УЧЕНЫЕ ЗАПИСКИ КАЗАНСКОГО УНИВЕРСИТЕТА
Том 152, кн. 4 Физико-математические науки *2010*

УДК 621.391.26

НЕПАРАМЕТРИЧЕСКОЕ ОЦЕНИВАНИЕ СМЕСИ ПЛОТНОСТЕЙ ВЕРОЯТНОСТЕЙ СИГНАЛОВ (ПОМЕХ)

Э.А. Ибатуллин

Аннотация

В работе для непараметрического оценивания смеси плотностей вероятностей сигналов (помех) предлагается использовать параллельную кластер-процедуру. Эта процедура позволяет достаточно точно оценивать параметры частных распределений смеси, в том числе и количество компонентов (классов), входящих в смесь. Все сказанное подтверждается полученными численными результатами.

Ключевые слова: непараметрическое оценивание, смесь плотностей вероятностей, параллельная кластер-процедура.

Введение

При статистическом синтезе устройств оптимальной обработки сигналов при действии помех необходимо располагать плотностями вероятностей сигналов и помех. Часто эти плотности вероятностей отличаются от гауссовской и имеют многомодовую природу. Таким образом, результирующая плотность вероятности может быть представлена как смесь плотностей вероятностей [1, 2]

$$f_r(\mathbf{X}, p, \mathbf{\Theta}) = \sum_{i=1}^{k} p_i f(\mathbf{X}/\mathbf{\Theta}_i), \tag{1}$$

где k – количество классов стохастических сигналов (помех); $f(\mathbf{X}/\mathbf{\Theta}_i)$ – частная плотность вероятности сигнала (помехи) $\mathbf{X}$ в i-м классе, характеризующемся набором параметров $\mathbf{\Theta}_i$; p_i – вес i-го класса в смеси:

$$\sum_{i=1}^{k} p_i = 1, \quad i = 1, \ldots, k, \quad \mathbf{\Theta}_i \neq \mathbf{\Theta}_j, \quad i \neq j.$$

Под классом сигналов (помех) будем понимать набор сигналов (помех) от одного источника, когда существенные параметры сигнала (помехи) изменяются случайным образом. Под существенным параметром сигнала (помехи) мы понимаем оцениваемый (измеряемый) параметр. Случайный характер сигнального (помехового) параметра появляется вследствие аппаратурных нестабильностей и измерительных ошибок. Например, случайным существенным параметром сигнала (помехи) может быть момент прихода сигнала (помехи) или его несущая частота.

Обычно вид функции $f(\mathbf{X}/\mathbf{\Theta}_i)$, $i = 1, \ldots, k$, в уравнении (1) предполагается известным. Неизвестными значениями будут p_i и $\mathbf{\Theta}_i$, $i = 1, \ldots, k$. Количество распределений k в смеси также может быть неизвестным. Задача состоит в оценке неизвестных параметров смеси распределений, описывающих классы сигналов (помех). В такой постановке задача решалась в [3, 4]. В настоящей статье рассматривается случай, когда виды плотностей распределений неизвестны, следовательно,

неизвестны k, p_i и $\mathbf{\Theta}_i$, $i = 1, \ldots, k$. Задача заключается в оценке числа классов, их средних значений и дисперсий, то есть в разрешении классов сигналов (помех). Дальнейшее изложение одинаково как для сигналов, так и для помех, поэтому для краткости будем рассматривать только сигналы.

1. Параллельная кластер-процедура

В том случае, когда виды плотностей распределений неизвестны, для разрешения классов сигналов возможно использование методов кластер-анализа, которые относятся к непараметрическим методам [2]. При кластер-анализе совокупность исследуемых сигналов разбивается на однородные в некотором смысле группы (кластеры). Разбиение производится в соответствии с некоторым критерием, характеризующим качество группирования. В роли такого критерия, например, может выступать условие минимальности величины, пропорциональной сумме внутриклассовых дисперсий [5]:

$$Q(S) = \sum_{i=1}^{k} \sum_{x_l \in S_i} \rho^2 \left[x_l, a_i\right], \tag{2}$$

где S_i – i-й класс, k – число классов (групп), $\rho\left[x_l, a_i\right]$ – расстояние между l-м сигналом и средним значением a_i i-го класса.

Кластер-процедуры реализуются с помощью итерационных алгоритмов. Если на каждом шаге одновременно (параллельно) используются все имеющиеся наблюдения, то кластер-процедура называется параллельной, в противном случае имеем последовательную процедуру. К параллельным кластер-процедурам относятся в первую очередь алгоритмы «переноса точек из класса в класс», отправляющиеся от некоторого начального разбиения $S^{(0)} = \left\{S_1^{(0)}, \ldots, S_k^{(0)}\right\}$, взятого произвольно или с помощью какого-либо из методов предварительной обработки исходных наблюдений. Вычисляется значение величины $Q(S)$. Затем каждое из наблюдений x_l поочередно перемещается во все кластеры, рассматривается как самостоятельный кластер, если число кластеров неизвестно, и оставляется в том положении, которое соответствует минимальному значению функционала качества Q. Работа алгоритма заканчивается, когда перемещения наблюдений перестанут приводить к улучшению (в смысле уменьшения Q) разбиения. Часто описанный алгоритм применяют несколько раз к одной и той же исходной совокупности наблюдений, исходя из различных начальных разбиений $S^{(0)}$, и выбирают в итоге наилучший (в смысле минимума Q) вариант разбиения.

Таковы общие предпосылки алгоритма параллельной кластер-процедуры. Более детальный анализ алгоритма позволил установить следующее. Для однозначного решения поставленной задачи требуется условие унимодальности частных плотностей вероятностей $f\left(\mathbf{X}/\mathbf{\Theta}_i\right)$, хотя при этом их виды остаются неизвестными. При неизвестном числе классов наряду с условием минимума Q необходимо также ввести еще дополнительные критерии – это минимальность расстояния между классами (Δa) и максимальность возможного числа классов (k_m). При этом в смешанной выборке сигналов производится их упорядочивание в порядке возрастания, что ускоряет процедуру перестановки сигналов из класса в класс. Далее из полученной последовательности формируются классы, сначала один, затем два, три и т. д. Решением считается разбиение на k классов, если при разбиении на $k+1$ класс выполняется одно из условий:

1) расстояние между классами станет меньше Δa;

2) число классов превысит k_m, при этом $Q_k < Q_{k-1}$.

Для величины, пропорциональной внутриклассовой дисперсии i-го класса, в соответствии с (2) можем записать

$$Q_i = \sum_{x_l \in S_i} \rho^2 (x_l, a_i) = \sum_{l=1}^{n_i} (x_{li} - a_i)^2, \tag{3}$$

где

$$a_i = \frac{1}{n_i} \sum_{l=1}^{n_i} x_{li}, \tag{4}$$

n_i – размер пачки сигналов x_{li}, отнесенных в i-й класс. В целях сокращения объёма вычислений при перестановке сигналов из класса в класс целесообразно получить рекуррентные выражения для a_i и Q_i. Такие выражения можно получить из (3) и (4). При добавлении сигнала x' в i-й класс имеем

$$a'_i = \frac{1}{n_i + 1} \sum_{l=1}^{n_i+1} x_{l_i} = \frac{1}{n_i + 1} \sum_{l=1}^{n_i} x_{l_i} + \frac{x'}{n_i + 1}. \tag{5}$$

Из выражения (4) вытекает, что

$$\sum_{l=1}^{n_i} x_{l_i} = a_i n_i, \tag{6}$$

поэтому

$$a'_i = \frac{n_i a_i + x'}{n_i + 1}. \tag{7}$$

Далее из (3) имеем

$$Q_i = \sum_{l=1}^{n_i} (x_{l_i} - a_i)^2 = \sum_{l=1}^{n_i} x_{l_i}^2 - 2a_i \sum_{l=1}^{n_i} x_{l_i} + n_i a_i^2. \tag{8}$$

С учетом (6) для Q_i получим:

$$Q_i = \sum_{l=1}^{n_i} x_{l_i}^2 - n_i a_i^2. \tag{9}$$

При добавлении сигнала x' выражение (9) примет вид

$$Q'_i = \sum_{l=1}^{n_i+1} x_{l_i}^2 - (n_i + 1)\, a'^2_i = \sum_{l=1}^{n_i} x_{l_i}^2 + (x')^2 - (n_i + 1) \left(\frac{n_i a_i + x'}{n_i + 1} \right)^2. \tag{10}$$

Прибавив и вычтя из (10) выражение $n_i a_i^2$ и произведя необходимые упрощения, приходим к выражению

$$Q'_i = Q_i + \frac{n_i (x' - a_i)^2}{n_i + 1}. \tag{11}$$

При исключении сигнала x' из i-го класса, поступая аналогично предыдущему случаю, получим:

$$a'_i = \frac{n_i a_i - x'}{n_i - 1}, \tag{12}$$

$$Q'_i = Q_i - \frac{n_i (x' - a_i)^2}{n_i - 1}. \tag{13}$$

2. Моделирование и результаты

На основе выражений (2)–(13) была разработана имитационная модель, составлена компьютерная программа и проведен вычислительный статистический эксперимент. Результаты получены для нескольких статистических экспериментов, при которых производились генерации различного количества классов нормально распределенных по времени запаздывания сигналов со среднеквадратичным отклонением, для каждого класса равным 0.3. Ошибки кластеризации, которые появлялись в результате работы алгоритма, определялись числом сигналов, ошибочно отнесенных в другой класс, к общему числу сигналов, сгенерированных в рассматриваемом классе, то есть подсчитывалось отношение $p_{ij} = n_{ij}/n_i$. Минимальное расстояние между классами было выбрано равным единице.

В первом эксперименте проводилась генерация одного класса с пачкой сигналов, равной 110, и средним, равным 5. При разбиении, включающем один класс, величина $Q = 11.205$, а среднее равно $a_1 = 4.9784$. При разбиении на два класса величина $Q = 4.011$, а средние получаются равными $a_1 = 4.7534$, $a_2 = 5.2691$. Хотя при двух классах значение величины Q, пропорциональной сумме внутриклассовых дисперсий, меньше, чем при одном классе, расстояние между средними получается меньшим, чем допустимое расстояние $\Delta x = 1$. Следовательно, при разбиении на два класса не выполняется критерий минимальности расстояния между классами. Таким образом, имеем один класс сигналов со средним $a_1 = 4.9784$. Относительная ошибка при определении среднего составляет 0.4%.

Во втором статистическом эксперименте проводилась генерация двух классов сигналов со средними 3 и 5. Здесь также $Q_2 < Q_1$, при разбиении на один и два класса, а при разбиении на три класса выполнены неравенства $Q_3 < Q_2 < < Q_1$. Средние значения при трех классах составили $a_1 = 2.7527$, $a_2 = 3.1849$, $a_3 = 5,0155$.. Здесь видно, что для 1-го и 2-го классов не выполняется критерий минимального расстояния. Следовательно, имеют место сигналы от двух классов со средними $a_1 = 2.9414$, $a_2 = 5.0155$ с максимальной относительной ошибкой определения среднего, равной 1.95%.

В третьем статистическом эксперименте проводилась генерация трех классов сигналов со средними 3, 5 и 6, причем между 2-м и 3-м классами расстояние равно минимальному. При разбиении на два класса $Q_2 = 29.208$, а при трех классах $Q_3 = = 8.619$. При разбиении на четыре класса $Q_4 = 7.217$, а средние равны $a_1 = 2.7609$, $a_2 = 3.1657$, $a_3 = 4.9544$, $a_4 = 6.0350$.. Замечаем, что между 1-м и 2-м классами нарушен критерий минимального расстояния. Отсюда, хотя $Q_4 < Q_3$, число классов равно трем со средними $a_1 = 2.9183$, $a_2 = 4.9544$, $a_3 = 6.0350$ и с максимальной относительной ошибкой определения среднего 2.72%. В этом эксперименте между 2-м и 3-м классами имеется пересечение на уровне 2.66σ. Это обстоятельство привело к ошибочному попаданию сигналов 2-го класса в 3-й и наоборот. Частота этих симметричных ошибок составляет $p_{23} = p_{32} = 0.083$. Эти результаты показывают, что параллельная кластер-процедура достаточно устойчиво работает при значительном пересечении классов.

Заключение

Параллельная кластер-процедура позволяет проводить оценивание параметров смеси вероятностных плотностей сигналов (помех), описывающих классы сигналов (помех).

При неизвестном числе классов сигналов (помех), наряду с критерием качества разбиения-суммой внутриклассовых дисперсий, необходимо использовать еще дополнительные критерии: минимальное расстояние между классами и максимально возможное число классов.

Результаты, полученные при статистическом моделировании, показывают работоспособность параллельной кластер-процедуры при значительном пересечении классов сигналов (помех).

Предложенный алгоритм имеет преимущество перед алгоритмами последовательной кластер-процедуры, заключающееся в меньшем времени, необходимом для обработки сигналов.

Summary

E.A. Ibatoulline. Nonparametric Estimation of a Mixture of Probability Densities of Signals (Interferences).

In this article, we suggest using the parallel clustering procedure for the nonparametric estimation of a mixture of probability densities of signals (interferences). This procedure makes it possible to estimate with an adequate accuracy the parameters of the marginal mixture distribution, as well as the number of components (classes) in a mixture. The efficiency of the suggested method is confirmed by the numerical results.

Key words: nonparametric estimation, mixture of probability densities, parallel clustering procedure.

Литература

1. *Yakowitz S., Spragins J.* On the Identifiability of Finite Mixture // Ann. Math. Stat. – 1968. – V. 39, No 1. – P. 209–214.
2. *Патрик Э.А.* Основы теории распознавания образов / Пер. с англ. – М.: Сов. радио, 1980. – 408 с.
3. *Ibatoulline E.A.* Parameter estimation of non-Gaussian probability density of signals and interferences // Proc. Int. Sym. EMC'98 ROMA. – Rome, Italy, 1998. – V. II. – P. 716–718.
4. *Ибатуллин Э.А.* Оценивание негауссовой плотности вероятности сигналов и помех итерационным методом максимального правдоподобия // Прием и обработка информации в сложных информационных системах. – Казань: Изд-во Казан. ун-та, 2001. – Вып. 20. – С. 20–29.
5. *Айвазян С.А., Бежаева З.И., Староверов О.В.* Классификация многомерных наблюдений. – М.: Статистика, 1974. - 240 с.
6. *Ibatoulline E.A.* Nonparametric estimation of the mixture of signal (interference) probability densities // Records of Int. Conf. on Modeling and Simulation AMSE'08. – Petra, Jordan, 2008. – P. 75–78.

Поступила в редакцию
02.04.10

Ибатуллин Эмир Аминович – доктор физико-математических наук, профессор Института физики Казанского (Приволжского) федерального университета.

E-mail: *Emir.Ibatoulline@ksu.ru*