

О выборе весов для подмножеств признаков при распознавании речи

Чучупал В. Я.

Москва, Вычислительный Центр РАН
chuchu@ccas.ru

При распознавании речи с использованием квазифонемных марковских моделей распознанная последовательность моделей $q_1^T = q_1, \dots, q_T$ для данной последовательности наблюдений $x_1^T = x_1, \dots, x_T$ 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

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

Наблюдение x_t обычно содержит несколько подмножеств параметров, $x_t = (x_t^1, x_t^2, \dots, x_t^K)$. Здесь K — число подмножеств, x_t^i — набор параметров i подмножества в момент t . Величина локального, в момент i , правдоподобия:

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

где через Θ обозначены параметры модели состояния q_i , а λ_j — весовой коэффициент соответствующего 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 веса μ или λ , 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 $f_i(x)$, $i = 1, \dots, K$ известны математические ожидания, то существуют такие константы $\lambda_0, \dots, \lambda_K$ и такое распределение вероятностей $P(x)$, что оно обладает максимальной энтропией на множестве всех вероятностных распределений, которые удо-

влетворяют заданным ограничениям и может быть выражено как:

$$P(x) = \frac{\exp(-\lambda_0 - \lambda_1 f_1(x) - \dots - \lambda_K f_K(x))}{\sum_x \exp(-\lambda_1 f_1(x) - \dots - \lambda_K f_K(x))}. \quad (3)$$

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

Если обозначить

$$\mu_0 = \frac{\exp(-\lambda_0)}{\sum_x \exp(-\lambda_1 f_1(x) - \dots - \lambda_K f_K(x))}, \quad \mu_1 = e^{-\lambda_1}, \dots, \mu_n = e^{-\lambda_K},$$

то уравнение (3) можно переписать в виде

$$\log P(x) = \sum_{j=1}^K \log \mu_j(x). \quad (4)$$

Алгоритм оценки весов подмножеств

Рассмотрим случай дискретных марковских моделей: когда параметры сигнала кодируются элементами кодовой книги: $x_t^i \rightarrow c_t^i$, или $x_t^i \in c_t^i$. 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:

$$P(\Theta|c_{i0}) = \underset{\substack{(c_1, \dots, c_K) \\ c_i = c_{i0}}}{\mathbb{E}} [P(\Theta|c_1, c_2, \dots, c_K)]. \quad (5)$$

Равенства $K_{\Theta, c_i} = P(\Theta|c_i)$, при всех Θ, c_i есть набор ограничений.

Введем (аналогично [4]) бинарные селекторные функции

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

Тогда ограничения (5) можно записать в виде:

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

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

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} :

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

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

- 2: **повторять**

- 3: Оценим вероятности (4):

$$\log P(\Theta | c_1, \dots, c_K) := \sum_{i=1}^K \log \mu^{(j)}(\Theta, c_j) - \log \left(\sum_{\Theta} \mu^{(j)}(\Theta, c_j) \right);$$

- 4: Вычислим новые значения ограничений:

$$\log K_{\Theta, c_{i0}} := \log \left(\sum_{\substack{(c_1, \dots, c_K) \\ c_i = c_{i0}}} P(c_1, \dots, c_K) P(\Theta | c_1, \dots, c_K) \right);$$

- 5: Переоценим $\log \mu$:

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

- 6: $j := j + 1$;

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

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:

$$\begin{aligned} \sum_{t=1}^T \log P(c_1^t, \dots, c_K^t | q^t) &= \\ &= \sum_{t=1}^T \log P(q^t | c_1^t, \dots, c_K^t) + \sum_{t=1}^T \log P(c_1^t, \dots, c_K^t) - \sum_{t=1}^T \log P(q^t). \end{aligned} \quad (8)$$

Априорные вероятности $P(q^t)$ оцениваются по обучающей выборке, далее, поскольку последовательность наблюдений известна, то сумма вероятностей $\sum_{t=1}^T P(c_1^t, \dots, c_K^t)$ 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.