УДК 681.5

А.Е. Архипов, Е.А. Бабенко

ПРИМЕНЕНИЕ РАНГОВЫХ СТАТИСТИК ПРИ ПОСТРОЕНИИ КРИТЕРИЯ СТРУКТУРНОЙ ИДЕНТИФИКАЦИИ АППРОКСИМАТИВНЫХ МОДЕЛЕЙ

Введение

Рассматривается задача структурно-параметрической идентификации аппроксимативных моделей в следующей постановке. Имеются данные наблюдений некоторого объекта, скалярный выходной сигнал y которого каким-то неизвестным образом зависит от векторного входного x. Данные представлены определённым количеством значений выходного сигнала при соответственных значениях входного — выборкой экспериментальных данных.

Требуется “восстановить” отображение, осуществляемое наблюдаемым объектом, а именно, построить аппроксимативную модель этого объекта.

Рассматривается такой случай, когда информация о характеристиках данных, в том числе статистических, может быть получена только из самих этих данных, и никаких других её источников нет.

Традиционно, в решении задачи идентификации аппроксимативных моделей выделяют три этапа: идентификация структуры, оценивание параметров, проверка адекватности полученной модели. Этим этапам, как правило, предшествует построение множества моделей-претендентов $\{\mu_s\}$, из которого выбирается наилучшая в том или ином смысле модель. Здесь предполагается, что такое множество уже построено. Первый из трёх перечисленных этапов гораздо менее изучен, чем последние два, поэтому излагаемые исследования посвящены именно ему.

Идентификация структуры модели

Будем считать, что каждая из моделей претендентов μ_s зависит от векторного параметра, размерность которого для разных моделей может быть разная, и при заданном значении параметра отображает пространство значений входной переменной в пространство значений выходной переменной.

Идеальным в определённом смысле был бы выбор структуры модели (и параметров тоже) по критерию минимума среднего риска. Но средний риск невозможно вычислить по конечной выборке экспериментальных данных. Из этой проблемы выросла теория минимизации среднего риска по экспериментальным данным (она изложена, например, в монографии [2]). Эта теория, в частности, вырабатывает практически вычислимые

показатели качества идентификации, минимизацией которых можно заменить минимизацию среднего риска. На этом пути в [1] и ряде других работ предлагается показатель вида

$$J_s = \gamma \frac{\hat{I}_s}{R^2} + (1-\gamma)\,\frac{1}{K}\sum_{k=1}^{K}\frac{\hat{D}\hat{a}_s^k}{(\hat{a}_s^k)^2},$$

где

$\hat{I}_s = \frac{1}{n}\sum\limits_{i=1}^{n}\left(y^i - \mu_s\left(\mathbf{x}^i,\ \mathbf{a}_s\right)\right)^2$ — эмпирический риск;

i — номер наблюдения;

$\mathbf{a}_s$ — значение параметра, минимизирующее эмпирический риск для модели μ_s;

γ — некоторое действительное число от 0 до 1;

$R = \max\left\{y^1,\ \ldots,\ y^n\right\} - \min\left\{y^1,\ \ldots,\ y^n\right\}$ — так называемый размах выборки;

K — размерность вектора параметров модели μ_s;

$\hat{D}\hat{a}_s^k$ — оценка дисперсии оценки $\hat{a}_s^k$ параметра a_s^k.

Данный показатель имеет эвристическую природу и обосновывается в [1] следующим образом. Если решаемая задача идентификации корректна (по Адамару: решение существует, единственно и устойчиво), то правильному выбору модели μ_s соответствует малое значение эмпирического риска $\hat{I}_s$ (высокая точность аппроксимации моделью экспериментальных данных), а малые вариации исходных данных приводят к малым вариациям оценок параметров (низкая дисперсия оценок параметров). Поэтому оба слагаемых показателя J_s сохраняются небольшими, значение показателя небольшое. Выбор модели слишком простой структуры будет сопровождаться ростом первого слагаемого, переусложнение модели повлечёт потерю устойчивости оценок параметров и повышение их разброса, что в обоих случаях приводит к росту значения показателя J_s. Таким образом, показатель J_s осуществляет компромиссный выбор лучшего варианта модели, адекватно учитывающего требования по точности приближения экспериментальных данных и устойчивости оценивания параметров модели.

Оценки дисперсии оценок параметров предлагается строить с помощью варьирования выборки, то есть искусственного построения псевдовыборок, статистически однородных с исходной выборкой, полученной в результате реального эксперимента. Существует множество методов варьирования выборки. Ниже используется метод варьирования строк матрицы (bootstrap в англоязычной литературе), который состоит в следующем: исходная выборка рассматривается как генеральная совокупность, из которой путём равновероятного выбора строк с возвращением формируется требуемое количество псевдовыборок. Случайный выбор очередного номера строки осуществляется генератором (псевдо)случайных чисел, равномерно распределённых на $\{1,\ \ldots,\ n\}$. В соответствии с очередным выпавшим номером строка за строкой формируются псевдовыбор-

 ISSN 1562-9945

ки. После построения L псевдовыборок оценка дисперсии произвольной статистики t, в данном случае — оценки параметра, строится следующим образом:

$$\hat{D}t = \frac{1}{L-1}\sum_{l=1}^{L}\left(t_l - \hat{M}t\right)^2,$$

где t_l — значение статистики t, вычисленное на l-й псевдовыборке,

$$\hat{M}t = \frac{1}{L}\sum_{l=1}^{L} t_l.$$

Показатель J_s хорошо зарекомендовал себя при решении практических задач. В [1] указывается, что ограничений в применении данного показателя не обнаружено, он может применяться при отборе моделей любого типа. Однако он содержит коэффициент γ, для выбора которого нет научно обоснованного метода. В зависимости от значения этого коэффициента минимум показателя будет приходиться на принципиально различные модели.

Выбирать коэффициент исключительно путём нормировки значений обоих слагаемых на заранее заданный диапазон (скажем, $[0,\ 1]$) по множеству моделей-претендентов нельзя, хотя такой способ и практикуется в литературе. Действительно, наличие “большого выброса” во множестве моделей-претендентов может привести к большому выбросу во множестве значений одного из слагаемых. В результате, после нормировки множество значений этого слагаемого будет состоять из одного значения на верхнем пределе (единице) и остальных значений около нижнего предела (нуля). Соответственно, вклад этого слагаемого в суммарное значение показателя будет около нулевой для всех моделей, кроме одной. Например, с первым слагаемым это произойдёт тогда, когда множество моделей-претендентов содержит модель с большой невязкой, а со вторым — когда присутствует модель с незначащим регрессором. Если множество моделей-претендентов генерируется автоматически по жёсткому заранее заданному правилу, то такую ситуацию будет встретить проще, чем нормальную ситуацию.

Чтобы эта методика была работоспособна, необходимо исключить большие выбросы во множестве моделей-претендентов, например, произведя предварительную селекцию моделей, или не допускать их попадания туда.

Поскольку последняя задача трудно разрешима при отсутствии дополнительных априорных сведений, предлагается иной подход к использованию показателя J_s, основанный на ранговых статистиках. Действительно, при выборе лучшей модели из конечного множества моделей-претендентов не нужно знать само значение показателя J_s, оно не имеет физического смысла, а нужно лишь упорядочить модели по значению этого показателя. Для этого предлагается ранжировать модели

отдельно по каждому из слагаемых, а затем взять “средний” ранг. Можно вычислить ранг модели по каждому из слагаемых, а затем взять, например, среднее арифметическое рангов. Эта методика также не имеет теоретического обоснования, но она, по крайней мере, снимает проблему соотнесения масштабов обоих слагаемых.

Для проверки адекватности предложенной методики проводится серия вычислительных экспериментов. Исследования проводятся по следующему плану:

1. Выбирается некоторая модель, которая считается истинной.
2. По истинной модели генерируется “естественная” выборка данных.
3. По “естественной” выборке данных производится идентификация аппроксимативной модели путём выбора наилучшей из множества моделей-претендентов.
4. Критерием пригодности процедуры идентификации считается совпадение структуры идентифицированной модели со структурой истинной модели.

Ранжирование моделей-претендентов производится после того, как для каждой из них решена задача параметрической идентификации. Параметры всех моделей-претендентов оцениваются по методу наименьших квадратов.

В качестве истинной модели взята линейная по параметрам модель, состоящая из 14 слагаемых. Слагаемые (регрессоры) представляют собой алгебраические выражения и элементарные функции. Истинная модель ориентирована на задачу структурно-параметрической идентификации модели процесса резания монолитными твердосплавными концевыми фрезами, которую автору предстояло решать. Значения факторов (независимых переменных), на основе которых вычисляются значения регрессоров, взяты из реального эксперимента. На истинные значения зависимой переменной накладывается аддитивный шум, который имеет независимые одинаково распределённые компоненты с нулевым средним. Дисперсия шума варьируется с целью определения пределов работоспособности исследуемой методики идентификации моделей.

Множество линейных по параметрам моделей-претендентов $\{\mu_1, \ldots, \mu_{17}\}$ содержит истинную модель μ_{14}. Сложность моделей (здесь — количество регрессоров) монотонно нарастает от модели μ_1 к модели μ_{17}, то есть модели $\mu_1, \ldots, \mu_{13}$ являются “недоусложнёнными”, μ_{14} — истинной, а μ_{15}, μ_{16}, μ_{17} — переусложнёнными.

Серия вычислительных экспериментов такой постановки показала, что средний ранг даёт истинную модель до тех пор, пока амплитуда шума не превышает примерно 25% от размаха значений зависимой переменной при условии, что регрессоры истинной модели являются равнозначимыми с точностью до 10% (по размаху значений).

На рисунках графически представлены результаты одного из таких вычислительных экспериментов, где номер ранга убывает с повышением качества модели. По графику на рис. 2 видно, что средний ранг достигает минимума на истинной модели μ_{14}.

 ISSN 1562-9945

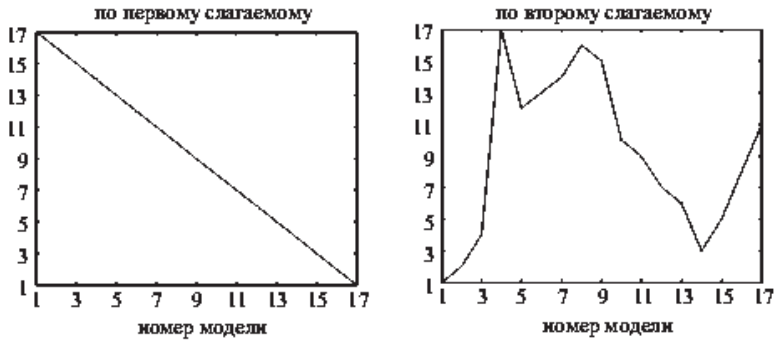

Рис. 1 – Ранжирование моделей по отдельным слагаемым

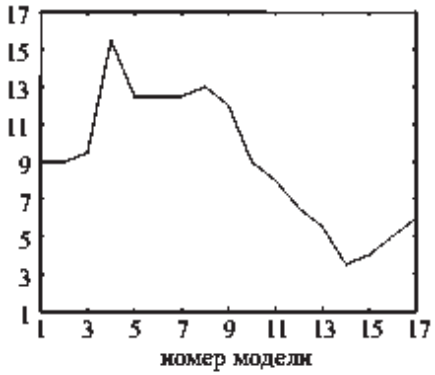

Рис. 2 – Средний ранг

Заключение

Полученные результаты подтверждают, что переход к ранговым статистикам в рассмотренном выше эмпирическом показателе качества идентификации позволяет снизить уровень искусственно вносимой информации в процедуру идентификации и получить удовлетворительные результаты. Свойства предложенной методики подлежат дальнейшим исследованиям.

Литература

1. Архипова С.А. Идентификация аппроксимативных моделей методом варьирования данных. — Дисс. на соискание учёной степени к.т.н. по специальности 05.13.03 "Системы и процессы управления". — Киев, 2000. — 216 с.
2. Вапник В.Н. Восстановление зависимостей по эмпирическим данным. — Москва: Наука, 1979. — 448 с.

 ISSN 1562-9945