УДК 681.513; 004.62

Е.А. Савченко

Предварительная обработка данных в задаче индуктивного моделирования

Решены задачи предварительной обработки выборки данных для повышения точности индуктивного моделирования – моделирования по выборке экспериментальных данных. Приведены задачи, возникающие при предварительной обработке данных, и способы их решения.

The problems of the preprocessing data sample improving the accuracy of the inductive modeling, i.e. the modeling by experimental data sample are solved. The tasks occurring while preprocessing of data and the ways to solve them are given.

Описано розв'язання задач попередньої обробки вибірки даних для підвищення точності моделювання – моделювання за вибіркою експериментальних даних. Подано задачі, що виникають при попередній обробці вибірки, та способи їх розв'язання.

Введение. Экспериментальные данные, предназначенные для моделирования, могут быть представлены в разных форматах, измеряться в разных диапазонах, содержать пропуски, выбросы и пр. Поэтому использование методов предварительной обработки выборки данных может быть не только средством повышения точности моделирования, но и обязательным этапом приведения выборки к виду, необходимому для дальнейшего использования. Выборка, содержащая пропуски в данных, не пригодна для дальнейшего моделирования.

Самоорганизация моделей с помощью индуктивного моделирования предъявляет определенные требования к выборке исходных данных. Во-первых, выборка должна содержать аргументы, влияющие на исходную переменную. Во-вторых, она должна быть полной, т.е. не содержать пропущенных значений. Если данные в выборке имеют разную размерность, желательно их нормировать.

Цель данной статьи – рассмотреть задачи, возникающие при предварительной обработке выборки данных и способы их решения.

Среди задач, возникающих при предварительной обработке выборки данных можно выделить задачи восстановления пропусков в данных, расширения выборки данных введением дополнительных переменных, оценки информативности переменных, оптимизации выборки данных.

Задача предварительной обработки выборки данных

Выделим следующие этапы предварительной обработки выборки исходных данных:

- выбор оптимального шаблона считывания данных;
- проверку наличия пропусков в выборке данных и их восстановление;
- нормирование (масштабирование) исходных данных;
- формирование (генерация) дополнительных аргументов;
- оценку и отбор информативных аргументов;
- оптимизацию размера выборки исходных данных путем поиска аналогов в предыстории.

Сначала выборка данных проверяется на наличие пропущенных значений. Для восстановления пропущенных данных в выборке применяется комбинаторный алгоритм МГУА с использованием различных шаблонов их считывания. По заданному шаблону формируется расширенная выборка данных, по которой строится лучшая модель и рассчитывается значение пропущенного элемента. Для каждого пропуска в данных рассчитывается собственная модель. Следующий этап – нормирование данных. Его целесообразно применять только тогда, когда данные имеют разную размерность или разные масштабы цифр. Затем формируются дополнительные аргументы. К входной выборке добавляются парные произведения исходных переменных, их запаздывания и др. Оценка информативности входных и сгенерированных переменных осуществляется расчетом значений модуля коэффициента корреляции каждой переменной с выходной величиной. Переменные с небольшим значением модуля могут быть исключены из множества эк-

вивалентных входных аргументов. Для оптимизации размера выборки рассчитывается значение евклидова расстояния вектор-строк таблицы данных и из всего множества наблюдений отбирается необходимое количество ближайших к исходному наблюдению.

На рис. 1 приведена блок-схема решения задачи предварительной обработки выборки данных.

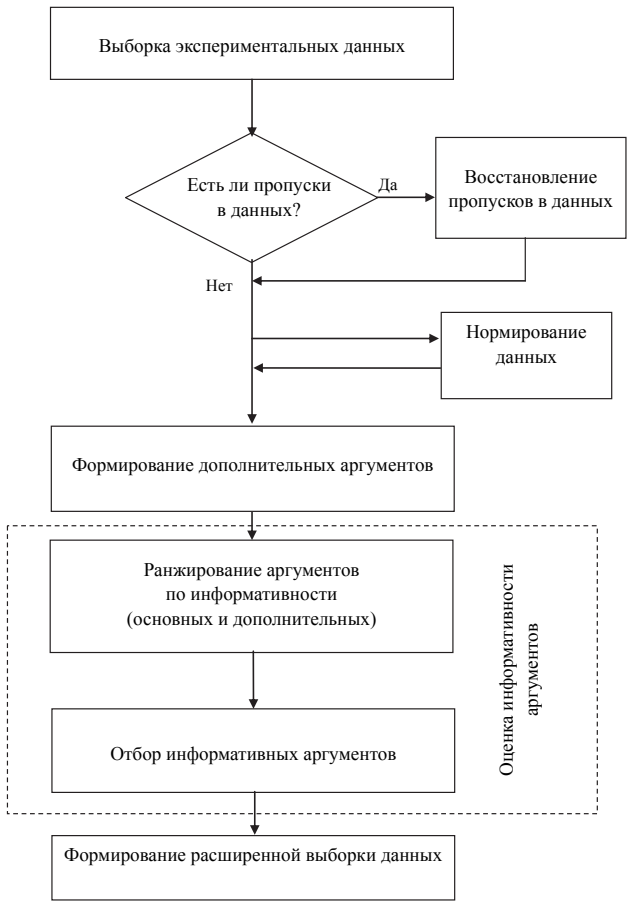
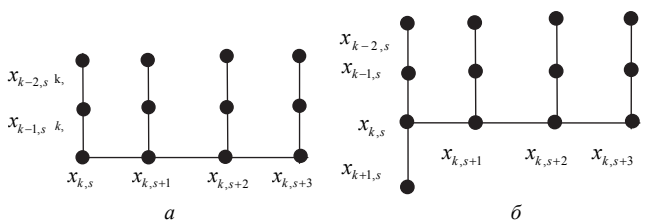


Рис. 1

Рассмотрим каждую из задач.

Восстановление пропущенных данных на основе МГУА с оптимизацией шаблона считывания данных

В теории разностных уравнений принято называть шаблоном односвязную геометрическую фигуру (граф) [1], который показывает, какие дискретные величины, измеренные в предыдущие моменты времени или в соседних точках пространства, влияют на значение выходной переменной в текущий момент времени. Тем самым определяется, какие элементы временного ряда входят в каждое очередное условное уравнения Гаусса. Каждому положению шаблона на выборке исходных данных соответствует одно условное уравнение. Шаблон перемещается сверху вниз по выборке (таблице) данных. Система условных уравнений преобразуется с помощью известной процедуры Гаусса в систему нормальных уравнений и используется затем для получения оценок коэффициентов моделей. Форма шаблона может быть различной и зависит от постановки задачи, например шаблоны для восстановления данных и для прогнозирования будут отличаться [2]. По способу получения уравнений из начальной выборки данных шаблоны могут быть явными и неявными. Модель зависимости выходной переменной от других переменных, полученная с использованием явных шаблонов – единая, т.е. получено одно уравнение, а с использованием неявных – система уравнений, поскольку выходными считаются все переменные поочередно. На рис. 2 приведен пример неявного *а*) и явного *б*) шаблонов.



а *б*

Рис. 2

Неявные шаблоны перемещаются на один шаг вперед одновременно, что образует совместную систему уравнений. Точность неявных шаблонов, так же как и устойчивость вычислений, выше явных, вследствие увеличения количества аргументов. Для пошагового прогноза явные шаблоны проще, поскольку для неявных приходится при расчете прогноза решать систему уравнений, матрица которых может быть плохо обусловленной. В общем случае для повышения точности следует отдавать предпочтение неявным шаблонам. Пошаговое перемещение шаблона позволяет получить выборку в виде временного ряда и превратить его

в расширенную выборку, каждая строка которой содержит все необходимые данные для составления условных уравнений Гаусса. Поскольку для восстановления данных форма каждого пропуска может быть иной, то и форма шаблонов для восстановления различных пропусков в данных будет отличаться. В [2] предложено три шаблона для восстановления пропусков в данных: крестообразный, диагональный и квадратный, позволяющие восстанавливать пропуски в данных различной формы, и приведены примеры восстановления пропусков в задаче медицинского мониторинга [2].

Нормирование выборки данных

Если данные в выборке имеют разную размерность, рекомендуется применять нормирование данных. Это необходимо для адекватного использования математических методов и применения компьютерных средств при вычислениях, связанных с большими и малыми величинами, а также для соответствия между количественными и качественными характеристиками данных. Нормирование – существенный фактор, влияющий на точность моделирования. Некорректный подход к нормированию может ухудшить точность моделей. Данные преобразуются к виду, удобному для сравнительного анализа так, чтобы каждое значение, поступающее на вход алгоритма, принадлежало интервалу [0,1]. Приведем основные формулы, используемые для нормирования [4]. Самый простой способ нормирования – по максимальному значению:

$$x_i = \frac{x_i}{x_{\max}}, \qquad (1)$$

где x_i – текущее значение аргумента; $x_{\max}$ – максимальное значение аргумента.

Следующая формула позволяет разместить данные на интервале от нуля до единицы:

$$x_i = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}} \quad \text{или} \quad x_i = \frac{x_{\max} - x_i}{x_{\max} - x_{\min}}, \qquad (2)$$

где $x_{\min}$ и $x_{\max}$ – соответственно минимальное и максимальное значения переменной. Оптимальным будет случай, когда значение данных равномерно заполняют интервал [0,1]. При решении задачи прогнозирования нормирование может быть неэффективным, когда нормированные значения будут нулевыми или сосредоточены у концов интервала. Нормированные по следующей формуле значения находятся в окрестности нуля, но не обязательно относятся к заданному отрезку:

$$x_i = \frac{x_i - \bar{x}}{\delta}, \qquad (3)$$

где $\bar{x}$ – выборочное среднее значение, δ – выборочное стандартное отклонение. При применении формулы (3) могут возникнуть проблемы из-за неопределенности границ отрезка изменения значений. Тогда необходимы дополнительные преобразования, гарантирующие более равномерное распределение значений, например:

$$x_i \to \frac{x_i - \bar{x}}{\delta} \to \frac{1}{1 + e^{-(\frac{x-\bar{x}}{\delta})}} \qquad (4)$$

Нормирование по следующей формуле используют нечасто, в основном для преобразования отрицательных значений в положительные:

$$x_i = \frac{1}{1 + e^{-x}}. \qquad (5)$$

Область ее значений – [1, + ∞]. Эта функция – вспомогательная, поскольку не избавляет переменную от размерности. Следует отметить, что использование функций нормирования, как правило, отражает входные значения в единичном гиперкубе. Выбор способа нормирования данных зависит от значений самой выборки и диапазона, в котором должны лежать нормированные значения. В статье использовано нормирование по формуле (1) как наиболее применимое для решаемого класса задач.

Расширение состава аргументов в результате формирования дополнительных переменных

После применения процедуры нормирования данных для повышения точности решаемой задачи множество входных аргументов может быть расширено путем введения дополнительных [5]. Основными, или первичными аргументами моделей называются переменные,

указанные в выборке экспериментальных данных. Дополнительными, или вторичными называются переменные, значения которых рассчитываются с помощью простых функциональных преобразований основных аргументов. Введение дополнительных аргументов расширяет область перебора при поиске оптимальной модели и может давать уменьшение значения ошибки и несмещенности модели. Наиболее распространены мультипликативные и аддитивные дополнительные аргументы. Можно предложить различные способы генерации дополнительных аргументов–кандидатов. Например, могут быть использованы: парные ковариации основных аргументов (произведение значений); суммы основных аргументов; обратные аргументы; парные ковариации обратных аргументов; суммы обратных аргументов; координаты первых аналогов; парные ковариации координат первых аналогов; выходные оценки переменных, полученных по алгоритмам МГУА, и др.

Расширение множества аргументов–кандидатов путем добавления новых аргументов служит одним из способов повышения точности и несмещенности модели. Снижение ошибки модели достигается в случае, когда среди дополнительных аргументов находятся информативные аргументы, и введение их в начальную выборку улучшает модель. Критерием информативности может служить уменьшение ошибки модели. Примеры улучшения точности моделей в результате введения дополнительных аргументов показаны в [6].

Входная выборка данных, как правило, – это временной ряд данных. Если в каждую строку такой выборки ввести предыдущие значения переменных как новые независимые аргументы, то получим расширенную выборку, которую можно использовать при составлении условных и нормальных уравнений Гаусса в методе наименьших квадратов. В такой выборке, согласно правилам линейной алгебры, можно свободно менять порядок следования строк, исключать некоторые строки и пр. После расширения выборки данных задачи выявления зависимости и прогнозирования случайных процессов решаются по одному общему алгоритму, в основе которого лежит комбинаторный алгоритм МГУА. Отличие этих задач заключается главным образом в выборе исходных переменных и координат пространства моделирования. Как показано в примере, в моделях для восстановления пропусков координатами пространства моделирования, кроме текущих и прошлых значений переменных, служат также их будущие значения [2].

Выбор информативных аргументов

При поиске оптимальной модели существенная роль принадлежит анализу информативности входных переменных. Все аргументы, независимо от способа их получения (основные или дополнительные), ранжируются по критерию информативности. В качестве критерия информативности аргументов–кандидатов может рассматриваться величина модуля коэффициента корреляции с выходной переменной [7]. Для отбора информативных аргументов можно предложить несколько способов.

1. Для непрерывных входных переменных критерием эффективности может служить модуль коэффициента корреляции оцениваемой переменной с выходной величиной. Автор моделирования, пользуясь своим опытом, может задать некоторый порог, и аргументы, информативность которых меньше заданного предельного значения, при построении модели не учитываются [7]. Может быть задано определенное количество аргументов, отобранных из ранжированных по модулю коэффициента корреляции аргументов. Наиболее информативные аргументы как основные, так и дополнительные используются для дальнейшего моделирования по комбинаторному алгоритму МГУА.

2. Для того чтобы время счета по комбинаторному алгоритму МГУА не превысило практически допустимой величины, из всего множества основных и дополнительных аргументов выбирается 20–25 переменных. Для их отбора в [8] предложен способ фильтрации групп информативных аргументов по комбинаторному алгоритму МГУА. Сначала ранжированный ряд основных и дополнительных аргументов делится на подмножества аргументов, которые

не пересекаются и каждое из которых содержит не более 20–25 аргументов. Для каждого подмножества по комбинаторному алгоритму МГУА ищут собственную модель. Все аргументы, вошедшие в полученные лучшие модели, объединяются и определяют собой пространство моделирования. В работе [18] предложен подобный способ: ранжированные по модулю коэффициента корреляции аргументы формируют два множества; первое содержит около 25 аргументов, второе – все остальные. По комбинаторному алгоритму МГУА строится оптимальная модель для первого подмножества, и аргументы, не вошедшие в эту модель, из подмножества удаляются, выборка данных дополняется до 25 переменных из второго множества аргументами с наибольшим значением модуля коэффициента корреляции, и вновь строится модель. Алгоритм продолжается до тех пор, пока не будут перебраны все аргументы входной выборки.

3. Для отбора информативных аргументов в [9] предложен новый весовой критерий, который учитывает степень влияния каждого аргумента на выходную величину, т.е. вес аргумента в каждой модели, куда он входит. Тестовые примеры показали, что даже на ненормированной выборке данных весовой критерий может выделить заданные истинные аргументы.

Оптимизация размера выборки данных на основе метода аналогов

Для поиска аналогов в предыстории процесса предложен метод аналогов [10], который заключается в отборе из всего множества наблюдений тех, которые будут ближайшими к текущему наблюдению в пространстве всех переменных. Расстояние между наблюдениями рассчитывается с помощью евклидовой метрики:

$$L_{ij}^2 = (x_{1,i} - x_{1,j})^2 + \\ + (x_{2,i} - x_{2,j})^2 + \ldots + (x_{m,i} - x_{m,j})^2, \quad (6)$$

где m – число аргументов, i и j изменяются от единицы до n. Все наблюдения ранжируются по расстоянию к выходному наблюдению. Наблюдение с наименьшим значением расстояния – первый аналог данного, следующее – вторым и т.д. Таким образом, из всего множества наблюдений будет отобрано некоторое количество аналогов выходного наблюдения. Выборка, состоящая из аналогов выходного наблюдения, используется при построении модели. Поскольку не всегда чем больше данных, тем лучше результат, исключение некоторых наблюдений при моделировании может улучшить его результат. Поэтому с помощью метода аналогов путем перебора различных вариантов может быть отобрано множество наблюдений, по которым результат будет самым лучшим. Например, при моделировании уровня глюкозы в крови по данным надомного мониторинга диабета, отбор из 60 наблюдений 40 ближайших к текущему наблюдению аналогов существенно улучшил прогноз на сутки вперед [11]. Таким образом, метод аналогов может быть применен для уменьшения количества наблюдений в выборке данных, т.е. оптимизации ее.

Заключение. В результате исследования решены задачи обработки данных, которые могут служить средством повышения точности моделирования и прогнозирования. Применение индуктивного подхода целесообразно на этапе построения модели для восстановления пропусков в данных и отбора информативных аргументов.

1. *Справочник* по типовым программам моделирования. – Киев: Техніка, 1980. – 184 с.
2. *Применение* алгоритмов МГУА для восстановления пропущенных данных и прогноза уровня глюкозы в крови при надомном мониторинге диабета / А.Г. Ивахненко, Е.А. Савченко, Г.А. Ивахненко и др. // Проблемы управления и информатики. – 2002. – № 3. – С. 123–133.
3. *Ivakhnenko A.G.*, *Ivakhnenko G.A*. Problems of Further Development of the Group Data Handling Algorithms. Part I // Pattern Recognition and Image Analysis. – 2000. – **10**, N 2. – P. 187–194.
4. *Снитюк В.Е.* Предварительная обработка данных. – *http://www.artint.com.ua*
5. *Ивахненко А.Г.*, *Ивахненко Г.А.*, *Савченко Е.А.* Концепция последовательных алгоритмических приближений (спусков) к точному решению интерполяционных задач искусственного интеллекта // КВТ. – 2000. – № 127. – С. 47–58.

6. *Обнаружение* закономерностей взаимодействия ионов с поверхностью по комбинаторному алгоритму МГУА / А.Г. Ивахненко, Е.А. Савченко, Г.А. Ивахненко и др. / Проблемы управления и информатики. – 2003. – № 2. – С. 80–89.
7. *Круг Г.М.*, *Круг О.Ю.* Математический метод классификации древней керамики // Тр. ин-та археологии АН СССР. – М.: Наука, 1965. – С. 317–323.
8. *Кошулько О.А.*, *Кошулько Г.А.* Багатоетапний комбінаторний алгоритм МГУА для моделювання об'єктів з великою розмірністю // Матеріали міжнар. конф. «Інтелектуальні системи прийняття рішень та проблеми обчислювального інтелекту» (ISDMCI'2009), Євпаторія. – Херсон: Ви-во ХНТУ, 2009. – **2**. – С. 331–332.
9. *Самойленко А.А*, *Степашко В.С.* Метод последовательного отсеивания неинформативных аргументов для эффективного решения переборных задач индуктивного моделирования // УСиМ. – 2013. – № 2. – С. 33–39, 46.
10. *Ивахненко Г.А.* Алгоритм комплексирования аналогов для самоорганизации дважды многорядных нейронных сетей // УСИМ. – 2003. – № 3. – С. 15–20.
11. *Савченко Е.А.* Экспресс-прогноз уровня глюкозы в крови с учетом аналоговых и временных характеристик // УСиМ. – 2003. – № 2. – С. 107–112.

Поступила 03.04.2015
Тел. для справок: +38 044 526-3028 (Киев)
E-mail: *savchenko_e@meta.ua*

●

Внимание!

Доступен сайт журнала: *usim.irtc.org.ua*, на котором размещен архив журнала с 2009 года.

На сайте Национальной библиотеки Украины имени В.И. Вернадского в рубрике «Наукова періодика України» также доступен архив журнала с 2009 года. Все научные издания, представленные на этом ресурсе на новой платформе, будут корректно индексироваться поисковой системой *Google Scholar*.

Журнал представлен в научно-метрической базе (*http*://*www.elibrary.ru*). Научная электронная библиотека содержит Российский индекс научного цитирования (РИНЦ), электронные научные публикации, информационные базы данных, а также сервис индивидуальной подписки на электронные версии научных изданий.