УДК 519.81

А.Н. ВОРОНИН

МЕТОДЫ КОМПЛЕКСИРОВАНИЯ ДАННЫХ

Аннотация. Рассмотрены методы комплексирования данных, позволяющие при ограниченном числе каналов получать максимально возможное количество доступной информации. Наряду с концепцией редукторов степеней свободы используется подход дискриминаторов степеней свободы, что дает возможность всем каналам в меру их информативности в текущей ситуации принимать участие в выработке кооперативного решения.

Ключевые слова: синергетика, каналы передачи данных, степени свободы, редуктор, дискриминатор, математическая статистика, байесовский подход, малая выборка.

СОСТОЯНИЕ ПРОБЛЕМЫ

В развитых информационных системах данные, характеризующие состояние x одного и того же объекта (процесса) О, передаются по нескольким каналам: $1, 2, \ldots, n$. Проблема состоит в определении относительной степени достоверности данных, поступающих по каждому из n каналов в текущий момент времени, и в выработке посредством механизма комплексирования К наиболее достоверной оценки x^* истинной характеристики x объекта (процесса) по имеющейся совокупности данных. На рис. 1 приведена схема компенсирования данных.

Рассмотрим примеры определения достоверности данных.

Пример 1. Оценка истинной высоты летящего самолета по нескольким высотомерам: барометрическому, бортовому радиолокационному, наземному радиолокационному и визуальному определению. Данные о высоте, поступающие по нескольким каналам, комплексируются и вырабатывается наиболее достоверная в определенном смысле оценка истинной величины высоты самолета.

Пример 2. Комплексирование данных экспертных оценок с учетом степени компетентности экспертов в рассматриваемом вопросе. Метод экспертного оценивания заключается в том, что для оценки некоторой количественной характеристики используются постулаты не одного, а нескольких лиц (экспертов), компетентных в данном вопросе. Предполагается, что истинное значение неизвестной количественной характеристики находится в диапазоне оценок экспертов и «обобщенное» коллективное мнение является более достоверным.

Пример 3. Комплексирование показаний приборов, имеющих различный класс точности (задача поставлена в [1]). При этом каждый прибор вносит свой вклад в результирующее показание в соответствии со своим классом точности.

Пример 4. Комплексирование сигналов для бистатической радиолокации малых небесных тел (метод описан в [2]). Для повышения точности измерений при исследовании параметров движения малых небесных тел используется бистатическая конфигурация радиолокационных систем. Информация от каждой из приемных антенн, разнесенных на значительные расстояния, подвергается обработке и сопоставлению их между собой так, чтобы результирующий сигнал был наиболее достоверным.

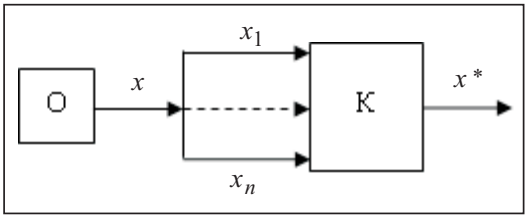

Рис. 1

В приведенных примерах для комплексирования данных рассматривается неизвестная количест-

венная характеристика как случайная величина, отражением закона распределения которой является сигнал данного канала. Для установления окончательной оценки данные всех каналов изучаются в совокупности и обрабатываются как некий исходный статистический материал. Обработка осуществляется с привлечением концепций математической статистики.

В обычной практике количество каналов получения комплексируемых данных относительно невелико, аналогично как и в случае малой выборки в задачах математической статистики [3, 4]. Таким образом, математическая модель рассматриваемого процесса представляется как оценка параметра распределения вероятностей случайной величины на основе выборки ограниченного объема. Поскольку в решении таких задач накоплен значительный опыт, дальнейшее исследование проведем в терминах математической статистики.

СОДЕРЖАТЕЛЬНЫЙ АНАЛИЗ ЗАДАЧИ

Оценки параметров распределения вероятностей случайной величины определяются на основе обработки статистического материала, представляющего собой совокупность экспериментальных значений изучаемой случайной величины.

При решении задач математической статистики в распоряжении исследователя всегда имеется лишь ограниченный статистический материал (выборка из генеральной совокупности), а оценка параметров требует распределения с возможно большей точностью. Это объясняется тем, что получение каждого нового элемента выборки — обычно сложный процесс, сопряженный со значительными трудностями технического или экономического характера. Трудности, связанные с вычислением, играют менее существенную роль. В подобных случаях О.К. Антонов говорил [5], что экономить на расчетах, оценивающих громадные экономические мероприятия, все равно, что экономить на прицеливании при выстреле.

Поэтому возникает задача: максимально использовать информацию о статистических свойствах изучаемой случайной величины и получить расчетные алгоритмы для вычисления уточненных оценок параметров распределения на основе статистического материала ограниченного объема. Так как результаты данного исследования могут быть применены не только для повышения информативности каналов получения данных в сложных информационных системах, но и в других случаях, то целесообразно формулировать и решать задачу в общих терминах математической статистики.

ПОСТАНОВКА ЗАДАЧИ

Рассмотрим непрерывную действительную случайную величину X, плотность распределения вероятностей которой $f(x|\theta)$ известна с точностью до неизвестного параметра θ. Задана совокупность из n независимых реализаций случайной величины X:

$$\bar{x} = \bar{x}^{(n)} = (x_1, x_2, \ldots, x_n), \tag{1}$$

где $x_1, x_2, \ldots, x_n$ — элементы выборки. Выборка (1) интерпретирует совокупность подлежащих комплексированию данных, полученную по n каналам.

Задача. По результатам случайной выборки (1) определить наилучшую в некотором смысле оценку θ^* неизвестного параметра θ распределения случайной величины X.

Так, если случайная величина X распределена нормально с известной дисперсией σ^2, то

$$f(x|\theta) = f(x|m_x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(m_x - x)^2}{2\sigma^2}\right] \tag{2}$$

и параметром, подлежащим оценке, является математическое ожидание: $\theta = m_x$.

Согласно [6] качество статистических оценок характеризуется следующими основными свойствами:

1) состоятельностью (сходимость по вероятности оценки к истинному значению параметра);

2) несмещенностью (отсутствие систематической погрешности оценки);

3) эффективностью (наименьшая дисперсия оценки).

Если в распоряжении исследователя имеется только та информация, которая содержится в изложенной выше постановке задачи, то для определения наилучших (состоятельных, несмещенных и эффективных) оценок используется предложенный Р. Фишером метод максимального правдоподобия [7]. Тогда оценка математического ожидания нормально распределенной случайной величины имеет вид

$$\theta^* = X_{\text{с}} = \frac{1}{n}\sum_{i=1}^{n} x_i, \tag{3}$$

где $X_{\text{с}}$ — средняя величина выборки (1), представляющая оценку математического ожидания m_x.

Часто для выборок малого объема оценки максимального правдоподобия типа (3) не обеспечивается удовлетворительная точность [8]. Это требует в конкретных приложениях разработки более эффективных процедур оценивания [3, 8]. Все они в какой-то степени связаны с привлечением дополнительной информации о статистических свойствах изучаемой случайной величины.

МЕТОД РЕШЕНИЯ

Действенным средством повышения эффективности статистического оценивания является байесовский подход [3]. Он заключается в том, что подлежащий оценке параметр θ рассматривается как реализованное значение случайной величины Θ. Всю доступную ему предварительную (до проведения экспериментов) информацию исследователь облекает в форму априорного распределения величины Θ, характеризующегося априорной плотностью распределения вероятностей $f_{\text{a}}(\theta)$. Эта функция считается известной до начала анализа данных, полученных экспериментально. Теорема Байеса комбинирует априорное распределение и данные наблюдений так, чтобы образовалось апостериорное распределение $f(\theta|x)$.

Физический смысл теоремы Байеса заключается в том, что если $f_{\text{a}}(\theta)$ — априорная плотность распределения вероятностей, приписываемая параметру θ перед проведением экспериментов, то $f(\theta|x)$ будет апостериорной плотностью, которую следует приписать θ после получения данных. Статистические оценки, вычисленные на основе апостериорного распределения, обладают лучшим качеством, чем оценки максимального правдоподобия, так как используют дополнительную информацию о неизвестном параметре θ в виде априорного распределения $f_{\text{a}}(\theta)$.

Наиболее значимым моментом при байесовском оценивании параметров является назначение функции априорной плотности. Функция должна соответствовать имеющейся предварительной информации. С одной стороны, вносятся только сведения, которые в априорных данных. Поэтому $f_{\text{a}}(\theta)$ выбирают, исходя из требования наличия максимально возможной энтропии (в шенноновском смысле [9]) при заданных условиях в виде конкретных априорных данных, рассматриваемых как ограничения [8]. С другой стороны, пренебрежение какой-либо объективной априорной информацией приводит к выбору менее информативной априорной плотности, что делает статистическую оценку менее эффективной.

Предлагается вводить априорную плотность с определяемым параметром, рассматриваемым как неизвестная константа. Такая априорная плотность вносит только те сведения, которые есть в априорных данных, и в то же время позволяет

использовать объективную априорную информацию о виде закона распределения оцениваемого параметра.

Найдем алгоритм вычисления уточненной оценки параметра θ (выкладки приведены в работе [1]):

$$\theta^* = \frac{\sum_{i=1}^{n} x_i f(x_i|\theta^*) f_{\mathrm{a}}(x_i|\theta^*)}{\sum_{i=1}^{n} f(x_i|\theta^*) f_{\mathrm{a}}(x_i|\theta^*)}. \qquad (4)$$

Отметим, что существует зависимость $\theta^* = \varphi(x_1, x_2, \ldots, x_n; \theta^*)$.

Как известно [10], уравнение в таком виде можно решать итерационным методом. Итерационная процедура следует из рекуррентной формулы

$$\theta^*[l] = \varphi(x_1, x_2, \ldots, x_n;\ \theta^*[l-1]),\ \ l \in [1, L],$$

где l — номер текущей итерации, L — число итераций.

Итерационный процесс заканчивается при выполнении условия $\theta^*[L] - \theta^*[L-1] \le \lambda_\theta$, где λ_θ — заданная точность вычисления оценки θ^*. Для анализа вопросов сходимости применяют известную теорему [10], в соответствии с которой для сходимости итерационного процесса достаточно на рассматриваемом интервале уточнения оценки θ^* соблюдать неравенство

$$|d\varphi(x_1, x_2, \ldots, x_n; \theta^*) / d\theta^*| < 1.$$

В сложных синергетических системах информация об одном и том же процессе (объекте, событии) обычно передается по нескольким каналам. Проблема состоит в том, чтобы определить, по каким каналам передаются более достоверные данные. В зависимости от этого требуется объединить (комплексировать) получаемые данные для выработки кооперативного решения о состоянии объекта. Традиционный путь предполагает выделение одного или нескольких наиболее информативных (доминирующих) каналов и отсечение остальных. Это осуществляется посредством механизма «редукторов степеней свободы» [11]. Преимущество такого способа в простоте, и часто он физически оправдан. Однако некоторые полезные нюансы относительно информации отсеченных каналов не участвуют в процессе выработки кооперативного решения.

Еще древние китайские мыслители считали, что любой выбор одного варианта из нескольких является ущербным, поскольку отвергает все остальные. Стержневым понятием китайской культуры всегда был выбор, при котором возможные ветви развития не отсекались, а срастались в единое целое, что признавалось единственно правильным решением.

Во многих случаях при синтезе синергетической системы комплексирования данных представляется целесообразным отказаться от концепции доминанты и вместо редукторов степеней свободы включать механизмы, позволяющие всем каналам получения данных участвовать в процессе формирования решения с весами, соответствующими степени их информативности в текущей ситуации (дискриминаторы степеней свободы). В результате вся доступная информация будет использована надлежащим образом.

Синергетический принцип комплексирования данных имеет много общего с идеями математической статистики [3]. Если синергетическая концепция слияния данных применяется для оценки характеристик процессов (объектов) по имеющейся совокупности данных, то математическая статистика изучает методы

оценки моментов распределения случайных величин по имеющейся совокупности элементов выборки. Общность проблем обеих теорий делает задачу исследования синергетических аспектов математической статистики актуальной как для синергетики, так и для развития статистических методов.

Рассмотрим задачу уточненного статистического оценивания математического ожидания m_x случайной величины X, распределенной по нормальному закону с плотностью распределения $f(x|\theta) = f(x|m_x)$, заданной формулой (2), по результатам случайной выборки (1), если известна дисперсия σ^2. Априорная информация включает оценку математического ожидания X_c, распределеннную также по нормальному закону с известной дисперсией σ^2 / n.

После выкладок [14] получим из (4) алгоритм вычисления оценки X_c, в соответствии с которым должна быть организована итерационная процедура:

$$X_c[l] = \frac{\sum_{i=1}^{n} x_i \exp\left[-\frac{(x_i - X_c[l-1])^2 (n+1)}{2\sigma^2}\right]}{\sum_{i=1}^{n} \exp\left[-\frac{(x_i - X_c[l-1])^2 (n+1)}{2\sigma^2}\right]}, \tag{5}$$

в качестве первого приближения целесообразно принять оценку максимального правдоподобия (3): $X_c[1] = \frac{1}{n}\sum_{i=1}^{n} x_i$, где l — номер текущей итерации.

В применении к задаче комплексирования данных алгоритм (5) выражает механизм дискриминаторов степеней свободы.

Результаты тестов показывают, что доверительный интервал, соответствующий уточненной оценке, меньше доверительного интервала оценки максимального правдоподобия. Наибольший выигрыш в эффективности получается при малых объемах выборки, поскольку с увеличением объема измерений относительный вклад априорной информации при получении оценок постепенно становится меньше, а байесовская оценка и оценка максимального правдоподобия асимптотически совпадают [12]. Поэтому вычислять уточненную оценку целесообразно главным образом при малых объемах выборки.

Важным свойством априорной плотности является то, что она не должна быть собственной плотностью, т.е. не обязательно ее интеграл должен быть равным единице [3]. В ряде случаев считаются вполне оправданными попытки использования псевдобайесовских оценок, при построении которых вместо недостающей априорной плотности вероятности оцениваемого параметра вводится другая плотность. Привлекает внимание возможность использования в качестве априорной плотности одной из так называемых потенциальных функций [13], частным случаем которых является нормальный закон распределения (2). Примерами могут также служить функции

$$f_1 = \frac{\alpha}{|X_c - x|},\ f_2 = \frac{\beta}{(X_c - x)^2},\ f_3 = \frac{\gamma}{1 + \delta(X_c - x)^2}, \tag{6}$$

где α, β, γ, δ — константы. Потенциальная функция характерна тем, что она монотонно убывает с удалением от значения X_c, т.е. является симметрично-четной относительно X_c. Если известно лишь, что оцениваемый параметр распределен в генеральной совокупности симметрично, то целесообразно получить уточненную оценку X_c, выбрав для априорной плотности достаточно простую потенциальную функцию.

Иногда для сокращения объема вычислений целесообразно намеренно заменить известный (например, нормальный) закон распределения другой, более простой потенциальной функцией. Так, если случайная величина распределена по равномерному закону, то оценка ее среднего значения подчиняется нормальному закону распределения. Выбрав, однако, в качестве априорной плотности первую из потенциальных функций (6), придем к следующему простому итерационному алгоритму вычисления оценки X_{c}:

$$X_{\text{c}}[l] = \sum_{i=1}^{n} \frac{x_i}{|X_{\text{c}}[l-1] - x_i| \sum_{j=1}^{n} \frac{1}{|X_{\text{c}}[l-1] - x_j|}},$$

$$X_{\text{c}}[1] = \frac{1}{n} \sum_{i=1}^{n} x_i;\ l \in [1, L],\ X_{\text{c}}[L] - X_{\text{c}}[L-1] \le \lambda_x. \tag{7}$$

Алгоритм (7) в применении к задаче комплексирования данных выражает механизм редукторов степеней свободы. Предложенная методика предусматривает индивидуальный подход к каждой реализации случайной величины (взвешивание в соответствии с апостериорной вероятностью ее появления), что позволяет [8] устранить потери информации при вычислении искомых оценок по малой выборке.

Важно отметить, что выработка оценки осуществляется посредством организации итерационного процесса, в котором элементы выборки на каждой итерации взаимодействуют. Аналогичным образом синергетика предусматривает процесс, характеризующийся самоуправлением и самоорганизацией в соответствии с поставленной целью. Здесь сложные процессы развиваются посредством коллективного взаимодействия компонент. Кооперация компонент позволяет использовать резервные возможности системы и существенно повышает степень эмерджентности (системный эффект).

ИЛЛЮСТРАЦИОННЫЕ ПРИМЕРЫ

Дано: количество каналов передачи данных $n \ge 3$ (число степеней свободы в синергетической системе комплексирования). Массив исходных данных представляется в виде матрицы-столбца

$$A^T = ||x_1 x_2 \ldots x_n||,$$

где x_i, $i \in [1, n]$, — данные о некоторой числовой величине x, полученные по i-м каналам (компоненты системы комплексирования).

Ставится задача: в зависимости от практических требований необходимо выявить наиболее информативный канал и получить достоверную оценку x^* величины x.

Пусть массив исходных данных представляется матрицей

$$A^T = \|x_1, \ldots, x_5\| = \|5.00,\ 6.50,\ 4.30,\ 5.20,\ 6.00\|.$$

Условие останова $X_{\text{c}}[L] - X_{\text{c}}[L-1] \le \lambda_x = 0.05$.

Оценка первой итерации:

$$X_{\text{c}}[1] = \frac{1}{n} \sum_{i=1}^{n} x_i = \frac{1}{5}(5.00 + 6.50 + 4.30 + 5.20 + 6.00) = \frac{27.00}{5} = 5.40.$$

Вариант 1. Рассмотрим задачу определения наиболее информативного канала и применим механизм редукторов степеней свободы, выражаемый алгоритмом (7).

Результаты расчетов по итерациям: $X_c[l] = 5{,}40$ при $l = 1$; $X_c[l] = 5{,}31$ при $l = 2$; $X_c[l] = 5{,}23$ при $l = 3$; $X_c[l] = 5{,}20$ при $l = 4$.

Таким образом, $x^* = X_c[L] = X_c[4] = 5{,}20$ и наиболее информативным каналом является четвертый.

Вариант 2. Рассмотрим задачу определения наиболее достоверной оценки величины x с помощью механизма дискриминаторов степеней свободы, выражаемого алгоритмом (5). При этом предполагается, что данная случайная величина распределена в генеральной совокупности нормально с известной дисперсией $2\sigma^2 = 1{,}5$.

Результаты расчетов по итерациям: $X_c[l] = 5{,}40$ при $l = 1$; $X_c[l] = 5{,}26$ при $l = 2$; $X_c[l] = 5{,}16$ при $l = 3$; $X_c[l] = 5{,}11$ при $l = 4$; $X_c[l] = 5{,}10$ при $l = 5$; $X_c[l] = 5{,}09$ при $l = 6$; $X_c[l] = 5{,}09$ при $l = 7$.

Таким образом, $x^* = X_c[L] = X_c[7] = 5{,}09$, что является наиболее достоверной оценкой при данных предположениях.

Итак, проведен сравнительный анализ аспектов, характерных как для методов математической статистики, так и для синергетических методов комплексирования данных. Результаты анализа применяются для повышения эффективности статистических оценок, вычисляемых по малой выборке, а также для выработки оценок характеристик объектов и процессов в синергетических системах комплексирования данных при ограниченном числе каналов.

СПИСОК ЛИТЕРАТУРЫ

1. Воронин А.Н., Зиатдинов Ю.К. Теория и практика многокритериальных решений: Модели, методы, реализация. — Saarbrucken (Deutschland); Lambert Academic Publishing, 2013. — 305 p.
2. Воронин А.Н. Метод комплексирования сигналов для бистатической радиолокации малых небесных тел // Тез. докл. 9-й международ. конф. «Системный анализ и управление». — М.: Изд-во МАИ, 2004. — С. 113–114.
3. Гаскаров Д.В., Шаповалов В.И. Малая выборка. — М.: Статистика, 1978. — 248 с.
4. Федулов А.А., Федулов Ю.Г., Цыгичко В.Н. Введение в теорию статистически ненадежных решений. — М.: Статистика, 1979. — 279 с.
5. Бешелев С.Д., Гурвич Ф.Г. Экспертные оценки. — М.: Наука, 1973. — 127 с.
6. Вентцель Е.С. Теория вероятностей. — М.: Наука, 1969. — 576 с.
7. Сейдж Э., Мелс Дж. Теория оценивания и ее применение в связи и управлении. — М.: Связь, 1976. — 496 с.
8. Кокс Д., Хинкли Д. Теоретическая статистика. — М.: Мир, 1978. — 560 с.
9. Шеннон К. Работы по теории информации и кибернетике. — М.: Изд-во иностр. лит., 1963. — 829 с.
10. Гутер Р.С., Резниковский П.Т. Программирование и вычислительная математика. — М.: Наука, 1971. — Вып. 2. — 273 с.
11. Колесников А.А. Синергетическая теория управления. — М.: Энергоатомиздат, 1994. — 344 с.
12. Мудров В.И., Кушко В.Л. Методы обработки измерений. — М.: Сов. радио, 1976. — 192 с.
13. Айзерман М.А., Браверман Э.М., Розоноэр Л.И. Метод потенциальных функций в теории обучения машин. — М.: Наука, 1970. — 384 с.
14. Воронин А.Н. Синергетические методы комплексирования данных // Кибернетика и системный анализ. — 2006. — № 2. — С. 24–30.

Поступила 12.06.2013