

Эволюционный метод восстановления пропусков в данных

*Снитюк В.Е., к.т.н., доц., Киевский национальный университет
им. Тараса Шевченко г. Киев snutyuk@gmail.com*

В статье выполнен анализ моделей и методов, предназначенных для восстановления отсутствующих данных. Предложен эволюционный метод, базирующийся на композиции использования нейронной сети и генетического алгоритма. Технология восстановления пропусков не требует выполнения ограничений, связанных с линейностью модели, распределением параметров и других.

Введение

Проблема обработки и восстановления пропущенных значений в данных свойственна многим практическим задачам. Чаще всего она возникает при идентификации зависимостей, априорная информация о значении параметров которых является неполной. Объективными причинами этого являются поломки оборудования при измерении значений технических характеристик процессов, потеря ретроспективной информации, экстремальный характер функционирования, ограниченный доступ и другие. Субъективные причины указывают на невозможность получения полной и точной информации вследствие влияния психологических аспектов и особенностей памяти.

Адекватная аналитическая обработка информации с пропусками осложняется из-за невозможности построения адекватных математических моделей и их использования для прогнозирования и решения сопутствующих задач.

Анализ известных методов восстановления пропусков

Наиболее распространенными методами обработки неполной информации в таблицах данных являются такие:

1. **Изъятие некомплектных строк из таблицы.** Строка или столбец таблицы данных называется некомплектным, если в нем отсутствует хотя бы одно значение. Метод применяется при большой размерности таблицы и незначительном количестве пропусков. В Сборник трудов VI-й Межд. конф. "Интеллектуальный анализ информации". – Киев. – 2006. – С. 262-271.

гих случаях метод ведет к смещенности оценок, поскольку строки с пропущенными значениями содержат новую информацию, необходимую для анализа.

2. **Заполнение пропусков средними значениями.** В результате применения такого метода несколько значений одного фактора оказываются одинаковыми, что указывает на его низкую точность.
3. **Метод ближайших соседей [1].** Находят строки таблицы, которые по определенному критерию (обычно, минимума декартового расстояния), являются ближайшими к строке с пропуском. Для его заполнения значения фактора у соседей усредняются с весовыми коэффициентами, обратно пропорциональными их декартовому расстоянию к строке, которая содержит пропуск. Метод точнее предыдущего, но он практически неприменим в случае большого количества пропусков и базируется на предположении о существовании связей между объектами.
4. **Регрессионный метод [1].** По комплектным данным строится уравнение линейной множественной регрессии и вычисляются пропущенные значения факторов. Метод невозможно применить, если количество пропусков в строке больше одного, что приводит к множеству решений, и кроме того, в реальных задачах зависимости, чаще всего, нелинейные, поэтому его точность является невысокой.
5. **Метод максимальной правдоподобности и EM-алгоритм [2].** Требуется проверка гипотез о распределении значений факторов. Применение осложняется в случае большого количества пропущенных значений фактора.
6. **Алгоритм ZET [3].** Главная идея алгоритма заключается в подборе „компетентной матрицы”, используя данные из которой дальше находят параметры зависимости, которая используется для прогнозирования пропущенного значения. Недостатком алгоритма является его локальность, поскольку для вычисления отсутствующего значения используются не все данные таблицы, а лишь их часть. Субъективизм определения размерности „компетентной матрицы” приводит к учету неинформативных „шумовых” факторов и смещению оценки неизвестного значения.
7. **Алгоритм ZetBraid [3].** Основное отличие от алгоритма ZET заключается в определении оптимального размера „компетентной матрицы”. Для оценки „качества” такой матрицы используется дисперсионный метод и метод „креста”. Все другие недостатки, в том числе и статистическая оценка неизвестного значения исключи-

тельно на основе корреляционно–регрессионного анализа, остаются.

8. **Метод Барлетта [1].** Итеративный метод, который применяется для заполнения пропусков в векторе значений результирующей характеристики в допущении, что значения входных факторов являются комплектными. Его недостатком является базирование на предположении о линейной зависимости, но отсутствие обоснования применимости метода наименьших квадратов приводит к ошибкам.
9. **Resampling [1].** Метод имеет те же недостатки, что и предыдущий. Он является итеративным и имеет две модификации. В первой из них некомплектные строки случайным образом заменяют на комплектные из исходной матрицы и рассчитывают уравнение регрессии. Во втором варианте уравнение регрессии получают из комплектной подматрицы, находят оценки неизвестных значений $Y_i, i = \overline{k, m}$ и заменяют пропуски на $Y_i + \varepsilon_i, i = \overline{k, m}$, где ε_i – значения ошибок предыдущих шагов. Ищут уравнение регрессии. После определенного количества итераций значения коэффициентов усредняют. Информационная избыточность на фоне малой мощности множества комплектных данных в первой модификации resampling и информационная недостаточность в композиции со случайным формированием значений исходной характеристики не позволяют получать приемлемые результаты. Кроме того, отсутствуют процедуры оптимизации метода.
10. **Моделирование неполных данных многообразиями малой размерности [4].** Методы, представляющие данное направление, разработаны учеными Красноярской школы нейроматематики под руководством проф. Горбаня О.М. Их главная идея заключается в том, что набор точек, который является многообразием при наличии пропусков, позволяет строить линейные и нелинейные приближения – модели, посредством которых возобновляют пропущенные значения. Результаты алгоритмизации этих методов и экспериментальных проверок засвидетельствовали достаточно высокую точность. Проведенные исследования указывают на удовлетворительное функционирование алгоритма при 10–15% пропусков. В то же время, математические изложения базируются на достаточно сильных предположениях о распределении входных данных, гладкости функций и обусловленности матрицы исходных значений. К недос-

таткам следует также отнести сложность реализации и верификации алгоритма.

Обобщая результаты анализа рассмотренных методов, делаем вывод об их низкой точности, наличии жестких требований к исходной информации, количеству пропусков, размерности матрицы данных, априорных предположениях о существующих зависимостях, сложности реализации, что свидетельствует о необходимости разработки методов, базирующихся на новых парадигмах. В статье предложен метод, интегрирующий в себе преимущества нейронных сетей для решения задачи идентификации, и генетического алгоритма для оптимизации.

Постановка задачи восстановления пропусков

В общей постановке задача восстановления пропусков в таблицах данных является такой:

Пусть $X = (X_1, X_2, \dots, X_n)$ – вектор входных факторов $Y = (Y_1, Y_2, \dots, Y_m)$ – вектор результирующих характеристик, p – количество экспериментов или периодов ретроспективы, $A = (a_{ij})_{i=1}^p_{j=1}^{n+m}$ – матрица исходной информации. Она имеет пропуски, обозначенные звездочками (табл. 1).

Таблица 1. Структура исходной информации

	X_1	X_2	X_3	·	X_{n-1}	X_n	Y_1	Y_2	·	Y_m
1	a_{11}	a_{12}	a_{13}	·	*	a_{1n}	a_{1n+1}	a_{1n+2}	·	a_{1n+m}
2	a_{21}	a_{22}	*	·	a_{2n-1}	a_{2n}	a_{2n+1}	*	·	a_{2n+m}
3	a_{31}	*	a_{33}	·	a_{3n-1}	a_{3n}	a_{3n+1}	a_{3n+2}	·	*
·	·	·	·	·	·	·	·	·	·	·
p-1	a_{p-11}	a_{p-12}	a_{p-13}	·	a_{p-1n-1}	*	a_{p-1n+1}	a_{p-1n+2}		a_{p-1n+m}
P	a_{p1}	a_{p2}	a_{p3}	·	a_{p11}	a_{pn}	a_{pn+1}	a_{pn+2}		a_{pn+m}

Задача восстановления пропусков в данных заключается в нахождении

$$\arg \min_* \|Y - F(X)\|, \quad (1)$$

где $F = (F_1, F_2, \dots, F_m)$ и $Y = (Y_1, Y_2, \dots, Y_m)$ – векторы значений, полученные по идентифицированным зависимостям

$$\mathcal{F}_i = F_i(X_1, X_2, \dots, X_n), i = \overline{1, m} \quad (2)$$

Снитюк В.Е. Эволюционный метод восстановления пропусков в данных и приведенные в табл. 1, соответственно. Задачу (2) детализируем и перепишем в виде

$$\arg \min_* \frac{1}{pm} \sum_{i=1}^p \sum_{j=1}^m (Y_{ij} - F_j(X_1^i, X_2^i, \dots, X_n^i))^2, \quad (3)$$

или

$$\arg \min_* \frac{1}{pm} \sum_{i=1}^p \sum_{j=1}^m (\hat{\epsilon}_{i,j+n} - a_{i,j+n})^2. \quad (4)$$

Если предположить, что зависимости (2) являются линейными, то есть

$$\hat{Y}_i = b_{i0} + b_{i1}X_1 + b_{i2}X_2 + \dots + b_{in}X_n, \quad (5)$$

то задача восстановления пропусков заключается в нахождении

$$\arg \min_* \|Y - BX\|, \quad (6)$$

где $Y = (a_{ij})_{i=1, j=n+1}^{p, n+m}$, $B = (b_{ij})_{i=1, j=0}^m, n$, $X = (a_{ij})_{i=1, j=1}^p, n$.

Решение задач (1)–(6) имеет первый этап, который, в общем случае, заключается в идентификации зависимостей. Отметим, что в задаче восстановления пропусков в таблицах данных процедуры идентификации и оптимизации итеративно повторяются. С некоторыми особенностями решается задача в зависимости от того, где находятся пропуски:

- только среди значений входных факторов;
- только среди значений результирующих характеристик;
- среди значений и входных факторов, и результирующих характеристик;
- среди значений таблицы, в которой входные факторы и результирующие характеристики не выделены.

Модели и метод эволюционного восстановления данных

Предположим, что пропуски имеются только среди значений входных факторов $X = (X_1, X_2, \dots, X_n)$, результирующая характеристика Y одна и существует зависимость

$$Y = F(X) = F(X_1, X_2, \dots, X_n). \quad (7)$$

А.Н Колмогоров и В.И. Арнольд доказали теорему [5-7] о том, что каждая непрерывная функция n переменных, заданная на единичном кубе n -мерного пространства, представима в виде

Сборник трудов VI-й Межд. конф. “Интеллектуальный анализ информации”. – Киев. – 2006. – С. 262-271.

$$f(x_1, x_2, \dots, x_n) = \sum_{q=1}^{2n+1} h_q \left[\sum_{p=1}^n \varphi_q^p(x_p) \right],$$

где функции $h_q(u)$ непрерывны, а функции $\varphi_q^p(x_p)$, кроме того, еще и стандартны, т.е. не зависят от выбора функции f . В терминах теории нейронных сетей (НС) теорема указывает на то, что любая непрерывная функция идентифицируема сетью с одним, как минимум, скрытым слоем нейронов с нелинейными функциями активации. Для идентификации (7) в качестве модели выберем прямосвязную НС с пороговым алгоритмом обратного распространения ошибки. Структура сети и ее элементный базис в экспериментах остается постоянной.

Поскольку входные образы для обучения нейронной сети имеют пропуски значений, то необходимо решить задачу параметрической оптимизации. В качестве метода оптимизации предложено использовать генетический алгоритм (ГА). В работе [8] доказана теорема:

Пусть выполнены следующие условия:

1. Последовательность популяций P^0, P^1, \dots , генерируема ГА, – монотонна, т.е.

$$\forall i \in N: \min\{f(a)/a \in P^{i+1}\} \leq \min\{f(a)/a \in P^i\}.$$

2. Для $\forall a, a'$ элемент a' достижим из a посредством мутации и кроссовера, т.е. через последовательность переходов в ряде структур.

Тогда глобальный оптимум функции f отыскивается с вероятностью 1:

$$\lim_{t \rightarrow \infty} P\{a^* \in P^t\} = 1.$$

Если использовать бинарное представление решений и для формирования популяции решений – элитный отбор, то теорема указывает на сходимость ГА по вероятности.

Для работы ГА необходимо сформировать генеральную и выборочную совокупность хромосом–решений. Хромосома состоит из фрагментов, которые отвечают пропускам в таблице данных:

$$Xr = \langle \text{пропуск } 1, \text{ пропуск } 2, \dots, \text{пропуск } K \rangle.$$

Данные в таблице без учета пропущенных значений нормируем. Если в качестве активационной функции будет использован гиперболический тангенс, то нормирование предпочтительнее осуществлять в отрезок $[-1; 1]$. Количество хромосом в генеральной совокупности определяется заданной точностью результата, в выборочной – исследователем.

На следующем шаге формируем обучающую и контрольную последовательность для обучения НС. Предлагается все образы с пропусками считать элементами обучающей последовательности. Для контрольной

Снитюк В.Е. Эволюционный метод восстановления пропусков в данных

последовательности их использование является проблематичным, поскольку невозможно использовать для верификации недостоверные значения. Соотношение мощности множеств образов обучающей и контрольной последовательности может быть разным, на что влияет соотношение образов с пропусками и без пропусков в исходной таблице.

Процедура восстановления пропусков будет такой:

Шаг 1. Инициализация K хромосом-решений выборочной последовательности.

Шаг 2. $K = 1$.

Шаг 3. Обучение НС на точках обучающей последовательности, где значения пропусков заполнены значениями K -й хромосомы.

При этом решается задача поиска

$$M_K = \min_W \frac{1}{2} \sum_{i=1}^{P_o} (\mathcal{E}_{in+1} - a_{in+1})^2,$$

где W – матрица весовых коэффициентов НС, P_o – количество образов в обучающей последовательности.

Шаг 4. Вычисление целевой функции ГА (fitness-function):

$$G_K = \frac{1}{2} \sum_{i=1}^{P_c} (\mathcal{E}_{in+1} - a_{in+1})^2,$$

где P_c – количество образов контрольной последовательности. Если $G_K < G_{\min}$, то переход на шаг 7.

Шаг 5. $K = K + 1$. Если $K > K_{\max}$, где K_{\max} – количество элементов в выборочной последовательности, то переход на шаг 6, иначе переход на шаг 3.

Шаг 6. Выполнение процедур кроссовера, мутации, определение и отбор хромосом следующей эпохи. Переход на шаг 2.

Шаг 7. Вывод результатов. Конец.

Экспериментальное моделирование

Для верификации разработанного метода проведена экспериментальное моделирование с использованием Matlab 7.0. В качестве начальных данных для моделирования определены две выборки. Данные первой выборки сгенерированы искусственно, значения входных факторов имеют равномерное распределение, а результирующая характеристика получена по формуле $Y = 3X_1^2 - 2X_2 + 4X_1X_2 - 7\sin X_3$. Вторая

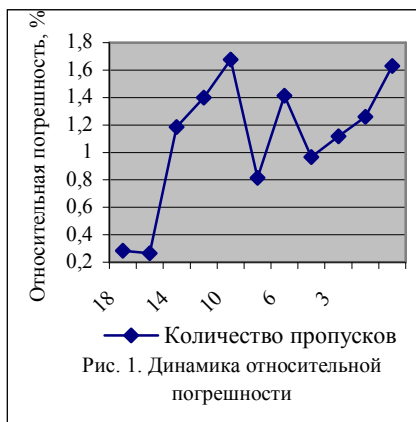
Сборник трудов VI-й Межд. конф. “Интеллектуальный анализ информации”. – Киев. – 2006. – С. 262-271.

выборка является официальной статистикой национального информационного центра энергетики США и содержит данные с 1949 до 2004 года [9], включающие производство твердого топлива, ядерной и другой энергии, импорт нефти и других энергоносителей, экспорт угля, газа, кокса и электроэнергии, потребление твердого топлива, ядерной и другой энергии, а также общее потребление.

Первая выборка насчитывала 25 образцов, из которых 20 отнесено в обучающую последовательность и 5 – в контрольную. Моделирование проводилось для разного количества пропусков при прочих равных условиях. Так, количество итераций обучения нейронной сети ограничено 50, а значение целевой функции составило 10. При моделировании установлено, что такая точность достигнута не была, и процесс обучения прекращался из-за ограничения на количество итераций. Результаты приведены в табл. 2, где N – количество пропусков, NP – процентное соотношения количества пропусков, F – значение целевой функции, Er – относительная погрешность (в процентах). Зависимость значения относительной погрешности от количества пропусков изображена на рис. 1.

Таблица 2. Данные экспериментов

N	NP	F	Er
18	30	1410	0,2841
16	27	1317	0,2657
14	23	257	1,1857
12	20	1157	1,3995
10	17	1056	1,6772
8	13	335	0,8149
6	10	205	1,4124
4	7	554	0,9669
3	5	138	1,1183
2	3	121	1,26
1	1,7	248	1,6291



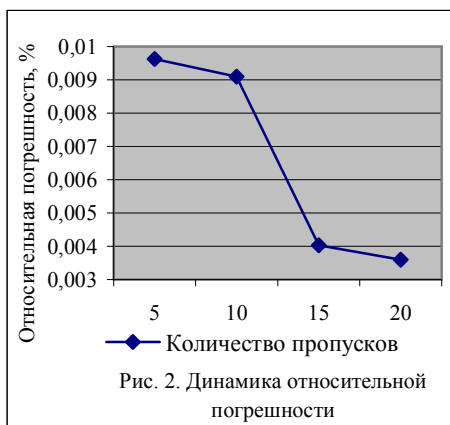
Статистическая информация, характеризующая состояние энергетики США, состояла из 11 входных факторов, одной результирующей характеристики и насчитывала 40 образцов. Из них 35 отнесено в обучающую последовательность и 5 – в контрольную. Количество итераций установлено 150, значение целевой функции – 1. На второй выборке итерации прекращались из-за достижения указанного значения ошибки. Максимальное количество итераций, в отличие от первого

случая, достигнуто не было. Результаты моделирования приведены в табл. 3 и на рис. 2.

Таблица 3. Данные экспериментов

N	NP	F	E _r
5	1,136364	0,93	0,00962
10	2,272727	0,88	0,0091
15	3,409091	0,39	0,00402
20	4,545455	0,35	0,0036

При моделировании для работы генетического алгоритма использовалась выборочная популяция из 20 элементов, количество эпох равнялось 100. Время моделирования на компьютере Intel Pentium M 2,0 ГГц составляло 30 минут и от количества пропусков не зависело. Полученные результаты свидетельствуют о достаточно высокой точности метода, которую можно повысить, если увеличить количество итераций обучения нейронной сети. Динамика относительной погрешности, приведенная на рис. 1 и рис. 2 указывает на то, что пластичность метода, или способность НС к обобщению возрастает при увеличении количества пропусков (до некоторого предела). Такая тенденция является необычной, ее объяснение требует дополнительных исследований.



Заключение

Разработанный эволюционный метод восстановления пропусков в данных имеет ряд преимуществ. Так, его использование не требует выполнения ограничений на исходную информацию. Таблица исходных данных может иметь произвольную размерность и структуру пропусков.

Перспективным представляется исследование эффективности использования НС с неитеративными алгоритмами обучения. Необходимо выяснить влияние распределения значений факторов на точность восстановления пропусков.

Как уже было указано выше, тенденция к увеличению точности идентификации с ростом количества пропусков также требует своего объяснения. С какой точностью возможно восстановления пропусков, если их количество составляет 50% всех значений в таблице данных? Каким условиям должны удовлетворять значения факторов, чтобы точ-

Сборник трудов VI-й Межд. конф. “Интеллектуальный анализ информации”. – Киев. – 2006. – С. 262-271.

ность результатов была максимальной? Ответы на эти вопросы позволяют сформировать методiku восстановления пропусков с использованием эволюционного подхода.

Литература

1. *Злоба Е., Яцкив И.* Статистические методы восстановления пропущенных данных // *Computer Modelling & New Technologies.* – 2002. – Vol. 6. – № 1. – Pp. 51-61.
2. *Хайкин С.* Нейронные сети: полный курс. – М.: “Вильямс”, 2006. – 1104 с.
3. Загоруйко Н.Г. Методы распознавания и их применение. – М.: Сов. радио, 1972. – 216 с.
4. *Россиев А.А.* Моделирование данных при помощи кривых для восстановления пробелов в данных. В кн. “Методы нейроинформатики” / Под ред. А.Н. Горбаня. – КГТУ: Красноярск, 1998. – С. 6-22.
5. *Колмогоров А.Н.* О представлении непрерывных функций нескольких переменных суперпозициями непрерывных функций меньшего числа переменных // Докл. АН СССР. – 1956. – Т. 108. – № 2. – С. 179-182.
6. *Арнольд В.И.* О функциях трех переменных // Докл. АН СССР. – 1957. – Т. 114. – № 4. – С. 679-681.
7. *Колмогоров А.Н.* О представлении непрерывных функций нескольких переменных в виде суперпозиции непрерывных функций одного переменного // Докл. АН СССР. – 1957. – Т. 114. – № 5. – С. 953-956.
8. *Harti R.E.* A global convergence proof for class of genetic algorithms.– Wien: Technische Universitaet, 1990. – 136 p.
9. Annual Energy Report 2004 / Energy Information Administration USA: Washington, 2004. – 435 p. <http://www.eia.doe.gov/aer>
10. *Люгер Ф. Дж.* Искусственный интеллект. Стратегии и методы решения сложных проблем. – М.: “Вильямс”, 2003. – 864 с.