О выборе весов для подмножеств признаков при распознавании речи

Чучупал В. Я.
chuchu@ccas.ru
Москва, Вычислительный Центр РАН

При распознавании речи с использованием квазифонемных марковских моделей распознанная последовательность моделей $q_1^T = q_1, \ldots, q_T$ для данной последовательности наблюдений $x_1^T = x_1, \ldots, x_T$ обычно определяется путем максимизации логарифма правдоподобия $\log P(x_1^T | q_1^T)$ на множестве всех допустимых последовательностей моделей [1]:

$$\log P(x_1^T|\Theta) = \arg\max_{q_1^T} \log P(x_1^T|q_1^T).$$

При этом

$$\log P(x_1^T|q_1^T) = \sum_{i=1}^{T} \log P(x_i|q_i). \tag{1}$$

Наблюдение x_i обычно содержит несколько подмножеств параметров, $x_i = (x_i^1, x_i^2, \ldots, x_i^K)$. Здесь K — число подмножеств, x_i^j — набор параметров j подмножества в момент i. Величина локального, в момент i, правдоподобия вычисляется как:

$$\log P(x_i|\Theta) = \sum_{j=1}^{K} \lambda_j \log P(x_i^j|\Theta) = \sum_{j=1}^{K} \mu_j(x_i^j, \Theta), \tag{2}$$

где через Θ обозначены параметры модели состояния q_i, а λ_j — весовой коэффициент j подмножества параметров.

Величина коэффициентов λ_j устанавливается из эвристических соображений и, как правило, эти коэффициенты выбираются равными, не зависящими от модели Θ или параметров [2]. Возникает вопрос: можно ли систематически и более общим образом выбрать λ или μ, например, как функции от моделей и текущих параметров?

Оценка вероятностей на основе максимума энтропии

Пусть x — случайная величина, и нас интересует распределение, либо плотность ее вероятностей. В соответствии с принципом максимальной энтропии [3], если для функций $f_i(x)$, $i = 1, \ldots, K$ известны математические ожидания, то существуют такие константы $\lambda_0, \ldots, \lambda_K$ и такое распределение вероятностей $P(x)$, что оно обладает максимальной энтропией на множестве всех вероятностных распределений, которые удовлетворяют заданным ограничениям и может быть выражено как:

$$P(x) = \frac{\exp\bigl(-\lambda_0 - \lambda_1 f_1(x) - \dots - \lambda_K f_K(x)\bigr)}{\sum\limits_x \exp\bigl(-\lambda_1 f_1(x) - \dots - \lambda_K f_K(x)\bigr)}. \tag{3}$$

Распределение $P(x)$ оптимально в том смысле, что кроме заданных ограничений никаких других зависимостей не предполагает.

Если обозначить

$$\mu_0(x) = \frac{\exp(-\lambda_0)}{\sum\limits_x \exp\bigl(-\lambda_1 f_1(x) - \dots - \lambda_K f_K(x)\bigr)},$$
$$\mu_1(x) = e^{-\lambda_1 f_1(x)}, \dots, \mu_K(x) = e^{-\lambda_K f_K(x)},$$

то уравнение (3) можно переписать в виде

$$\log P(x) = \sum_{j=0}^{K} \log \mu_j(x). \tag{4}$$

Алгоритм оценки весов подмножеств

Рассмотрим случай дискретных марковских моделей: когда параметры сигнала кодируются элементами кодовой книги: $x_t^i \to c_i^t$, или $x_t^i \in c_i^t$. Здесь c_i — элемент кодовой книги C^i для описания i-го подпространства параметров.

Оценка условной вероятности $P(\Theta|c_{i0})$ модели Θ при параметрах, принадлежащих коду c_{i0}, есть среднее значение:

$$P(\Theta|c_{i0}) = K_{\Theta,c_{i0}} = \mathop{\mathsf{E}}_{\substack{(c_1,\dots,c_K)\\ c_i=c_{i0}}} \bigl[P(\Theta|c_1, c_2, \dots, c_K)\bigr]. \tag{5}$$

Равенства (5), при всех Θ, c_i, можно рассматривать как набор ограничений.

Введем (аналогично [4]) бинарные селекторные функции

$$f_{\Theta,c_i}(x) = \begin{cases} 1, & \text{если } x \in \Theta \text{ и } x \in c_i; \\ 0, & \text{в противном случае.} \end{cases} \tag{6}$$

Тогда ограничения (5) можно записать в виде:

$$\sum_{c_1,\dots,c_K} P(c_1, \dots, c_K) \sum_{\Theta} P(\Theta|c_1, \dots, c_K) f_{\Theta,c_i}(x) = K_{\Theta,c_i} P(H_{c_i}), \tag{7}$$

Алгоритм 1. Оценка оптимальных весов подмножеств для дискретного случая.

Вход: параметры модели Θ, $j = 0$;

1: Оценим по частотам обучающей выборки ограничения (7) и начальные значения $\log\mu$ для всех i, c_{i0}:

$\log K^{(0)}_{\Theta,c_{i0}} := \log P(\Theta|c_i)$;

$\log\mu^{(0)}(\Theta, c_{i0}) := \log K_{\Theta,c_{i0}} + \log P(H_{c_{i0}})$;

2: **повторять**

3: Оценим вероятности (4):

$\log P(\Theta|c_1,\ldots,c_K) := \sum\limits_{i=1}^{K} \log\mu^{(j)}(\Theta, c_j) - \log\Big(\sum\limits_{\Theta} \mu^{(j)}(\Theta, c_j)\Big)$;

4: Вычислим новые значения ограничений:

$\log K_{\Theta,c_{i0}} := \log\Big(\sum\limits_{\substack{(c_1,\ldots,c_K)\\ c_i=c_{i0}}} P(c_1,\ldots,c_K)P(\Theta|c_1,\ldots,c_K)\Big)$;

5: Переоценим $\log\mu$:

$\log\mu^{(j+1)}(\Theta, c_i) := \log\mu^{(j)}(\Theta, c_i) + \log K^{(0)}_{\Theta,c_i} - K^{(j)}_{\Theta,c_i}$;

6: $j := j + 1$;

7: **пока** $\log\mu^{(j+1)}(\Theta, c_i) - \log\mu^{(j)}(\Theta, c_i) \geqslant \varepsilon$;

где $P(H_{c_i})$ обозначает вероятность того, что наблюдаемое значение кода для i-го подпространства параметров есть c_i.

Алгоритм 1 вычисления $\mu(\Theta, x^j)$ по обучающей выборке основан на алгоритме GIS (Generalized Iterative Scaling) [5].

Чтобы полученные оценки можно было использовать обычным для распознавания речи образом, перепишем (1), используя формулу Байеса:

$$\sum_{i=1}^{T} \log P(c_i^1,\ldots,c_i^K|q_i) = \\ = \sum_{i=1}^{T} \log P(q_i|c_i^1,\ldots,c_i^K) + \sum_{i=1}^{T} \log P(c_i^1,\ldots,c_i^K) - \sum_{i=1}^{T} \log P(q_i). \quad (8)$$

Априорные вероятности $P(q_i)$ оцениваются по обучающей выборке, далее, поскольку последовательность наблюдений известна, то сумма вероятностей $\sum_{i=1}^{T} P(c_i^1,\ldots,c_i^K)$ постоянна, и при поиске максимума её можно игнорировать.

Работа выполнена при поддержке РФФИ, проект № 07-01-00657а.

Литература

[1] *Jelinek F.* Statistical methods for speech recognition. MIT Press, 1998.

[2] *Young, S.* The HTK BOOK. Ver. 2.1. Cambridge University, 1997.

[3] *Janes, E.T.* Probability theory: the logic of science. Cambridge University Press, 2006.

[4] *Rosenfeld R.* A maximum entropy approach to adaptive statistical language modeling // Computer Language and Speech. — Vol. 10, No. 3. — Pp. 187–228.

[5] *Darroch J. N., Ratcliff D.* Generalized iterative scaling for log-linear models // The annals of Mathematical Statistics, 1972. — No. 43. — Pp. 1470–1480.