РОССИЙСКАЯ АКАДЕМИЯ НАУК
СИБИРСКОЕ ОТДЕЛЕНИЕ
А В Т О М Е Т Р И Я

2004, том 40, № 3

УДК 681.5 : 58

В. И. Красинский

(Новосибирск)

ПРИМЕНЕНИЕ ТЕХНОЛОГИИ ИСКУССТВЕННЫХ НЕЙРОННЫХ СЕТЕЙ ДЛЯ ПРОГНОЗА ФЛОРИСТИЧЕСКИХ ПАРАМЕТРОВ ГЕРБАРНЫХ ЭТИКЕТОК

Представлены результаты классификации документов ботанической базы данных гербарных этикеток по совокупности двух качественных признаков на основе одного из математических методов искусственного интеллекта – теории искусственных нейронных сетей. Прогнозировались значения целевого качественного признака других документов этой базы данных, что позволило выявить несколько ошибок в документах.

Введение. Модель искусственных нейросетей (ИНС) появилась в 40-е годы XX века для объяснения способов решения задач животными и человеком. Универсальность ИНС как аппроксиматоров функций произвольного вида доказана в работах [1, 2]. Решение задачи классификации и распознавания объектов в какой-либо предметной области с помощью ИНС состоит из двух этапов. Сначала сеть последовательно получает набор входных признаков обучающих примеров-прецедентов с известными ответами о значениях целевого признака, в этом режиме настраиваются параметры сети. На втором этапе (эксплуатация) на вход подаются признаки-условия объекта, а сеть дает ответ-прогноз о значении целевого признака, т. е. распознает ситуацию.

В настоящее время имеются компьютерные программы для имитации нейросетей, их использование экономит время исследователя при разработке модели и позволяет сосредоточиться на решении задачи в предметной области. В данной работе в качестве инструментария был использован пакет программ "STATISTICA Neural Networks" (SNN), имеющий хорошие возможности для конструирования различных типов нейросетей и анализа качества их работы [3]. Программы этого пакета могут работать с разнотипной (количественной и качественной) входной информацией.

Постановка задачи в предметной области. При создании электронной базы данных (БД) каталога гербария в Центральном сибирском ботаническом саду (ЦСБС) СО РАН впервые предложено кодировать экологические и флористические параметры места сбора растения и совместно с ботаниками разработаны правила кодирования документа гербарной этикетки [4]. Фактически при этом создается тезаурус сведений, содержащихся в гербарных этикетках. При накоплении в электронной БД значительного объема документов появилась возможность применения методов искусственного интел-

лекта (ИИ) для анализа сложной, часто недоопределенной, ботанической информации. В работе [5] показано успешное применение теории нечетких множеств (НМ): впервые решена задача распознавания образов на пересекающихся (толерантных) классах, в частности надежная диагностика растений. Разработанный метод был применен и для поиска ошибок в документах БД. По модели нечеткого дескриптора по списку ключевых слов предсказана степень ошибок описаний мест сбора растений [6]. Для разработки моделей на основе НМ, особенно в гуманитарных науках, требуется создание тезаурусов и баз данных по предметной области. Также необходимо привлечение знаний специалистов для формулирования гипотез. На основе теории нечетких множеств можно делать выводы из недоопределенных фактов, но при известной четкой или нечеткой таксономии этих фактов. Технология нейросетей в отличие от теории НМ дает возможность формировать и проверять таксономию. Поэтому оба эти направления ИИ важны для решения сложных задач анализа биологической информации.

Цель данной работы – проверить таксономические и прогнозные возможности нейросетевой технологии для выявления ошибок в массиве ботанической гербарной информации. Проверка проводилась на массиве гербарных этикеток с известной классификацией по отношению к основному типу растительности. Для классификации документов и последующего распознавания типа растительности в описаниях этикеток входными параметрами являются номинальные переменные PHIE (нефлористическая характеристика фитоценоза) и PHIF (тип формаций). Выходная переменная для классификации и прогноза – переменная PHIT (основной тип растительности). Был проверен и третий входной параметр – высота сбора растения (в совокупности с PHIE и PHIF), но этот признак оказался малозначимым на представленном материале для классификации основного типа растительности и поэтому не применялся при окончательном конструировании сети. В этом результате проявилось важное свойство нейросетей – ранжирование признаков по их значимости для классификации предъявленных объектов.

Исходные данные. Первичные описания документов – это обычные тексты, поэтому формализованные поля документов БД имеют номинальный тип. Количество обучающих примеров при создании искусственных нейросетей индивидуально для каждой задачи и определяется опытным путем. Плохие эксплуатационные свойства нейросети могут получиться как при недообучении (интерполяция сложной функции по малому числу экспериментальных точек), так и при переобучении – сеть начнет реагировать на случайные колебания входных сигналов (плохие сглаживающие свойства). Считается, что во многих прикладных задачах при обработке качественных признаков обучающая выборка должна быть порядка нескольких сотен объектов. Из 13 тысяч документов БД гербария NS для нейросетевого анализа были отобраны документы, содержащие в поле основного типа растительности PHIT одно из четырех значений: «лес», «луг», «степь», «тундра» – это наиболее распространенные типы растительности Сибири. Из этого промежуточного набора были отобраны 1543 документа таких, что гистограммы по каждому значению (терму) анализируемых номинальных флористических показателей PHIT, PHIE, PHIF содержали не менее 20 документов. Далее, для объективности анализа из этих 1543 документов с помощью датчика случайных чисел были отобраны 200 документов в качестве обучающей выборки (ее фрагмент приведен в приложении) и 100 документов в контрольный файл.

Распределение 200 документов обучающей выборки: для обучения – 100 документов, для тестирования – 50 документов, для дополнительной проверки – 50 документов. Гистограммы этого файла по номинальным значениям анализируемых трех переменных следующие:

PHIE	Count
долинный	20
каменист	61
остепнен	20
заболоче	26
сырой	6
щебнисты	2
парковый	2
деградир	7
субальпи	2
мелкотра	3
лесной	1
пойменны	22
луговой	5
засоленн	11
альпийск	12
ИТОГО	200

PHIF	Count
лиственн	34
кустарни	28
злаковый	47
разнотра	21
осоковый	16
осочковы	12
ивовый	5
лишайник	4
еловый	3
тополевы	10
ковыльны	11
полынный	3
пырейный	1
ячменевы	2
чиевый	1
мятликов	2
ИТОГО	200

PHIT	Count
лес	52
степь	66
луг	77
тундра	5
ИТОГО	200

Характеристики нейросетевой модели. Значения номинальных признаков PHIE и PHIF каждого документа обучающей выборки предъявлялись сети как условия, а значение признака PHIT предъявлялось как известный ответ. Внутренний алгоритм обучения сети (по методу обратного распространения ошибки) уточнял на каждом примере значения весов связей между нейронами, минимизируя значение функции штрафа на выходе. В итоге обучения ответы сети приближены к правильным на всей обучающей выборке и сеть готова к эксплуатации.

В процессе определения структуры и обучения нейросети на описанном файле из 200 документов программа SNN [3] построила 24 сети разных типов (многослойные персептроны MLP, радиально-базисные RBF, вероятностные PNN) с разными характеристиками, из которых как наилучшую по аппроксимации 100 предъявленных документов выбрала трехслойную сеть MLP с общим уровнем ошибки error = 0,09. Отметим, что нулевая ошибка не

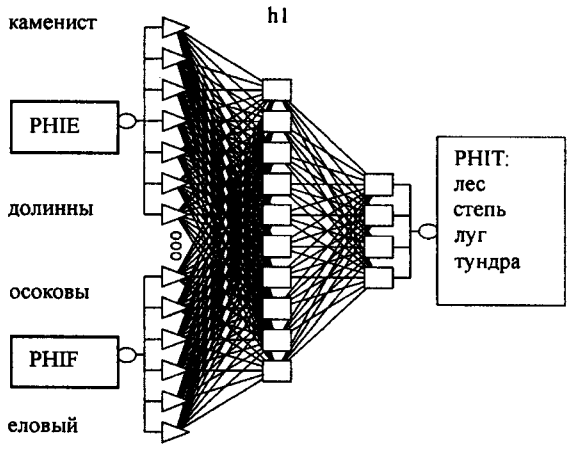

Схема нейросети для классификации гербарных этикеток

может быть достигнута в принципе, поскольку в математическом смысле решается задача интерполяции сложной функции неизвестного вида по ограниченному количеству наблюдений. Оценка программой качества обучения сети – «Excellent». Время этого анализа-обучения составило около 2 мин на процессоре Intel Pentium/360 МГц, ОП = 128 Мбайт.

Первый слой построенной сети – входной, с количеством входов 31 по совокупному числу термов двух входных номинальных переменных. Наличие конкретного значения признака в документе кодируется как «1», отсутствие – как «0». Таким образом, осуществляется преобразование номинальных признаков в нормализованную числовую форму. В этом же числовом интервале анализируются и выходные переменные. Второй (скрытый) слой состоит из 10 нейронов. На рисунке приведена схема построенной сети MLP для классификации и прогноза параметров гербарных этикеток.

В сетях типа MLP нелинейное преобразование вход–выход простого нейрона чаще всего задается логистической функцией [7]:

$$f(s)=\frac{1}{1+\mathrm{e}^{-as}}.$$

Для многовходового нейрона показатель степени as заменяется на скалярное произведение векторов $\mathbf{W}^T\mathbf{X}$, где $\mathbf{W}$ – вектор весовых коэффициентов входов, а $\mathbf{X}$ – вектор входных значений. Это скалярное произведение осуществляет функцию взвешенного суммирования входных сигналов нейрона.

Для дополнительного управления величиной выходов нейронов во внутренних слоях сети часто применяют аддитивные смещения (threshold), которые являются дополнительными варьируемыми слагаемыми на входах нейронов. Значения компонентов вектора $\mathbf{W}$ и смещений подбираются в процессе обучения сети методом обратного распространения ошибки по критерию минимизации квадрата ошибки на совокупности обучающих примеров [7].

Имеются эмпирические соотношения для определения количества L нейронов промежуточного (скрытого) слоя. Одно из них:

$$N/10-n-m\le L\le N/2-n-m,$$

Таблица 1

Входы нейронов	Нейроны выходного слоя			
	лес	степь	луг	тундра
Смещение	-1,419	2,995	0,433	2,365
h1 #01	4,123	-5,311	-4,126	2,148
h1 #02	-6,776	0,007	0,449	3,027
h1 #03	0,939	-1,382	-3,750	1,430
h1 #04	-1,734	5,389	-4,665	-0,166
h1 #05	-0,375	6,888	-2,988	-3,839
h1 #06	-3,796	-4,426	7,746	-2,375
h1 #07	-2,405	3,080	-2,643	0,304
h1 #08	-1,753	-1,451	1,520	1,196
h1 #09	-1,615	-1,488	2,893	-2,433
h1 #10	5,489	-0,519	-2,347	-3,531

где N – число объектов обучающей выборки; n – размерность входного сигнала; m – размерность выходного сигнала. В анализируемом примере $N=100$, $n=31, m=4$, и приведенное соотношение приобретает вид $-25<L<15$. В реальности $L<15$, т. е. число нейронов скрытого слоя L, удовлетворяющее этому неравенству, весьма вариабельно, и алгоритм обучения использует это число как один из параметров, минимизирующих ошибку классификации.

Для обученной сети получилось значение $L=10$. В табл. 1 приводятся весовые коэффициенты входов нейронов выходного слоя.

Итоги работы нейросети. В режиме распознавания (прогноза значений параметра PHIT) 100 документов обучающего файла и 100 документов контрольного файла выявились три ошибочных, по мнению сети, документа-этикетки (табл. 2).

Таблица 2

Номер документа	Описание места сбора	Кодированное PHIE	Кодированное PHIF	Кодированное PHIT	Прогноз PHIT
3330	Каменистый разнотравно-злаковый луг	каменист	злаковый	луг	степь
6267	Заболоченная осоковая тундра	заболоче	осоковый	тундра	луг
884	Лишайниково-кустарниковая каменистая тундра	каменист	кустарни	тундра	степь

Таблица 3

№ п/п	Выходной нейрон	Значение выхода
1	лес	0,657
2	степь	0,0001
3	луг	0,382
4	тундра	0,014

Ведущие специалисты-ботаники подтвердили правильность диагноза нейросети для этих трех этикеток: каменистой чаще всего бывает степь, а заболоченный и осоковый – это обычно луг, т. е. в описаниях трех мест сбора растений были допущены ошибки.

В качестве примера в табл. 3 приводятся значения выходов нейронов выходного слоя для одной из распознаваемых этикеток, у которой значения входов сети следующие: PHIE = «заболоченн», PHIF = «еловый».

Прогнозный ответ нейросети по этому документу – «лес» (правильный ответ), с уровнем ошибки 0,26. Порог принятия решения о принадлежности документа к одному из четырех классов настроен в алгоритме на уровень >0,5, т. е. в представленном примере сеть весьма сильно сомневается в ответе, поскольку относительно большой вес имеет и вариант ответа – «луг».

Числовые выходы нейронов выходного слоя сети в некоторых моделях используются как значения функций принадлежности (ФП) неизвестных входных объектов к нечетким множествам. Так, если расплывчатый реальный тип растительности считать лингвистической переменной с четырьмя термами (см. табл. 3), то предъявленный к распознаванию объект имеет степень 0,66 принадлежности к терму «лес» и степень 0,38 принадлежности к терму «луг». Остальные два значения близки к нулю. В подобных моделях нейросеть является вычислителем значений функций принадлежности новых объектов к нечетким множествам на основе критерия многомерной похожести на совокупность объектов, включенных в обучающую выборку. Итоговый вывод по задаче (прогноз, сигнал управления) производится методами теории НМ, при этом возможен учет экспертных мнений специалистов предметной области. Подобные гибридные модели получили название «Neuro-Fuzzy Synergism», они весьма перспективны для решения сложных научных и технических задач.

Заключение. Методы искусственного интеллекта интенсивно внедряются в различные отрасли науки и техники, поскольку позволяют решать сложные реальные задачи, которые трудно формализуются классическими методами аналитической математики и теории вероятностей. Методы искусственного интеллекта позволяют включить в анализ знания специалистов-экспертов предметной области. Алгоритмы, оперирующие с нечеткими предикторами, вычисляемыми на основе нейросетевых методов, внедряются для управления сложными технологическими процессами. Возможность реализации новых способов анализа информации связана с технологическим прорывом в области микроэлектроники в последние годы.

Приведенные в работе результаты показывают возможность успешного применения нейросетевой технологии в такой гуманитарной науке как ботаника для классификации объектов и прогноза некоторого их признака по совокупности других качественных признаков объектов. Возможно, в частности, выявление ошибок в базах данных. Подобным образом можно решать и другие классификационные задачи ботаники и экологии, поскольку трактовка «ошибки» – это вопрос интерпретации пользователем результатов

расчетов компьютерной программы в ответ на предъявленную входную информацию.

ПРИЛОЖЕНИЕ

Фрагмент обучающей выборки

Номер документа	GEN	SPE	HEI высота	PHIT тип	PHIE нефлорист.	PHIF формация	PHIA ассоциация
95	POTENTIL	PARADOXA	1400	лес	долинный	лиственн	еловый
10870	POA	PRATENSI	0	лес	долинный	лиственн	еловый
41	ORTHILIA	OBTUSATA	1400	лес	долинный	лиственн	еловый
10878	CALAMAGR	NEGLECTA	0	лес	долинный	лиственн	еловый
5752	TROLLIUS	ASIATICU	0	лес	долинный	ивовый	березовы
11014	FESTUCA	RUBRA	0	лес	долинный	ивовый	
452	RIBES	ALTISSIM	1100	лес	каменист	лиственн	
9300	ELYMUS	MUTABILI	0	лес	остепнен	лиственн	
9035	KOELERIA	CRISTATA	970	степь	каменист	полынный	разнотра
3286	SERRATUL	MARGINAT	0	степь	каменист	полынный	злаковый
6904	YOUNGIA	TENUIFOL	1600	степь	каменист	мятликов	
2619	ARTEMISI	COMMUTAT	0	степь	деградир	осочковы	
11233	LEYMUS	SECALINU	0	степь	мелкотра	злаковый	карагано
2977	ARTEMISI	TOMENTEL	0	степь	мелкотра	злаковый	разнотра
2978	ARTEMISI	TOMENTEL	0	степь	мелкотра	злаковый	полынный
11229	HELICTOT	SCHELLIA	0	степь	луговой	злаковый	разнотра
2971	SERRATUL	CENTAURO	0	степь	луговой	злаковый	разнотра
5574	THALICTR	PETALOID	0	степь	луговой	разнотра	овсянице
3265	ARTEMISI	OBTUSILO	0	луг	засоленн	пырейный	чиевый
11190	PUCCINEL	TENUISSI	0	луг	засоленн	ячменевы	разнотра
11188	LEYMUS	PABOANUS	0	луг	засоленн	ячменевы	вострецо
2953	TARAXACU	DEALBATU	0	луг	засоленн	чиевый	вострецо
6393	SALIX	VESTITA	1950	луг	альпийск	кустарни	
6392	JUNIPERU	PSEUDOSA	1950	луг	альпийск	кустарни	
3720	DORONICU	ALTAICUM	2100	луг	альпийск	злаковый	разнотра
3722	SENECIO	PRATICOL	2100	луг	альпийск	злаковый	разнотра
6748	BERBERIS	SIBIRICA	2400	тундра	каменист	лишайник	
6749	PEDICULA	AMOENA	2400	тундра	каменист	лишайник	
6267	POLEMONI	VILLOSUM	2400	тундра	заболоче	осоковый	
6240	VALERIAN	CAPITATA	2180	тундра	щебнисты	лишайник	разнотра

Примечание. В колонках GEN–SPE в сокращенном виде записан вид конкретного растения, хранящегося в гербарии NS ЦСБС СО РАН.

СПИСОК ЛИТЕРАТУРЫ

1. **Галушкин А. И.** Теория нейронных сетей. М.: Радиотехника, 2000.
2. **Горбань А. Н.** Обобщенная аппроксимационная теорема и вычислительные возможности нейронных сетей // Сиб. журн. вычисл. мат. 1998. **1**, № 1. С. 12.
3. **Нейронные сети.** STATISTICA Neural Networks: Пер. с англ. М.: Горячая линия – Телеком, 2001.
4. **Красноборов И. М., Красинский В. И., Артемов И. А.** Ботанические компьютерные базы данных и анализ флористической информации // Тр. IV междунар. симп. по результатам международной программы биосферного мониторинга «Эксперимент Убсу-Нур». М.: ИНТЕЛЛЕКТ, 1996. С. 81.
5. **Красинский В. И.** Диагностика объектов, характеризующихся разнотипными признаками, по отношению к пересекающимся классам: Автореф. дис. ... канд. техн. наук /НИОХ СО РАН. Новосибирск, 2002. 19 с.
6. **Красинский В. И.** Предсказание ошибок в документах базы данных на основе нечеткого дескриптора по ключевым словам // V рабочее совещание по электронным публикациям EL-PUB2000. Новосибирск, 2000. http://www-sbras.nsc.ru/ws/el-pub-2000/
7. **Омату С., Халид М., Юсоф Р.** Нейроуправление и его приложения: Пер. с англ. М.: ИПРЖР, 2000. Кн. 2.

Центральный сибирский ботанический сад СО РАН,
E-mail: vkras@ngs.ru

Поступила в редакцию
3 июля 2003 г.