

УДК 504.064.36; 004.8.032.26

НЕЙРОСЕТЕВОЙ КЛАССИФИКАТОР ДАННЫХ ЭКОЛОГИЧЕСКОГО МОНИТОРИНГА И ИХ РЕГРЕССИОННЫЙ АНАЛИЗ

А.Л. Ляхов, С.П. Алёшин

Полтавский национальный технический университет имени Юрия Кондратюка

e-mail: aleshsp@ukr.net

Введение

Для эффективного анализа влияния техногенной нагрузки на окружающую среду в регионе необходимо знать внутреннюю структуру экологических факторов, их влияние на показатели качества и динамику их изменений. Для этого целесообразно найти аналитическую зависимость входных экологических факторов с индикаторами экологии в регионе. Принятие решений по коррекции экологии зависит от значения детерминирующей силы каждого фактора в отдельности и его вклада при совместном многофакторном воздействии всего массива факторов на целевую функцию. Если построить информационно-аналитическую модель анализа данных экологического мониторинга с оценкой этих показателей, то при прочих равных условиях, полученные знания могут существенно повлиять на доминантные индикаторы техногенной нагрузки в регионе [1,2,3]. Техногенное давление на окружающую среду и человека можно характеризовать рядом показателей экологического мониторинга, а индикатором этого воздействия рассматривать, например, статистику заболеваний среди жителей этого региона [3]. Поэтому построение аналитических моделей техногенной нагрузки в регионе для поддержки принятия решений является актуальной задачей как с научной, так и с практической точки зрения.

1. Анализ состояния вопроса и направление исследований

Публикации в этой области исследований, несмотря на многофакторность, разную силу воздействия на среду и случайный характер факторов в функции техногенной нагрузки, свидетельствуют о принципиальной возможности решения подобной задачи. Имеет перспективы направление исследований на основе интеллектуальных технологий извлечения знаний из массива данных по результатам мониторинга экологической нагрузки [3]. Основу этого направления составляет применение интеллектуальных технологий, построенных на принципах обучения и реализованных в формате пакетов технического анализа [3,4,5,6]. Анализ публикаций по теме позволяет сделать вывод о целесообразности постановки и решении задачи моделирования процесса оценки и управления состоянием экологии в регионе с целью минимизации техногенной нагрузки. Необходимо построить адекватную математическую модель, связывающую данные экологического мониторинга почвы, воды, воздуха с индикаторами состояния здоровья жителей региона и создать продуктивную систему принятия решений. Это особенно важно при ограниченных ресурсах и необходимости их продуктивного точечного использования с максимальной эффективностью. Однако в условиях большой размерности вектора входных данных (экологические показатели) и сложности взаимосвязей с индикаторами техногенной нагрузки (статистика заболеваемости) построить многомерную регрессию для всего массива данных с хорошей производительностью, как правило, не удастся [3,5,7]. Поэтому логично предположить, что нахождению многофакторной зависимости, как решению задачи многомерной

регресии, должна предшествовать процедура структурирования исходных данных с целью формирования однородных подмножеств. Тогда значительно легче добиться заданных границ производительности при построении модели регрессионного анализа для каждого из полученных подмножеств. Для построения модели анализа структуры входных данных предлагается использовать среду пакет технического анализа Statsoft и обеспечить адаптацию эмуляторов базовых функций к задаче регрессионного анализа техногенной нагрузки региона в рамках установленных ограничений.

2. Постановка задачи и метод ее решения

По результатам экологического мониторинга (массив показателей (Таблица 2)) необходимо построить математическую модель внутренней структуры факторов техногенной нагрузки, которая по заданному входному вектору наблюдений выдавала бы отклик с заданной точностью:

$$F: X \rightarrow Y_{opt}, X \subset \mathbb{R}^m, Y_{opt} \subset \mathbb{R}, \quad (1)$$

X — множество выборок признаков описания экологического состояния;

Y_{opt} — верифицированное выходное значение функции регрессии.

В нейросетевом формате эта задача может быть представлена в виде [5]:

$$y^k(x^m) = \sum_{j=1}^H v_j (w_{j1}x_1 + w_{j2}x_2 + \dots + w_{jm}x_m + u_j) \quad (2)$$

где y^k - вектор состояний; k -номер индикатора состояний (дыхание, кровь, % обращений к врачу и т.д.); (x^m) - вектор факторов; m - размерность входного вектора данных – (NO, SO₂, CO и т.д.); H – мощность обучающей выборки; v_j - параметры нейросети; $w_{j1}, w_{j2}, \dots, w_{jm}$ – весовые коэффициенты нейросетевой модели.

Нахождение однородных подмножеств во входном массиве данных рассмотрим как задачу кластерного анализа. Имеется конечный массив обучающих выборок $X^n = \{x_1, x_2, \dots, x_n\} \subset X$. Выбран вид функции расстояния между объектами $\rho(x, x^*)$ (евклидово расстояние). Необходимо разделить массив данных на непересекающиеся подмножества, так, чтобы каждое состояло из объектов, близких по метрике $\rho(x, x^*)$, а объекты разных подмножеств существенно отличались между собой, то есть:

$$\sum_i^n d(X_i, \bar{X}_j) \Rightarrow \min; \quad (3)$$

$$\sum_j^K d(\bar{X}_h, \bar{X}_j) \Rightarrow \max \quad (4)$$

$$\vec{X} = (x_1, \dots, x_n) \in X; \quad \vec{K} = (k_1, \dots, k_j) \in K$$

X — множество выборок признаков описания инвестиционных состояний; K —

множество номеров кластеров; $\vec{X} = (x_1, \dots, x_m) \in X$; $\vec{K} = (k_1, \dots, k_j) \in K$ - множества значений входных факторов и номеров классов; $h, j = 1, 2, \dots, K$ – средние значения в

кластерах; $d(\bar{X}_i, \bar{X}_j)$ – расстояние между объектом и центром кластера; $D(\bar{X}_h, \bar{X}_j)$ – расстояние между центрами кластеров. Решив задачу кластерного анализа и добившись однородности данных выборок в каждом подмножестве, зависимость «вход—выход», будет представлять собой непрерывную функцию [7]. Построим для каждого из кластеров отдельную нейронную сеть для многомерной регрессии, получим физическую модель техногенной нагрузки как реализацию функции (1). Прогноз динамики заболеваний и выбор оптимальных управляющих факторов для достижения целевого состояния имеют детерминированную связь через массив синаптических весов обученной нейронной сети.

3. Результаты моделирования в среде STATISTICA Neural Networks

Воспользуемся обобщенными показателями экологического мониторинга на примере Черкасского региона [3], составим исходную таблицу данных в пакете технического анализа Statsoft (Statistika 6.1) и проведем кластеризацию данных в пространстве наблюдений (строк) по известной методике [7] для стандартизованных данных. Результатом стандартизации является приведение всех переменных к единой шкале: данные изменяются в пределах нуля в диапазоне ± 3 , причем большая часть всех значений будет принадлежать интервалу $(-1, 1)$. Процедура стандартизации не изменяет структуру взаимодействий между переменными, не влияет на структуру кластеров:

Элементы кластера номер 1 (РЕЗУЛЬТАТ ЭКОЛОГИЧЕСКОГО МОНИТОРИНГА) и расстояния до центра кластера. Кластер содержит 95 набл.						
	3	4	5	6	7	
Расст.	356,3160	902,6745	348,7101	201,4843	329,9767	271

Элементы кластера номер 2 (РЕЗУЛЬТАТ ЭКОЛОГИЧЕСКОГО МОНИТОРИНГА) и расстояния до центра кластера. Кластер содержит 232 набл.						
	Var1	2	101	102	103	1
Расст.	710,9604	710,9604	39,85736	80,80621	343,3665	194

Элементы кластера номер 3 (РЕЗУЛЬТАТ ЭКОЛОГИЧЕСКОГО МОНИТОРИНГА) и расстояния до центра кластера. Кластер содержит 15 набл.						
	26	51	76	118	119	1
Расст.	1259,087	886,3760	948,9191	435,3337	661,3876	281

Таблица 1. Кластеры в данных экологического мониторинга

Можно утверждать, что все множество данных хорошо разделяется на три кластера. Алгоритмы, реализующие процедуры группировки данных, обеспечивают однородность данных внутри кластеров и максимальную отдаленность самих кластеров [7]. Следовательно, условия (3) и (4) при выполнении процедуры формирования однородных групп в исходных данных соблюдены.

Далее реализуем процедуру (1) как задачу многомерной регрессии внутри каждого кластера. Наблюдения соответствующего кластера используем в задаче многомерного регрессионного анализа. На втором этапе реализуем процедуру построения многомерной регрессии в режиме «обучения с учителем» [3]. Анализ профилей моделей (Таблица 2) позволяет выделить сети с приемлемыми для практики производительностью и ошибками на обучающих, контрольных и тестовых множествах. После выполнения процедуры обучения получим ансамбль нейросетевых моделей. Найденная регрессионная зависимость устанавливает связь входных факторов и выходных индикаторов системы через массивы синаптических коэффициентов пяти обученных нейросетевых моделей.

Таблица 2. Профили построенных моделей и показатели их продуктивности

Результаты работы моделей (Полтавська область.stw)										
N	Архитектура	Производительность обуч.	Контр. производительность	Тест. производительность	Ошибка обучения	Контрольная ошибка	Тестовая ошибка	Входы	Скрытые(1)	Скрытые(2)
1	РБФ 17:17-4-1:1	0,096655	0,066709	0,143559	0,079432	0,050451	0,051091	17	4	0
2	МП 10:10-5-1:1	0,085117	0,103619	0,469839	0,033707	0,028942	0,067853	10	5	0
3	Линейная 16:16-1:1	0,000000	0,071421	0,256897	0,000000	0,022746	0,050861	16	0	0
4	Линейная 17:17-1:1	0,000000	0,069870	0,251321	0,000000	0,022252	0,049757	17	0	0
5	МП 5:5-3-1:1	0,039115	0,081631	0,258584	0,013231	0,021307	0,038857	5	3	0

Таким образом, можно считать, что сеть успешно распознала структуру обучающего множества и пригодна для использования в прогнозировании значений зависимой переменной. Математическая модель связывает факторы и состояния техногенной нагрузки региона с приемлемыми показателями производительности. Вариации входными факторами позволяют получать различные значения функции, что обеспечивает системного менеджера данными для принятия решений.

Выводы. Анализ экологической нагрузки региона как решение задачи распознавания кластеров и построение регрессии в каждом из них реализуется в среде стандартных эмуляторов пакета технического анализа StatSoft в пределах заданных показателей адекватности. Практическая значимость результатов исследования состоит в оперативном анализе рисков и повышении объективности при принятии решений по оптимизации техногенной нагрузки в регионе.

Литература

1. Бусленко Н.П. Моделирование сложных систем. 2 – изд. – М.: Наука, 1978. – 400с.
2. Морозов А.А., Теслер Г.С. Ситуационное управление и системы поддержки принятия решений //Збірник доповідей науково – практичної конференції. ПІММС НАН України / Системи підтримки прийняття рішень. Теорія і практика. – Київ – 2005. – С.5-9.
3. Голуб С.В. Методологія створення автоматизованих систем багаторівневого соціоекологічного моніторингу: автореф. дис. на здобуття наукового ступеня д-ра техн. наук: 05.13.06 / С.В. Голуб. – Київ, 2008. – 35 с.
4. Хайкин С. Нейронные сети: полный курс, 2-е издание. Пер. с англ. - М.: Издательский дом «Вильямс», 2006 г. - 1104с.
5. Морозов А.А. Состояние и перспективы нейросетевого моделирования СППР в сложных социотехнических системах / В.П. Клименко, А.Л. Ляхов, С.П. Алёшин // Математичні машини і системи. – 2010. - № 1.- С. 127 – 149.
6. Барабаш Ю. Л. Коллективные статистические решения при распознавании.- М.: Радио и связь , 1983.- 224с.
7. Боровиков В.П. STATISTICA NN - Техническое описание. М.: Мир, 1999. – 239 с.