, 1996, San Diego, CA, Lawrence Erlbaum Associates, 1996, pp. 515-522.
7. Становление химии как науки. Всеобщая история химии / Под ред. Ю.И.Соловьева. — М.: Наука, 1983. — 464с.
8. Гилл Ф., Мюррей У., Райт М. Практическая оптимизация, — М.: Мир, 1985. — 509с.
9. Горбань А.Н., Миркес Е.М., Свитин А.П. Полуэмпирический метод классификации атомов и интерполяции их свойств // Математическое моделирование в биологии и химии. Новые подходы. — Новосибирск: Наука. Сиб. отделение, 1992. — с.204–220.
10. Горбань А.Н., Миркес Е.М., Свитин А.П. Метод мультиплетных покрытий и его использование для предсказания свойств атомов и молекул // Журнал физической химии. — 1992. — Т.66, №6. — с.1503–1510.
11. Физико-химические свойства элементов. — Киев: Наукова думка. 1965 — 808с.
12. Свойства элементов. В 2-х частях. Ч.1. Физические свойства. Справочник. — М.: Металлургия, 1976. - 600с.

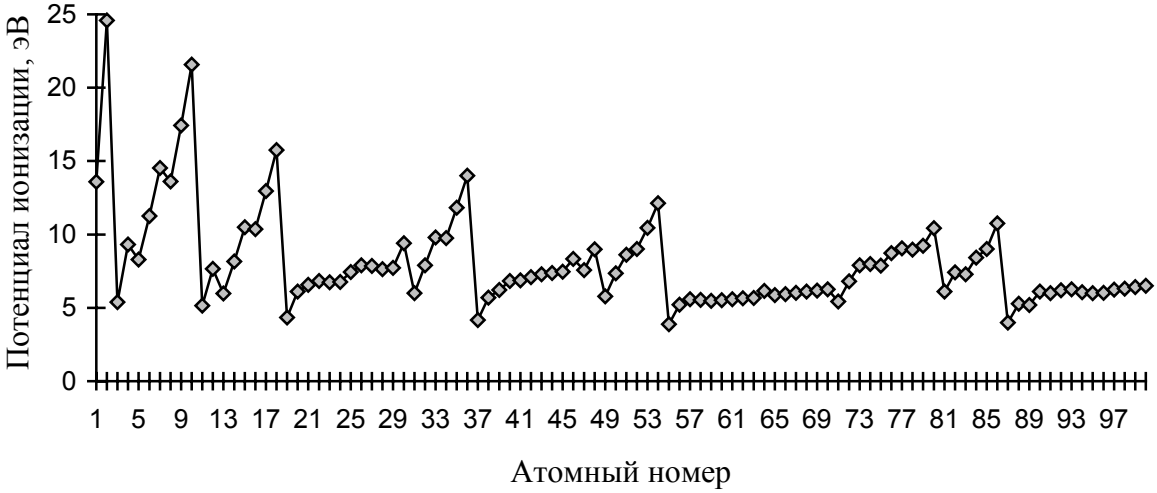

Рис. 1. Зависимость 1-го ПИ от атомного номера

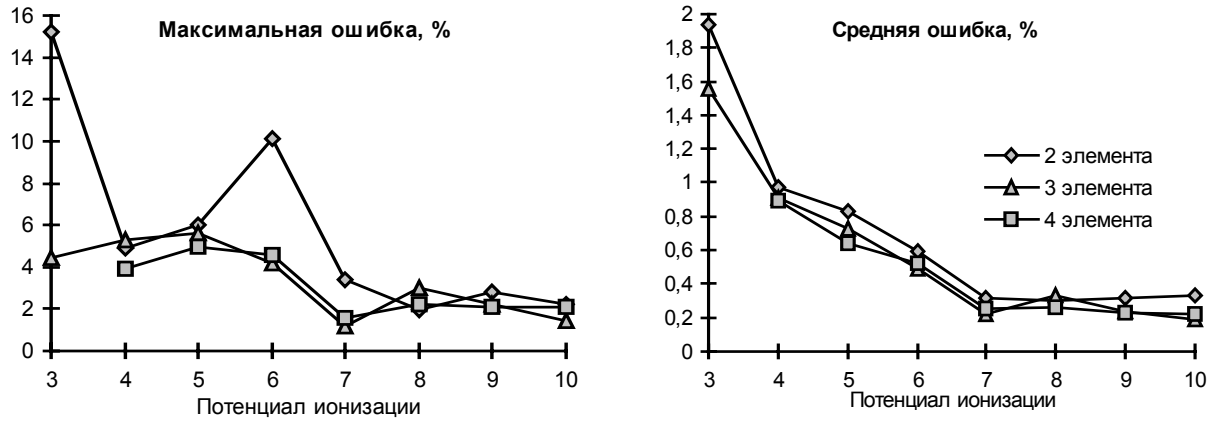

Рис. 2. Зависимость ошибки прогноза 3-10 ПИ от числа элементов в опорной группе. Опорные группы и регрессионные зависимости для каждого ПИ строились по предыдущим ПИ

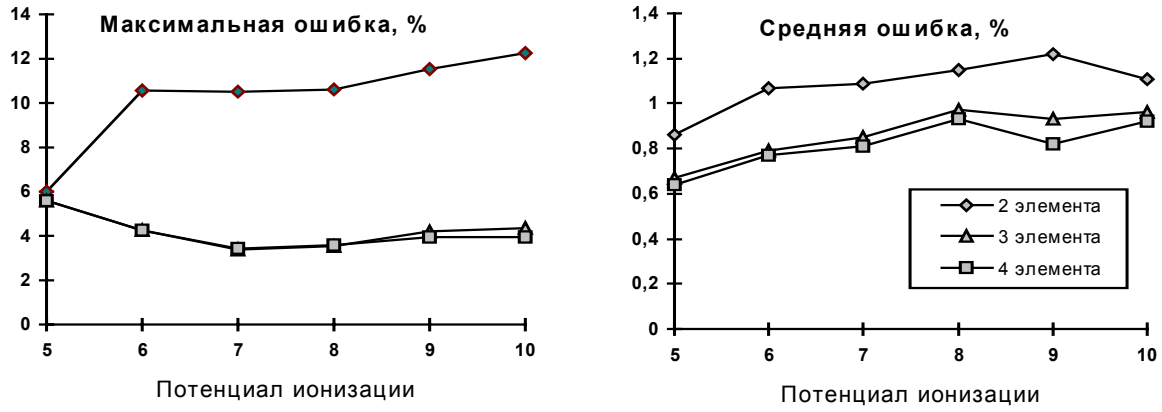

Рис. 3. Зависимость ошибок прогноза 5-10 ПИ от числа элементов в опорной группе. Опорные группы и регрессионные зависимости строились по первым четырем ПИ

Таблица 2.

Прогноз высших потенциалов ионизации отдельных химических элементов

Атомный номер	Элемент	5-й ПИ	6-й ПИ	7-й ПИ	8-й ПИ	9-й ПИ	10-й ПИ
59	Pr	50,7					
60	Nd	49,2	69,6				
61	Pm	53,6	67,7	97,1			
62	Sm	55,9	72,9	87,9	123,7		
63	Eu	56,3	76,3	93,9	110,8	153,6	
64	Gd	61,9	77,2	98,4	117,7	135,9	186,9
65	Tb	67,4	84,9	99,8	123,8	142,9	163,8
66	Dy	48,3	92,2	110,2	125,9	151,5	171,3
67	Ho	52,6	65,8	119,5	138,0	154,5	181,6
68	Er	54,5	72,1	84,6	149,7	169,1	185,7
69	Tm	54,9	74,5	93,4	106,1	182,8	203,5
70	Yb	52,4	74,8	96,1	117,3	128,6	219,9
71	Lu	57,8	71,7	96,2	120,6	143,9	154,0
72	Hf	63,1	79,2	92,2	121,0	147,8	173,2
73	Ta		85,8	102,4	115,2	147,4	177,1
74	W			110,5	128,2	140,7	176,6
75	Re				139,0	156,9	168,7
76	Os					170,1	188,6
77	Ir						203,7