УДК: 519.254

# Восстановление смесей распределений в задачах идентификации личности

**Л.В. Рябова, ассистент кафедры средств защиты информации института информационно-диагностических систем Национального авиационного университета, lubanau@ukr.net ,**
**М.Е. Самойленко, *аспирант, НАУ,***
***Щербак Т.Л., к.т.н., доцент, НАУ, Німченко Т.В., к.т.н., доцент, НАУ.***

Большая часть биометрических классификаций проведена на основании байесовских алгоритмов анализа, который основывается на знании априорных вероятностей классов и законов распределения вероятностей признаков в каждом классе. На практике нам известна только обучающая выборка объектов. Будем считать элементы выборки независимыми случайными величинами, имеющими одинаковое распределение. Требуется по выборке оценить плотность этого распределения.

Существуют три основных подхода к оцениванию плотности распределения: непараметрический, параметрический и восстановление смесей распределений. Первые два метода разработаны и изучены достаточно давно и скурпулезно как в теоретической части, так и в области практического применения. На практике крайне важна корректная интерпретация полученных результатов, по этому возникает задача определения неизвестных параметров смеси на основе анализа данных, для которых строится модель. Между тем, для описания сложных стохастических систем ( например, автоматизированная система идентификации личности, автоматизированный комплекс технических средств защиты информации, финансовый рынок и т.п. ) часто используется модель, основанная на конечной смеси вероятностных распределений, то есть

$$P_{\Theta}^{x}(x)= \sum_{i=1}^{k} p_i \psi_i(x,t_i) \qquad (1)$$

где $k \geq 1$ –известное натуральное число, $\psi_1, \ldots \psi_k$ – известные плотности распределений, неизвестный параметр $\Theta=(p_1, \ldots, p_k; t_1, \ldots, t_k)$, причем $p_i \geq 0$, $\sum_i p_i = 1$, $i=1, \ldots, k$ – многомерные параметры. Плотности $\psi_1, \ldots, \psi_k$ – обычно называют компонентами смеси (1), а параметры $p_1, \ldots, p_k$ – веса соответствующих компонент. Т.е., если "форма" классов имеет достаточно сложный вид, не "поддающийся" описанию одним распределением, то применяют методы восстановления смесей распределений – описывают класс несколькими распределениями. Функции правдоподобия принадлежат параметрическому семейству распределений $\Theta( x;\Theta )$ и отличаются только

значениями параметра $p_i(x)=Ø(x;\Theta)$. Классическая постановка задачи состоит в том, что нам известна выборка $X_m$ – независимых случайных наблюдений смеси $p_i(x)$, известно число $k$ компонент смеси функции $Ø$.

Требуется найти оценку параметров $Ø=(w_1,\ldots,w_k;\Theta_1,\ldots,\Theta_k)$.

К сожалению, попытка разделить смесь, используя принцип максимума правдоподобия приводит к чрезвычайно громоздкой задаче оптимизации. Обойти эту проблему позволяет алгоритм EM (expectation-maximization) . Восстановление смесей распределений, состоит из итеррационного повторения двух шагов. Первоначально на Е-шаге искусственно вводится вектор скрытых переменных $G$, для которого по текущему приближению вектора параметров $\Theta$ вычисляется ожидаемое значение (expectation) вектора скрытых переменных. На М-шаге решается задача максимизации правдоподобия (maximization) и находится следующее приближение вектора по текущим значениям векторов $G$ и $\Theta$ .

Итеррации останавливаются, когда значения функционала $Q(\Theta)$ , где

$$Q(\Theta)=\sum_{i=1}^{m}\ln\sum_{j=1}^{k}w_j p_j(x_i)\to\max \qquad (2)$$

или скрытых переменных перестают существенно изменяться. Хотя алгоритм EM сходится при достаточно общих предположениях, скорость сходимости может существенно зависеть от "хорошего" выбора начального приближения. До сих пор предполагалось, что число компонент известно заранее [ 1, 2 ]. На практике это не так, как правило, интересует корректная интерпретация полученных результатов. По этому были разработаны предназначенные для устранения этих проблем информационные критерии, основанные на функции правдоподобия. Однако в ряде практических значимых моделей нарушаются условия регулярности (например, для конечной смеси нормальных законов), что приводит к необходимости накладывать дополнительные искусственные технические условия для корректности использования информационных критериев. На практике вычислительный EM- алгоритм существенно усложняется. Для устранения указанных недостатков были предложены более мощные критерии проверки гипотез о числе компонент смеси. Для формализации задачи использованы две модели непрерывных распределений: добавлений и расщепления компоненты.

Модель добавления компоненты реализуется следующим способом. Предполагается, что каждое из независимых наблюдений $(X_1, \ldots, X_n)$ имеет плотность в виде конечной $k$-компонентной смеси некоторых законов распределения вида ($\Theta \in [0,1]$ ).

$$p(x;\Theta)=(1-\Theta)f(x)+\Theta g(x),\ \sum_{i=1}^{k}p_i=1 \qquad (3)$$

Отметим, что если плотности $\psi_i(x)$, $i=1,\ldots,k$ функция $f(x)$ строго положительна, то критерий проверки гипотезы о числе компонент смеси основан на статистике

$$Q_1 = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left( \frac{f(x_i)}{g(x_i)} - 1 \right) \qquad (4)$$

Данная модель ориентируется на проверку значимости произвольной компоненты с возможным малым весом. Решая задачу об уменьшении числа в подгоняемой модели смеси вероятностных распределений, важно не исключить из рассмотрения практически важные компоненты, ошибочно объединив их в одну компоненту. Это означает, что в смеси присутствуют компоненты с близкими значениями параметров, и том числе и весов. Тогда этому случаю соответствует модель расщепления компонент. Критерий проверки гипотезы о числе компонент в модели расщепления основан на статистике

$$Q_2 = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{g(x_i)}{f(x_i)} \qquad (5)$$

предполагая, что каждое независимых наблюдений $\Theta \in [0,1]$ имеет плотность вида

$$p(x,\Theta) = f(x) + \Theta g(x) \qquad (6)$$

Отметим, что для нахождения критического значения в обоих моделях статистик предполагается знание значений фишеровской информации, определяемой из соотношений

$$I_1 = \int_{-\infty}^{\infty} \frac{g^2(x)}{f(x)} dx - 1 \ldots\ldots I_2 = \int_{-\infty}^{\infty} \frac{g^2(x)}{f(x)} dx \qquad (7)$$

Результаты проведенного тестирования означают, что предложенные критерии могут использоваться и при уменьшенных объемах выборки без существенных ограничений. При этом, отметим относительную простоту практического использования этих критериев. Проведенное сравнение показывает, что статистические критерии являются весьма эффективными и удобными для применения на практике в силу высокой точности и скорости работы.

**Литература**

1. Akaike H. Information theory and an extension of the maximum lakelihood principile.// Second International Symposium on Information Theory.- Budapest, 1973. P. 267-281.

2. Schwartz G. Estimating the dimension of a model // The Annals of Statistics.- 1978.- Vol. 6.- P.461 – 464.