УДК 519.25:681.3.03

А.П. Алпатов, В.И. Кузнецов, А.П. Сарычев

СТАТИСТИЧЕСКОЕ МОДЕЛИРОВАНИЕ ЗАВИСИМОСТЕЙ ЭНЕРГЕТИЧЕСКИХ ВОЗМОЖНОСТЕЙ И СТОИМОСТИ НОСИТЕЛЕЙ ОТ ИХ ТЕХНИЧЕСКИХ ХАРАКТЕРИСТИК

Неотъемлемой составной частью исследований перспектив развития сложных технических систем, в частности – космических, является прикладной статистический анализ, поскольку именно его методы, результаты и выводы используются для решения таких задач, как: а) прогнозирование ожидаемых значений отдельной переменной-характеристики проектируемого объекта как функции от ряда заданных характеристик (на основе построенной регрессионной статической зависимости); б) прогнозирование или хотя бы установление тенденции изменения переменной-характеристики, представленной временным рядом (на основе построенной авторегрессионной модели); в) отнесение отдельного, например, нового объекта к тому или иному классу существующих объектов (на основе решающего правила, построенного по обучающим выборкам). Одной из задач системных исследований космической техники является прогнозирование общих тенденций и перспектив развития. Необходимость прогноза остро ощущается как на этапах проектирования, проведения испытаний и модернизации технических объектов и систем, так и при разработке космических программ [1-4].

Эти задачи актуальны в исследованиях транспортных космических систем (ТКС). Основу современных ТКС составляют ракеты-носители космических аппаратов (КА). К настоящему времени сформировался мировой рынок услуг по выведению КА и обслуживающий его мировой парк ракет-носителей (РН). Основными "потребительскими" характеристиками РН являются энергетические возможности (масса выводимого на орбиту полезного груза) и стоимость пуска. В настоящей работе исследуются статистические зависимости между указанными "потребительскими" характеристиками и основными техническими характеристиками РН мирового парка.

 ISSN 1562-9945

1 Введение

Особенностью исследований транспортных космических систем является большое число учитываемых характеристик объектов и многообразие определяющих их внешних факторов, что, безусловно, усложняет решение перечисленных статистических задач. Это проявляется в отсутствии точных априорных гипотез об исследуемых объектах, что может характеризоваться следующими видами структурной неопределенности: 1) по числу однородных групп наблюдений, объективно существующих в исходной выборке данных; 2) по количеству и составу входных переменных в регрессионных моделях зависимостей выходных переменных от входных переменных 3) по степени статистической зависимости между аддитивными случайными составляющими в выходных переменных в системе регрессионных моделей; 4) по количеству и составу признаков в задачах статистической классификации.

Первоочередными задачами статистического анализа данных, проводимого в рамках прогнозирования тенденций и перспектив развития РКТ в условиях структурной неопределенности, являются задачи автоматической классификации (кластерного анализа). Общая постановка задачи автоматической классификации совокупности объектов состоит в требовании разбиения этой совокупности на некоторое число однородных в определённом смысле классов, причём число классов, вообще говоря, заранее неизвестно. Исходная информация о каждом объекте из классифицируемой совокупности представлена, как правило, значениями многомерного признака, а понятие однородности основано на предположении, что геометрическая близость в пространстве признаков двух объектов означает близость их "физических" состояний, их сходство.

Разработаны два метода оценивания энергетических возможностей ракет-носителей и стоимости пусков по их основным техническим характеристикам в условиях структурной неопределенности.

В первом методе результаты решения задачи кластеризации в пространстве входных переменных (технических характеристик) и результаты решения задачи кластеризации в пространстве выходных переменных использованы для прогнозирования выходных переменных. Этот метод может быть применен, например, для

получения оценок энергетических возможностей новых ракет-носителей по их заданным входным характеристикам в тех случаях, когда установление значений выходных переменных для этих новых объектов сопряжено с длительным временем или высокой стоимостью обследования, либо требует реального функционирования объекта. Другой пример применения этого метода – проверка правдоподобности заявленных энергетических возможностей и стоимости пусков новых ракет-носителей.

В основу второго метода положено построение системы регрессионных уравнений, описывающих зависимости выходных переменных объектов от множества входных переменных:

$$y(k) = f_k(X(k)) + \mathrm{o}(k),$$

где $y(k)$ – k-я выходная переменная; $k = 1,2,...,h$ – номер выходной переменной; h – их число; $X(k)$ – множество входных переменных для $y(k)$; $X(k)$ для разных $y(k)$ могут быть, вообще говоря, различными; $\mathrm{o}(k)$ – аддитивные составляющие выходных переменных. Как показано в [5], факт статистической зависимости между $\mathrm{o}(k)$ может быть использован при одновременном оценивании коэффициентов системы регрессионных моделей для выходных переменных. Учёт этой статистической зависимости позволяет точнее восстановить незашумленные ненаблюдаемые выходы объекта, т.е. позволяет улучшить качество системы регрессионных моделей.

Отыскиваемые во втором методе регрессионные зависимости для разных групп (классов, кластеров) ракет-носителей могут отличаться как по значениям коэффициентов, так и по структуре регрессионных моделей. Этот факт может быть использован для решения задачи статистической классификации (распознавания) нового объекта, т.е. для отнесения его к тому или иному известному классу [6].

2. Описание класса моделей

Задача оценивания коэффициентов в системе регрессионных моделей традиционно решается отдельно для каждой выходной переменной, и затем полученные модели лишь формально объединяются в систему. Если ошибки наблюдения выходных переменных моделируемого объекта статистически независимы, такое решение вполне оправдано. Интуитивно же ясно, что факт статистической зависимости между ошибками наблюдения выходных

переменных моделируемого объекта может быть использован для более точного восстановления незашумленных (ненаблюдаемых) значений выходов объекта. Вместе с тем эта возможность вовсе не очевидна, и, как известно (см., например [7, с. 488-491]), даже в случае шумов с произвольной (т.е., вообще говоря, недиагональной) ковариационной матрицей схема независимого оценивания коэффициентов регрессионных уравнений остаётся правильной. Но, оказывается, она правильна только при условии, что у всех выходных переменных общее множество регрессоров. Реальные объекты могут описываться системами регрессионных моделей более широкого класса, т.е. такими системами, в которых выходные переменные могут определяться, вообще говоря, разными множествами регрессоров. Оценивание коэффициентов именно для такого класса систем регрессионных моделей и рассматривается в данной работе.

Пусть модель функционирования исследуемого объекта имеет вид:

$$y(k) = \overset{o}{y}(k) + o(k) = \sum_{j=1}^{m(k)} \overset{o}{и}_j(k)\, \overset{o}{x}_j(k) + o(k), \quad k = 1,2,...,h, \qquad (1)$$

где k – номер выхода объекта; h – число выходов объекта; $y(k)$ и $\overset{o}{y}(k)$ – измеренный с ошибкой k-й выход объекта и незашумленный (ненаблюдаемый) k-й выход объекта соответственно; $o(k)$ – случайная ненаблюдаемая ошибка измерения k-го выхода объекта; $\overset{o}{x}_j(k)$ – j-й вход объекта из множества входов $\overset{o}{X}(k) \neq \varnothing$ ($\varnothing$ – пустое множество), участвующих в формировании k-го выхода объекта; $\overset{o}{\mathbf{и}}(k) = (\overset{o}{и}_1(k), \overset{o}{и}_2(k), ..., \overset{o}{и}_{m(k)}(k))^{\mathrm{T}}$ – вектор неизвестных, не равных нулю коэффициентов; $m(k)$ – число входов, принадлежащих множеству $\overset{o}{X}(k)$.

Будем считать, что в результате наблюдения объекта для каждого его k-го выхода получены: 1) $\overset{o}{\mathbf{X}}(k)$ – $(n \times m(k))$-матрица n наблюдений $m(k)$ входов множества $\overset{o}{X}(k)$, имеющая полный ранг, равный $m(k)$; 2) $\mathbf{y}(k)$ – $(n \times 1)$-вектор соответствующих наблюдений

выхода $y(k)$. Тогда в соответствии с моделью (1) для вектора наблюдений $\mathbf{y}(k)$ выполняется

$$\mathbf{y}(k) = \overset{\circ}{\mathbf{y}}(k) + \mathbf{o}(k) = \overset{\circ}{\mathbf{X}}(k)\overset{\circ}{\mathbf{и}}(k) + \mathbf{o}(k),\ k = 1,2,...,h, \qquad (2)$$

где $\overset{\circ}{\mathbf{y}}(k)$ – $(n\times 1)$-вектор значений незашумленного (ненаблюдаемого) k-го выхода объекта; $\mathbf{o}(k)$ – $(n\times 1)$-вектор случайных ненаблюдаемых ошибок измерения k-го выхода объекта.

Пусть относительно $\mathbf{o}(k),\ k = 1,2,...,h,$ выполнены предположения

$$E\{\mathbf{o}(k)\} = \mathbf{0}_n,\ \ E\{\mathbf{o}(k)\mathbf{o}^T(k)\} = \mathrm{y}(k)\ \mathbf{I}_n,\ \ k = 1,2,...,h\,; \qquad (3)$$

$$E\{\mathbf{o}(k)\mathbf{o}^T(q)\} = \mathrm{y}(k,q)\ \mathbf{I}_n,\ \ k,q = 1,2,...,h,\ \ k \neq q\,, \qquad (4)$$

где $E\{\cdot\}$ – знак математического ожидания по всем возможным реализациям случайных векторов $\mathbf{o}(k)$ и $\mathbf{o}(q)$; $\mathbf{0}_n$ – $(n\times 1)$-вектор, состоящий из нулей; $\mathrm{y}(k,k)$ – неизвестная конечная величина, дисперсия случайной величины $\mathrm{o}(k)$; $\mathrm{y}(k,q)$ – неизвестная конечная величина, ковариация случайных величин $\mathrm{o}(k)$ и $\mathrm{o}(q)$; $\mathbf{I}_n$ – единичная $(n\times n)$-матрица.

3. Решение практической задачи

Рассмотрим задачу оценивания энергетических возможностей и стоимости пусков ракет-носителей по их основным техническим характеристикам.

Множество выходных переменных образовано следующими характеристиками:

$y(1)$ – стоимость пусков (заявленная, млн. долл.);

$y(2)$ – максимальная масса полезного груза, выводимого на низкую круговую орбиту (Low Earth orbit, LEO-орбита, кг).

Множество входных переменных образовано такими характеристиками:

x_1 – время с первого пуска (в годах);

x_2 – стартовая масса ракеты-носителя (т);

x_3 – широта местности расположения космодрома (град.);

x_4 – стартовая тяга – тяга у земли первой ступени (кгс);

x_5 – удельный импульс у земли первой ступени (с);

x_6 – удельный импульс в пустоте верхней ступени (с).

 ISSN 1562-9945

Информация о ракетах-носителях мирового парка получена из базы данных, разработанной и пополняемой по открытым источникам в отделе системного анализа и проблем управления Института технической механики НАНУ и НКАУ. При моделировании использованы данные по 84 модификациям ракет-носителей.

Задача состоит в том, чтобы построить систему регрессионных уравнений, которая позволяла бы дать оценку стоимости пуска и максимальной массы полезного груза, выводимого на LEO-орбиту, по значениям входных переменных ракеты-носителя.

Предлагаемая схема оценивания энергетических возможностей ракет-носителей и их стоимости пусков состоит из двух этапов. Первый этап представляет собой решение двух задач кластеризации: 1) в пространстве входных переменных и 2) в пространстве выходных переменных. Результаты этих двух кластеризаций, проведенных независимо друг от друга, сравниваются по составу кластеров (по номерам попавших в них наблюдений): устанавливается соответствие между полученными кластерами в пространстве входных переменных X и в пространстве выходных переменных Y. Наблюдения, которые входят в один и тот же кластер в пространстве выходных переменных и, одновременно, в один и тот же кластер в пространстве входных переменных, объединяются в новые "общие" кластеры. Группа ракет-носителей, попавших в один кластер как в пространстве X, так и в пространстве Y, может претендовать на то, что их геометрическая близость означает близость их "физических" свойств (их сходство), и, можно ожидать, что зависимости выходных переменных от входных переменных для них будут одинаковы. И, следовательно, для такой группы ракет-носителей постановка задачи оценивания энергетических возможностей ракет-носителей и стоимости пусков по информации об их входных переменных является корректной.

Далее, на втором этапе схемы, возможны два варианта решения поставленной задачи оценивания. Первый из них основан на поиске так называемого наблюдения-аналога, т.е. такой ракеты-носителя, которая наиболее близка к исследуемой ракете-носителю в пространстве входных переменных. По этой схеме в качестве оценок значений выходных переменных исследуемой ракеты-носителя берутся значения выходных переменных у наблюдения-аналога.

Второй вариант основан на построении системы регрессионных моделей для выходных переменных в зависимости от входных переменных. Системы моделей строятся отдельно для каждого "общего кластера" по наблюдениям, вошедшим в данный "общий кластер". Далее эти системы моделей применяются для получения оценок значений выходных переменных исследуемой ракеты-носителя. Какую именно систему моделей применять к данной исследуемой ракете-носителю определяется тем, к какому кластеру в пространстве входных переменных она относится.

Решение задач кластеризации первого этапа в пространстве входных переменных и в пространстве выходных переменных проводилось следующим образом. Данные по каждой переменной предварительно центрировались относительно её выборочного среднего значения и нормировались на выборочное среднеквадратическое отклонение. В качестве меры близости между наблюдениями применялось расстояние Евклида. В качестве способа вычисления расстояния между подмножествами объектов и правила присоединения объекта выбран алгоритм "ближайшего соседа".

По данным о ракетах-носителях, вошедших в "общий" кластер, построена оптимальная система регрессионных моделей. Факт принадлежности этих ракет-носителей одному общему кластеру делает такую попытку построения вполне корректной.

Ввиду ограниченности выборки наблюдений ($n = 84$) в качестве критерия качества моделей применен усредненный критерий регулярности метода группового учета аргументов (критерий скользящего контроля), исследованный в [8].

Пусть матрица регрессоров $\mathbf{V}$ – $(n \times s)$-матрица наблюдений входов, принадлежащих множеству V, s – их число. Усредненным критерием регулярности в МГУА называется случайная величина

$$УКР = \sum_{i=1}^{n} (y_i - \mathbf{v}_i^{\mathrm{T}} \hat{\mathbf{d}}_{(i)})^2, \qquad (5)$$

где $\hat{\mathbf{d}}_{(i)}$ – МНК-оценка регрессии выходной переменной y на V, рассчитанная по выборке, которая получается из исходной выборки в результате исключения из нее наблюдения с номером i:

$$\hat{\mathbf{d}}_{(i)} = (\mathbf{V}_{(i)}^{\mathrm{T}} \mathbf{V}_{(i)})^{-1} \mathbf{V}_{(i)}^{\mathrm{T}} \mathbf{y}_{(i)}, \qquad (6)$$

 ISSN 1562-9945

т.е. в (6) для $((n-1)\times s)$-матрицы $\mathbf{V}_{(i)}$ и для $(n-1)$-вектора $\mathbf{y}_{(i)}$ выполняется

$$\mathbf{V}^{\mathrm{T}} = [\mathbf{V}_{(i)}^{\mathrm{T}} \mid \mathbf{v}_i], \quad \mathbf{y}^{\mathrm{T}} = (\mathbf{y}_{(i)}^{\mathrm{T}}, y_i), \tag{7}$$

где $\mathbf{V}_i$ – $(s\times 1)$-вектор значений входов, соответствующих y_i – наблюдению выхода с номером i (выполнение (7) для любого i можно обеспечить простой перестановкой столбцов матрицы $\mathbf{V}^{\mathrm{T}}$ и элементов вектора $\mathbf{y}$).

Поскольку поиск модели проводился в классе нелинейных моделей, входные и выходные переменные центрировались и нормировались по правилу $z_{норм} = (z-a)/b$, где a – среднее значение переменной, а b – ее среднеквадратичное отклонение (см. таблицу 1).

Таблица 1

Нормировочные коэффициенты для центрирования и нормирования входных и выходных переменных

Коэффициенты	x_1	x_2	x_3	x_4	x_5	x_6	$y(1)$	$y(2)$
центрирования (a)	14,024	337,43	30,821	231360,0	278,62	358,78	75,964	8085,7
нормирования (b)	11,102	252,41	14,303	219960,0	46,976	69,079	69,404	6641,6

Получена модель оптимальной сложности, которая имеет вид

$$\begin{gathered} y(1) = -0{,}254 + 0{,}269\,x_2^3 + 0{,}311\,x_6, \\ y(2) = 0{,}802\,x_2 + 0{,}339\,x_4\,x_5 + 0{,}375\,x_5 + 0{,}125\,x_6. \end{gathered} \tag{8}$$

Попытки усложнения этой системы двух регрессионных уравнений (по количеству и составу членов в модели) приводят к увеличению усредненного критерия регулярности для каждого регрессионного уравнения.

Качество построенных регрессионных уравнений можно охарактеризовать среднеквадратическим отклонением модельных значений от наблюдаемых значений выходных переменных:

$$\hat{\mathrm{y}}_k = \left(\sum_{i=1}^{n} (y_i(k) - \hat{y}_i(k))^2 / (n - m_k) \right)^{1/2}, \tag{9}$$

где $y_i(k)$ и $\hat{y}_i(k)$ – i-е наблюдаемое значение и i-е модельное значение k-й выходной переменной соответственно; i – номер наблюдения; k – номер выходной переменной; n – число наблюдений; m_k – число коэффициентов в модели для k-й переменной ($m_1 = 3$, $m_2 = 4$).

В полученной системе моделей значения этого показателя для выходных переменных соответственно равны: $\sigma_1 = 0,371; \quad \sigma_2 = 0,291$.

Как известно, статистики σ_1 и σ_2 в (9) являются заниженными оценками ожидаемых среднеквадратических ошибок моделей на новых данных. Свободными от этого недостатка являются значения, получаемые по усредненному критерию регулярности, равные 0,386 и 0,299 для первого и второго регрессионных уравнений соответственно.

Другим показателем качества регрессионных моделей является выборочный множественный коэффициент корреляции, который является обычным коэффициентом корреляции между наблюдаемыми и модельными значениями выходной переменной:

$$R_k = \left(\sum_{i=1}^{n} \left(y_i(k) - \bar{y}(k) \right) \left(\hat{y}_i(k) - \bar{\hat{y}}(k) \right) \right) \Big/ \left(\sum_{i=1}^{n} \left(y_i(k) - \bar{y}(k) \right)^2 \sum_{i=1}^{n} \left(\hat{y}_i(k) - \bar{\hat{y}}(k) \right)^2 \right)^{1/2},$$

где $\bar{y}(k)$ и $\bar{\hat{y}}(k)$ – средние значения наблюдаемых и модельных значений k-й выходной переменной соответственно.

В полученной системе моделей значения множественного коэффициента корреляции для двух выходных переменных равны $R_1 = 0,93$ и $R_2 = 0,96$.

По всем показателям качества построенные модели могут быть признаны хорошими и могут быть рекомендованы для оценивания энергетических возможностей и стоимости пусков ракет-носителей по их входным характеристикам. Для этого достаточно в полученные модели подставить значения входных переменных исследуемой ракеты-носителя. Поскольку входные и выходные переменные моделей предварительно центрировались и нормировались, то порядок применения моделей должен быть следующим: а) нормировать и центрировать значения входных переменных, воспользовавшись таблицей нормировочных коэффициентов; б) подставить нормированные и центрированные значения входных переменных в модель; в) от полученных модельных значений перейти к ненормированным и нецентрированным прогнозным значениям выходных переменных, снова воспользовавшись таблицей нормировочных коэффициентов.

 ISSN 1562-9945

4. Выводы

Разработаны два метода оценивания энергетических возможностей ракет-носителей и стоимости пусков по их основным техническим характеристикам на основе кластеризации и системного регрессионного анализа. Решена практическая задача оценивания энергетических возможностей ракет-носителей и стоимости пусков на основе информации из базы данных отдела системного анализа и проблем управления Института технической механики НАНУ и НКАУ.

ЛИТЕРАТУРА

1. Алпатов А.П., Камелин А.Б., Кунцевич В.М., Черемных О.К. Перспективные научные космические исследования в Украине // Сборник трудов Первой украинской конференции по перспективным космическим исследованиям (Киев, 8-10 октября 2001 г.). – Киев: 2001. – С. 5–10.
2. Алпатов А.П., Антоненко М.Е., Визер Т.Ф., Головин Ю.Н., Дорошкевич В.К., Иванов В.И., Ковалев Б.А., Мостипан И.Ф., Орешкин В.И., Сазина Н.П. Анализ причин переносов сроков пусков. Информационно-аналитический бюллетень "Ракетная и космическая техника. Транспортные космические системы": Препринт. НАНУ и НКАУ. Институт технической механики; 2. Днепропетровск: 2003. – 9 с.
3. Будник В.С., Дорошкевич В.К., Ковалев Б.А., Кузнецов В.И. Развитие системного подхода в исследованиях объектов ракетно-космической техники // Техническая механика. – 2001. – № 2. – С. 122–133.
4. Гусынин В.П., Гольдштейн Ю.А., Дорошкевич В.К., Кузнецов В.И., Кучугурный Ю.П. Многокритериальный сравнительный анализ объектов ракетно-космической техники // Космічна наука і технологія. – 2005. – Т. 11. – № 1/2. – С. 3–9.
5. Сарычев А.П. Оценивание коэффициентов в системах регрессионных моделей // Проблемы управления и информатики. – 2003. – № 4. – С. 74–82.
6. Сарычев А.П. Классификация объектов наблюдений, описываемых системами регрессионных уравнений с детерминированными коэффициентами // Штучний інтелект. – 2005. – № 3. – С. 43–56.
7. Рао С.Р. Линейные статистические методы и их применения.– М.: Наука, 1968. – 548 с.
8. Сарычев А.П. Усредненный критерий регулярности метода группового учета аргументов в задаче поиска наилучшей регрессии // Автоматика. – 1990. – № 5. – С. 28–33.

Получено 31.01.07