УДК 624.131;725.3:681.3.06

ИСПОЛЬЗОВАНИЕ АЛГОРИТМОВ DATA MINING ДЛЯ РЕШЕНИЯ ПРОГНОЗНЫХ ЗАДАЧ ПРИ СТРОИТЕЛЬСТВЕ МЕТРОПОЛИТЕНА

Л.А. Строкова

Томский политехнический университет
E-mail: geyer@tpu.ru

Рассмотрен пример применения алгоритмов деревьев решений и искусственных нейронных сетей для задач прогнозирования величины осадки по данным натурных наблюдений. Показано преимущество самоорганизующихся карт Кохонена по сравнению с регрессионным анализом данных мониторинга. Сделан вывод об эффективности использования алгоритмов Data Mining для решения прогнозных задач инженерной геодинамики.

Ключевые слова:
Прогнозирование осадки, проходка горных выработок, деревья решений, самоорганизующиеся карты Кохонена.

Планирование и проектирование сооружений в условиях городской застройки с наличием метрополитена требует привлечения знаний мониторинга о негативных инженерно-геологических процессах, в том числе, о величине осадки, вызванных строительством метро. Прогнозирование величины осадки, одна из многих задач инженерно-геологической практики, при решении которой преодолеваются различные типы неопределенностей, связанные, например, со случайным характером пространственной изменчивости параметров грунтового массива, неизвестностью закона распределения признаков, с неточностью средств измерения, отсутствием представительного ряда наблюдений и других. Соответственно, встает под сомнение применимость классической теории вероятности для решения прогнозных задач. В последние годы при работе с массивами разнородных данных стала использоваться технология Data Mining [1]. Рассмотрим практическое использование алгоритмов Data Mining на платформе Deductor для прогнозирования величины оседания поверхности, вызванной проходкой метрополитена.

Исходными данными для эксперимента послужили материалы мониторинга за оседанием поверхности, вызванные строительством метрополитена в г. Мюнхен. Многолетний мониторинг за осадками ведется сотрудниками Технического университета г. Мюнхена под руководством И. Филлибека (Dr. Fillibeck). Автор выполняла работы по математическому моделированию оседания поверхности [2]. Моделирование осуществлено по 40 поперечникам, для которых известно фактическое оседание поверхности. Параметры массива данных мониторинга можно разделить на 2 группы. *Первая группа* – это геометрические параметры, характеризующие расположение туннеля в горном массиве (глубина залегания туннеля Z, отношение мощности перекрывающих пород над туннелем к диаметру проходки H/D, отношение расстояния между осями двойных туннелей к диаметру A/D, величина осадки, ширина корыта оседания). *Вторая группа* – геологические параметры: состав, возраст, условия залегания горных пород от дневной поверхности до подошвы туннеля.

Моделирование при помощи программы PLAXIS проводилось с целью нахождения закономерностей между параметрами мониторинга за осадкой и на основании этого иметь возможность предсказывать ожидаемую величину осадки для новых линий метрополитена. Для решения этой задачи вначале по традиции были применены методы регрессионного анализа. Для уменьшения числа переменных весь массив данных был разбит на 3 выборки, внутри которых были одинаковыми состав пород и способ проходки. В первую выборку было отобрано 24 разреза, в которых туннель расположен в четвертичных отложениях.

Использование статистических методов для выявления зависимости между величиной осадки S_{max} и геометрическими параметрами H/D, A/D показало, что связь между этими параметрами незначительная (коэффициент детерминации $R^2<0,1$, рис. 1). Связь между параметрами величиной осевшего грунта над туннелем в результате проведения горных работ V_s и глубиной залегания туннеля Z оказалась более значимой (коэффициент детерминации $R^2=0,48$, рис. 2). Эти методы позволили установить только диапазон изменения параметров с разным доверительным уровнем.

Использование сведений о составе пород также не позволило выявить какие-либо однозначные закономерности в данных по мониторингу, которые позволили бы прогнозировать величину осадки в новых условиях. Тогда было выдвинуто предположение об эффективности использования для решения этой задачи «машинное обучение». Для интеллектуального анализа данных Data Mining, были выбраны такие алгоритмы как деревья решений и самоорганизующиеся карты Кохонена. Данные методы хорошо зарекомендовали себя во многих областях науки и техники, поскольку обладают свойством адаптивности, обобщения, извлечения знаний и моделирования сложных нелинейных зависимостей в массивах данных [3, 4].

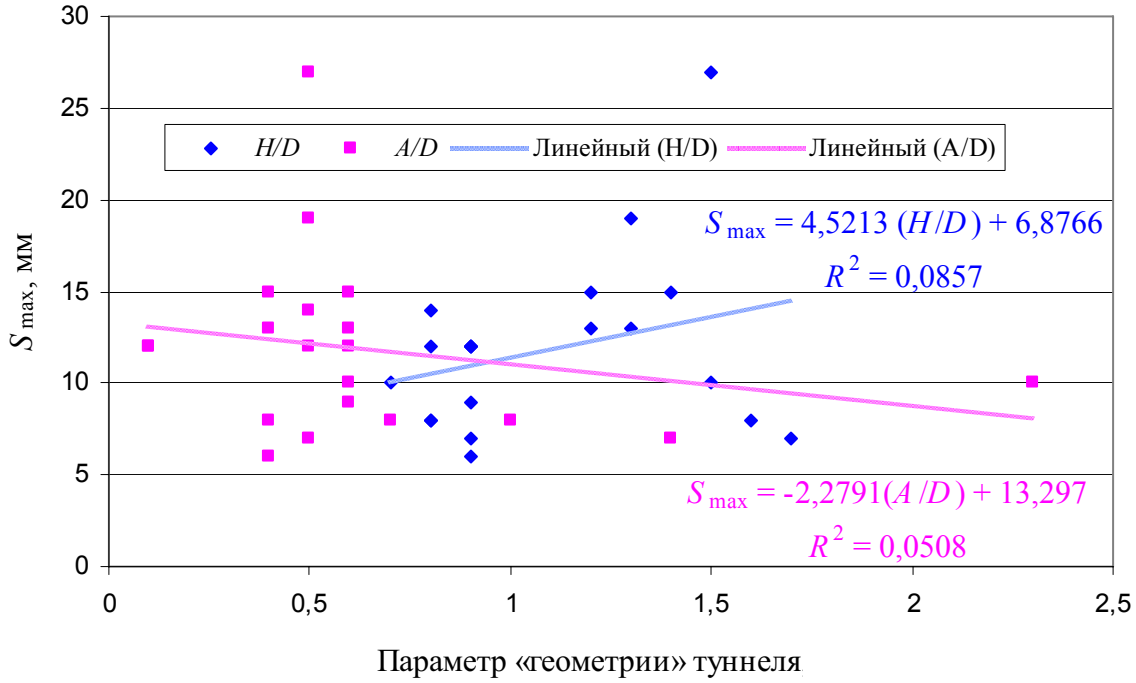

Рис. 1. *Диаграмма рассеивания*

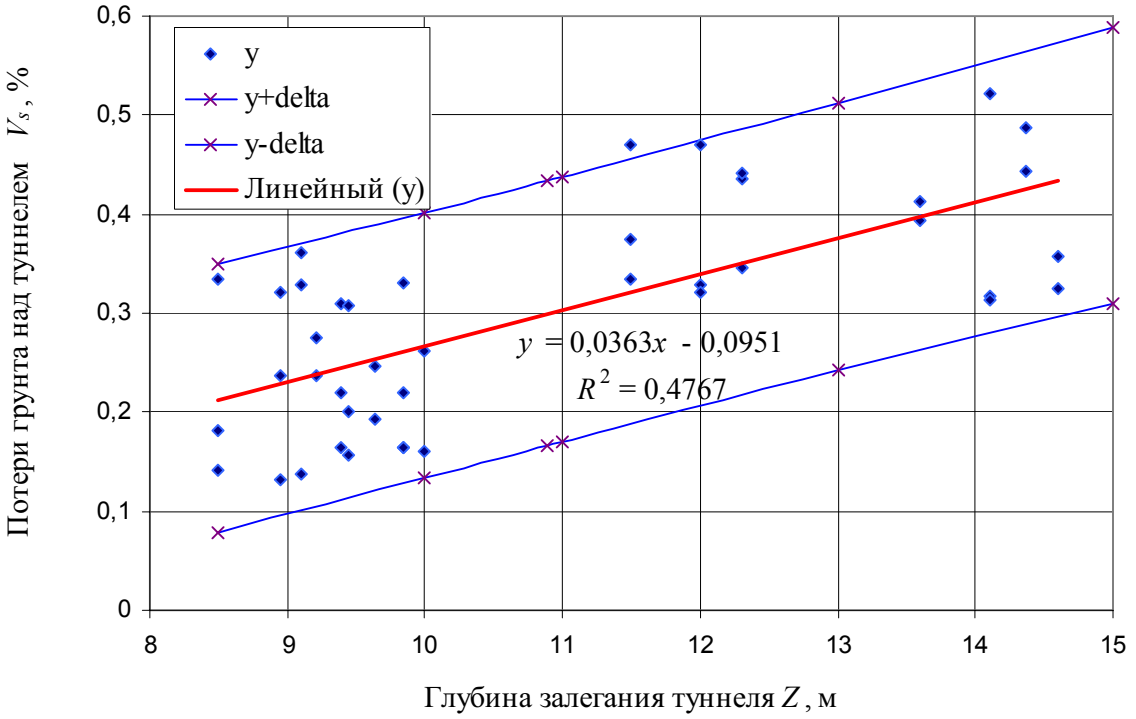

Рис. 2. *Доверительная область изменения V_s от Z с доверительным уровнем 90 %*

Алгоритм дерево решений (рис. 3) отобрал 11 решающих правил о прогнозируемой величине осадки по 4 значимым факторам (табл. 1). Значимость факторов представлена в табл. 2. Самым значимым из них является отношение H/D, с уровнем значимости 49 %, что еще раз подтвердило данные математического моделирования с вариацией параметров [2]. Однако было распознано только 12 случаев из 24. Скорее всего, массив данных мониторинга не содержит некоторых значимых параметров.

Метод самоорганизующихся карт Кохонена смог распознать 23 случая из 24. Прогнозная величина осадки в мм, может быть определена по карте Кохонена с поддержкой 5...22 % при достоверности решения 33...100 %.

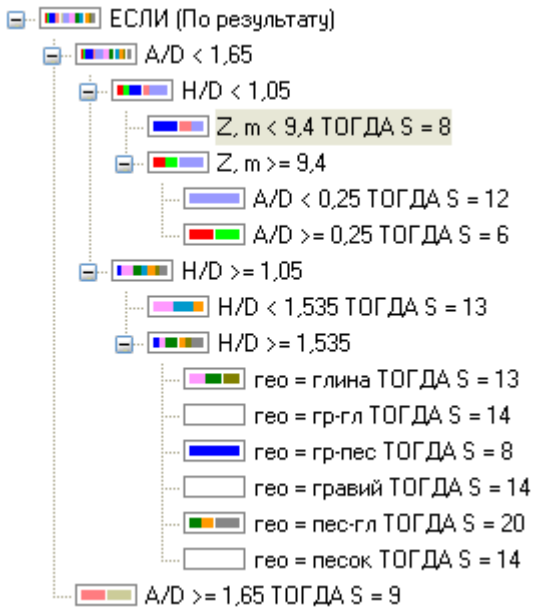

Рис. 3. *Дерево решений для определения величины осадки*

Таблица 1. *Решающие правила для определения возможной осадки*

№	Условие	Следствие (осадка, S_{max}, мм)	Поддержка		Достоверность	
			%	Кол-во	%	Кол-во
1	A/D<1,65 И H/D<1,05 И Z, m<9,4	8	17,38	4	50,0	2
2	A/D<1,65 И H/D<1,05 И Z, m>=9,4 И A/D<0,25	12	8,70	2	100,0	2
3	A/D<1,65 И H/D<1,05 И Z, m>=9,4 И A/D>=0,25	6	8,70	2	50,0	1
4	A/D<1,65 И H/D>=1,05 И H/D<1,535	13	21,74	5	40,0	2
5	A/D<1,65 И H/D>=1,05 И H/D>=1,535 И гео = глина	13	13,04	3	33,3	1
6	A/D<1,65 И H/D>=1,05 И H/D>=1,535 И гео = гравий-глина	14	0,0	0	0,0	0
7	A/D<1,65 И H/D>=1,05 И H/D>=1,535 И гео = гравий-песок	8	4,35	1	100,0	1
8	A/D<1,65 И H/D>=1,05 И H/D>=1,535 И гео = гравий	14	0,0	0	0,0	0
9	A/D<1,65 И H/D>=1,05 И H/D>=1,535 И гео = песок-глина	0	17,39	4	50,0	2
10	A/D<1,65 И H/D>=1,05 И H/D>=1,535 И гео = песок	14	0,0	0	0,0	0
11	A/D>=1,65	9	8,70	2	50,0	1

Последовательно анализируя различные карты входных признаков (рис. 6), видим, что определяющим для разделения участков с разной величиной осадки является отношение H/D. Несмотря на то, что обучение проходило без учителя, алгоритм Кохонена сгруппировал большинство участков в три крупных кластера с высокой поддержкой порядка 20 % и достоверностью около 50 % (рис. 4, 5). Самый большой кластер представлен участками, в которых туннель проходит в гравийно-галечниковых отложениях на небольшой глубине до 10 м. Величина осадки составляет 6...9 мм. В отдельный кластер объединены участки, в которых туннели пройдены в песчано-глинистых грунтах, залегают на глубине 18...20 м, здесь величина осадки максимальна и составляет 20 мм. Малое расстояние между осями двойных туннелей к диаметру A/D на всех участках, независимо от типа горных пород и глубине расположения выработки приводит к увеличению осадки до 12...15 мм. Небольшой кластер объединяет участки, у которых туннель расположен в глинистых грунтах на глубине 15...20 м, величина осадки составляет 13...17 мм.

Таблица 2. *Значимость факторов при прогнозировании величины осадки*

Номер атрибута	Атрибут	Значимость, %
2	H/D	48,86
1	A/D	22,96
4	гео	17,98
3	Z, m	10,20

Результаты классификации вышеописанными способами удобно сравнивать по таблицам сопряженности (табл. 3, 4).

Таблица 3. *Матрица сопряженности метода самоорганизующихся карт Кохонена (распознанные случаи – светлая заливка, нераспознанные – темная)*

	Классифицировано											
Фактически	6	7	8	9	10	12	13	14	15	17	18	20	Итого
6	1												1
7		1											1
8			3										3
9				1						1			2
10					2								2
12						3							3
13							3						3
14								2					2
15									2				2
17										2			2
18											1		1
20												2	2
Итого	1	1	3	1	2	3	3	2	2	3	1	2	24

Из сравнения таблиц сопряженности видно, что алгоритм «самоорганизующиеся карты Кохонена» более наглядно классифицировал участки имеющегося массива данных мониторинга. Для более достоверного прогноза необходимо привлечь дополнительную информацию о других участках строительства метрополитена.

Полученное множество логических правил составило основу базы знаний для прогнозирования величины осадки поверхности, вызванной проходкой метрополитена и принятия мер по минимизации повреждений существующих зданий и фундаментов. Традиционная математическая статистика,

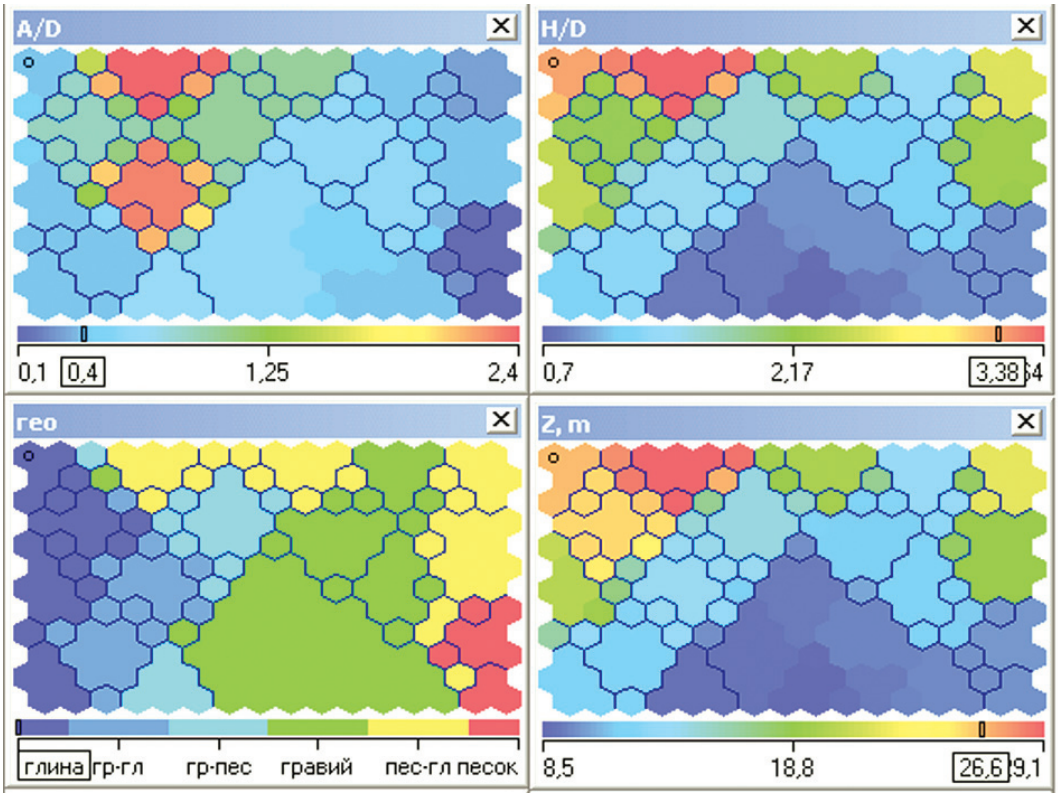

Рис. 4. *Самоорганизующиеся карты Кохонена по входным параметрам*

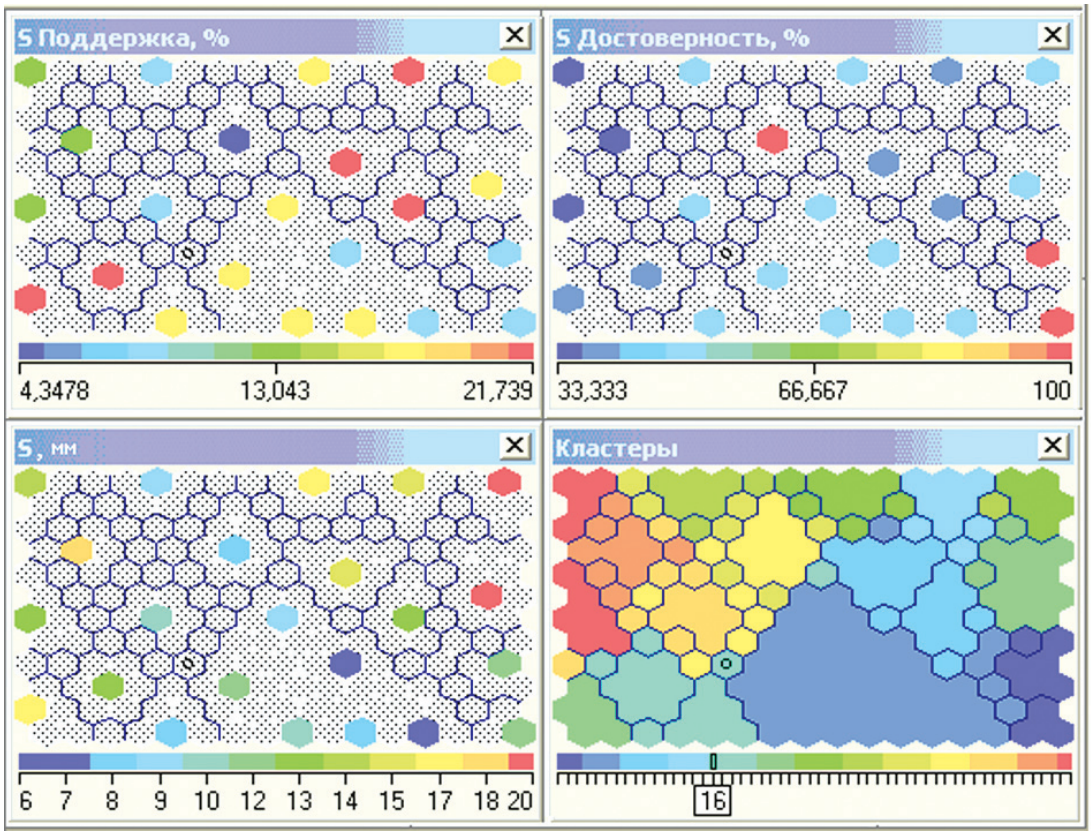

Рис. 5. *Самоорганизующиеся карты Кохонена по выходным параметрам*

Таблица 4. *Матрица сопряженности метода дерева решений (распознанные случаи – светлая заливка, нераспознанные – темная)*

	Классифицировано												
Фактически	6	7	8	9	10	12	13	14	15	17	18	20	Итого
6	1												1
7	1												1
8			3										3
9			1	1									2
10			1	1									2
12			1			2							3
13							3						3
14							1					1	2
15							2						2
17							1					1	2
18							1						1
20												2	2
Итого	2	0	6	2	0	2	8	0	0	0	0	4	24

как основной инструмент анализа данных, не справилась с этой задачей. Главная причина – концепция усреднения по выборке, приводящая к операциям над фиктивными величинами (типа средней глубины залегания туннеля, среднего расстояния между осями двойных туннелей и т. п.). Технология Data Mining лучше справилась с поставленной задачей. Обнаруженные закономерности обладают высокой поддержкой и достоверностью.

Рассмотренный пример представляет собой лишь иллюстрацию процесса разработки базы знаний для принятия решений по планированию и проектированию сооружений в условиях плотной городской застройки современными средствами Data Mining. Для окончательного воплощения в полезный инструмент принятия решений полученный прототип обязан пройти всестороннюю проверку на более обширном массиве данных с возможностью внесения необходимых корректив. Вместе с тем, представляется достаточно показательной приведенная демонстрация преимуществ новейших технологий обнаружения знаний в данных мониторинга для решения задач прогнозирования опасных геологических и инженерно-геологических процессов.

СПИСОК ЛИТЕРАТУРЫ

1. Дюк В.А., Самойленко А.П. Data Mining: учебный курс. – СПб.: Питер, 2001. – 368 с.
2. Строкова Л.А. Моделирование оседания поверхности при проходке туннеля щитовым способом // Известия Томского политехнического университета. – 2008. – Т. 312. – № 1. – С. 45–50.
3. Паклин Н.Б. Анализ геофизических данных // Бурение и нефть. – 2005. – № 5. – С. 38–40. [Электронный ресурс]. – режим доступа: http://www.basegroup.ru/practice/geophysics.htm. – 11.11.2008.
4. Учебные материалы об аналитической платформе «Deductor» компании BaseGroup Labs. [Электронный ресурс]. – режим доступа: www.basegroup.ru. – 11.11.2008.

Поступила 17.11.2008 г.