УДК 004.891

И.А. Драгун

Предобработка данных для нейросетевой оценки операционного риска

Цель настоящего исследования – создание использующего нейросетевые технологии компьютерного программного продукта, позволяющего врачу улучшить диагностику и оценить степень тяжести состояния больного перед операцией и тем самым определить риск хирургического вмешательства.

В рамках разработки данной экспертной системы для оценки операционного риска нами собрана база данных 645 больных желчнокаменной болезнью, содержащая значения качественных и количественных инструментальных и лабораторных показателей больных в 10-дневный предоперационный период. В таблице 1 представлены количественные и качественные признаки и проценты наличия значений признаков для двух групп больных желчнокаменной болезнью: группа A_1 – 446 пациента, выписанные из больницы после операции, и группа A_2 – 199 пациента, умершие после операции.

Предварительная обработка данных для нейросетевого анализа заключается в представлении всех имеющихся данных в числовом виде, причем все значения должны быть нормированы в интервале, соответствующем области значений активационной функции нейросети [1, 2]. Основными особенностями медико-биологических данных являются, во-первых, наличие как количественных, так и качественных значений показателей, во-вторых, зачастую для отдельного пациента не удается собрать весь необходимый набор показателей.

Учитывая объем статистических данных, при проверке качества обучения нейросети необходимо использовать процедуры перекрестного тестирования и скользящей проверки [2]. Также из таблицы 1 видно, что исходная информация нуждается в восстановлении или правдоподобном заполнении недостающих (пропущенных) данных.

Вообще говоря, ситуация с отсутствующими, по каким-либо причинам утраченными данными характерна не только для медико-биологической статистики, поэтому в ряде нейросимуляторов реализовано заполнение пропущенных значений средними либо наиболее вероятными значениями в выборке. Существуют также отдельные программные решения для заполнения попусков в таблицах, например, разработка лабораторией Института вычислительного моделирования СО РАН неравновесных систем №1.3 «Итерационный факторный анализ: метод главных компонент» (FAMaster) [3]. Процесс заполнения пропусков в таблицах основан на итерационном построении интерполяционных моделей известных данных, сначала линейной, а затем на ее основе и нелинейной модели, при этом возможна интерполяция полиномом либо сплайн-интерполяция [3].

Таблица 1

№	Тип	Название	A1	A2	№	Тип	Название	A1	A2
1	кач.	Осложнения	100,0	100,0	18	кол.	АСАТ	73,8	71,0
2	кач.	Данные ЭКГ	69,3	70,0	19	кол.	АЛАТ	73,5	71,0
3	кач.	Цвет	88,8	76,0	20	кол.	Общий белок	85,4	79,0
4	кач.	Прозрачность	85,2	73,0	21	кол.	Мочевина крови	93,0	87,5
5	кач.	Реакция	84,5	76,5	22	кол.	Креатинин крови	71,7	53,0
6	кол.	Полных лет	100,0	99,0	23	кол.	Патрий крови	78,3	84,5
7	кол.	Вес	91,3	62,0	24	кол.	Калий крови	77,8	85,5
8	кол.	Температура при поступлении	76,7	53,5	25	кол.	Сахар крови	76,7	76,5
9	кол.	Температура перед операцией	47,1	70,5	26	кол.	Гемоглобин	92,2	86,5
10	кол.	Белка	78,0	70,5	27	кол.	Лейкоциты	91,7	86,0
11	кол.	Удельный вес	85,0	72,0	28	кол.	ЛИИ	89,2	71,5
12	кол.	Лейкоциты	78,3	69,0	29	кол.	СОЭ	88,1	79,0
13	кол.	Эпителий	72,2	52,0	30	кол.	Длительность кровотечения	80,5	55,0
14	кол.	Билирубин общий	88,6	87,5	31	кол.	Время свертывания	80,7	53,5
15	кол.	Билирубин прямой	87,0	81,0	32	кол.	Протромбиновый индекс	85,4	83,0
16	кол.	Билирубин непрямой	86,5	81,0	33	кол.	Акт на 10 мин.	65,0	68,0
17	кол.	Тимоловая проба	64,8	68,0	34	кол.	Фибриноген по Рутбергу	67,9	70,5

Таблица 2

	Распределения близки к нормальным				Распределения близки к логнормальным			
	FAMaster		Модификация		FAMaster		Модификация	
	A1	A2	A1	A2	A1	A2	A1	A2
Минимальное расстояние	0,0061	0,0172	0,0008	0,0069	0,0078	0,0244	0,0048	0,0067
Максимальное расстояние	1,4075	0,5779	0,3466	0,4534	1,0082	5,9434	0,4986	0,3990
Среднее расстояние	0,2605	0,2204	0,1078	0,1482	0,2139	1,0527	0,1004	0,1350
Стандартное отклонение расстояния	0,1182	0,0303	0,0078	0,0159	0,0809	5,5212	0,0122	0,0144
Изменение среднего значения расстояния	0,0123	0,0119	0,0064	0,0159	0,0166	0,0973	0,0059	0,0150
Изменение стандартного отклонения расстояния	0,0543	0,0079	0,0053	0,0087	0,0328	2,0689	0,0057	0,0076
Процент восстановленных значений, лежащих вне диапазона исходных данных	13,0435	8,7302	0,0000	0,7937	27,0115	38,3333	1,7241	1,6667

Для представления в числовом виде качественную информацию кодируют. Качественные признаки принято делить на три категории: упорядоченные, неупорядоченные и частично упорядоченные [1, 2]. При предварительной обработке качественных данных признаки 3, 4, 5 таблицы 1 закодированы как упорядоченные, а 1 и 2 следует считать неупорядоченными признаками. Пропущенные значения качественных признаков заполняются наиболее вероятным значением в группе.

Использование наиболее вероятных значений вместо отсутствующих «усредняет» данные конкретного пациента и тем самым лишает нейросеть возможной дополнительной информации об обучающей группе. При восстановлении пропущенных значений числовых признаков программой FAMaster мы получили высокий процент «выпавших» из интервала изменения значений. Поэтому для наиболее правдоподобного заполнения пропусков нами предлагается следующая модификация алгоритмов FAMaster: при расчете линейного приближения отсутствующие значения заменять наиболее вероятными, что способствует повышению качества восстановления и скорости счета.

Качество заполнения пробелов легко оценить по Евклидову расстоянию между тестовыми и восстановленными данными. В таблице 2 представлены значения некоторых характеристик заполнения пробелов в двух наборах тестовых данных посредствам программы FAMaster и модифицированным алгоритмом для двух состояний. Для заполнения пробелов проводилась сплайн-интерполяция. Для удобства представления исходные данные пронормированы в интервале от 0 до 1.

Литература

1. Миркес Е.М. Нейрокомпьютер. Проект стандарта. Новосибирск, 1998.

2. Горбань А.Н. Нейроинформатика / А.Н. Горбань, В.Л. Дунин-Барковский и др. Новосибирск, 1998.

3. Моделирование данных при помощи кривых для восстановления пробелов в таблицах / Россиев А.А. // Методы нейроинформатики. Красноярск, 1998.