Краткий обзор приложения метода условных случайных полей в области распознавания речи

Леднов Д.А., *кандидат технических наук*

Цель этой работы — обзор математических основ модели условных случайных полей, а также идеологическое сравнение этой модели с другими известными направлениями обработки речи.

• условные случайные поля • скрытые модели Маркова • модель Байеса • модель максимума энтропии

This paper is aimed to survey conditional random fields mathematical bases and to compare them with other direction of speech processing.

• conditional random fields • hidden Markov's models • Byres' model • model maximum entropy

Введение

Последние десять лет в области распознавания речи развивается направление, которое может стать, с одной стороны, достойной альтернативой скрытым марковским моделям (СММ), а с другой стороны, включить в себя все преимущества СММ. Это направление известно как модель условных случайных полей. Цель этой работы состоит в том, чтобы показать математические основы этого направления и идеологически сравнить его с другими направлениями обработки речи. Здесь приводится сравнение метода условных случайных полей с методом Байеса, скрытыми марковскими моделями и методом максимальной энтропии. В качестве базовых работ этого обзора выбраны работы [6,7].

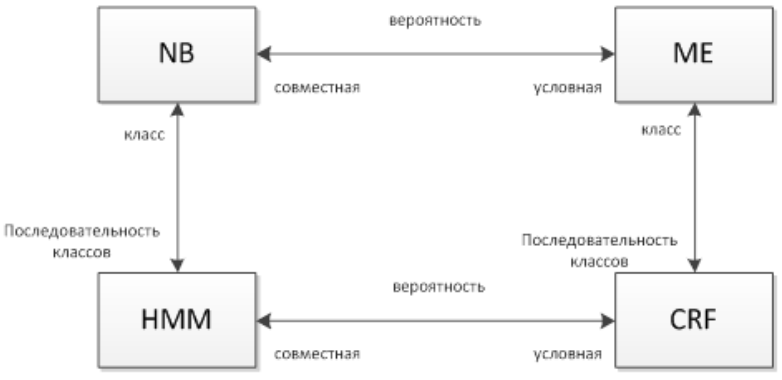

Рис. 1. Классификация статистических моделей в области классификации паттернов. На рисунке использованы английские аббревиатуры — NB — Nave Bayes, HMM — hidden Markov's Models, ME — Maximum Entropy Model, CRF — Conditional Random Field

Модель Байеса

Модель Байеса связана с тем, что в процессе моделирования необходимо построить представление условной вероятности $p(Y|X)$ с входным вектором наблюдений $X = (x_1, x_2, .., x_m)$ и условно зависимой Y переменной класса, подлежащей определению. Эта условная вероятность может быть представлена с помощью правила Байеса:

$$p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)}.$$

Делитель $p(X)$ не важен для классификации и может быть интерпретирован как нормализационная константа. Числитель – это совместная вероятность $p(X|Y)p(Y) = p(Y,X)$.

Общая декомпозиция этой вероятности может быть выполнена с помощью формулы полной вероятности:

$$p(Y,X) = p(Y)p(Y|x_1)\prod_{i=2}^{m} p(x_i|x_{i-1},..,x_1,Y)$$

На практике часто допускается, что входные переменные условно независимы. Это допущение известно, как предположение Байеса. На основе этого предположения получим:

$$p(Y,X) = p(Y)\prod_{i=1} p(x_i|Y).$$

Последнее выражение равносильно высказыванию, что совместная вероятность появления вектора наблюдения и класса пропорциональна произведению вероятностей того, что определённое значение каждой компоненты вектора наблюдений порождается переменной класса.

Скрытая модель Маркова

Модель Байеса допускает, что весь вектор наблюдения принадлежит единственному классу. В модели Маркова, в отличие от модели Байеса, дана последовательность наблюдений, и поэтому необходимо идентифицировать последовательность классов:

$$p(Y,X) = \prod_{i=1}^{m} p(y_i)\, p(x_i|y_i)$$

$$p(Y,X) = p(y_1)p(x_1|y_1)\prod_{i=2}^{m} p(y_i|y_{i-1})\, p(x_i|y_i).$$

По сравнению с моделью Байеса вводится условная зависимость между классами — переходная вероятность, которая определяет зависимость между классами в цепи. Как и в случае модели Байеса, в качестве характеристики модели выступает совместная вероятность (*см. рис. 1*).

Модель максимума энтропии

В работе [1] было введено понятие условной энтропии:

$$H(Y|X) = -\sum_{(X,Y)\in Z} p(Y,X)\log p(Y|X),$$

где множество $Z = X \times Y$, которое нашло практическое приложение в работах [2,3] для задач идентификации классов.

Базовая идея максимальной энтропии — найти такую модель $p^*(X|Y)$, которая делает энтропию максимальной на обучающих данных, т.е.

$$p^*(Y|X) = \underset{p(Y|X)\in P}{argmax}\, H(Y|X),$$

где P — множество всех моделей, содержащихся в обучающих данных.

Введём некоторые функции случайных переменных $f_i(X,Y)$. По сути, функции $f_i(X,Y)$ определяют некоторые известные нам связи между случайными величинами. Математические ожидания этих функций можно рассчитать двумя способами: с одной стороны, можно предположить, что известна эмпирическая оценка совместной вероятности $\hat{p}(Y,X)$, и тогда справедливо:

$$\hat{E}(f_i) = \sum_{(X,Y)\in Z} \hat{p}(Y,X) f_i(Y,X), \quad (1)$$

а с другой стороны, можно предположить, что известна сама совместная вероятность, и тогда справедливо:

$$E(f_i) = \sum_{(X,Y)\in Z} p(Y,X) f_i(Y,X) = \sum_{(X,Y)\in Z} \hat{p}(X) p(Y|X) f_i(Y,X), \quad (2)$$

где $\hat{p}(X)$ — эмпирическая оценка случайной величины X. Очевидно, что в лучшем случае эти два математических ожидания должны быть равны друг другу.

Ещё надо отметить, что выполняется нормальное статистическое условие $\sum_Y p(Y|X) = 1$, где суммирование ведётся по всему пространству переменных *Y*.

Всё сказанное выше означает, что для поиска модели $p^*(Y|X)$ мы можем поставить задачу

$$\Lambda(p,\lambda) = H(Y|X) + \sum_{i=1}^{m} \lambda_i \left(\hat{E}(f_i) - E(f_i)\right) + \lambda_{m+1}\left(\sum_Y p(Y|X) - 1\right). \quad (3)$$

оптимизации с неопределёнными коэффициентами Лагранжа:

$$H(Y|X) = -\sum_{(X,Y)\in Z} \hat{p}(X) p(Y|X)\log p(Y|X),$$

Аппроксимируем выражение для энтропии:

$$\frac{\partial}{\partial p(Y|X)} H(Y|X) = -\hat{p}(X)(1 + \log p(Y|X)),$$

$$\frac{\partial}{\partial p(Y|X)} \sum_{i=1}^{m} \lambda_i \left(\hat{E}(f_i) - E(f_i)\right) = \sum_{i=1}^{m} \lambda_i \hat{p}(X) f_i(Y,X)$$

и проведём дифференцирование всех членов функционала (3):

$$\frac{\partial}{\partial p(Y|X)}\Lambda(p,\lambda) = -\hat{p}(X)(1+\log p(Y|X)) + \sum_{i=1}^{m}\lambda_i\hat{p}(X)f_i(Y,X) + \lambda_{m+1} = 0.$$

Тогда

$$p(Y|X) = exp\left(\sum_{i=1}^{m}\lambda_i f_i(Y,X)\right)exp\left(\frac{\lambda_{m+1}}{\hat{p}(X)} - 1\right),$$

Если провести экспоненцирование последнего выражения, то получим:

$$p(Y|X) = \frac{exp(\sum_{i=1}^{m}\lambda_i f_i(Y,X))}{\sum_Y exp(\sum_{i=1}^{m}\lambda_i f_i(Y,X))}. \quad (4)$$

и, учитывая нормальное статистическое условие, окончательно найдём оптимальную модель, максимизирующую условную энтропию:

Модель максимума энтропии, как и модель Байеса, предназначена для идентификации одного класса и для этой идентификации используется условная вероятность (*см. рис. 1*). Модель условных случайных полей является расширением модели максимума энтропии в условиях описания последовательности наблюдений, как последовательности классов.

Графическое представление

Отвлечёмся от изложения статистических моделей идентификации классов и их последовательностей и рассмотрим графическое представление совместной вероятности, которое предложил Бишоп в работе [3].

$$p(\vartheta) = \prod \Psi_s(\vartheta_s),$$

Допустим, что совместную плотность распределения вероятностей множества случайных величин можно факторизовать, т.е. представить в виде:

где $\Psi_s(\vartheta_s)$ — факторы. Пусть, например совместную вероятность $p(x,y,z)$ можно представить в виде $p(x,y,z) = p(x)p(y)p(z|x,y)$, тогда все вероятности в правой части формулы есть факторы. На *рис. 2* показано графическое представление этой факторизации $\psi 1 = p(x), \psi 2 = p(y), \psi 3 = p(z|x,y)$.

В таком графическом представлении можно представить скрытую модель Маркова (*см. рис. 3*).

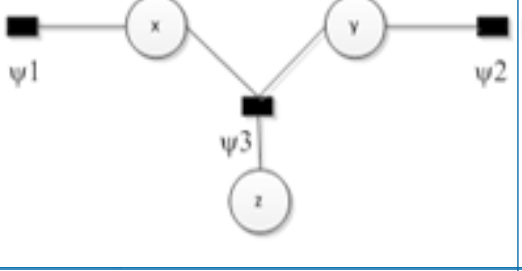

Рис.2.Графическое представление факторизации совместной вероятности p(x,y,z)=p(x)p(y)p(z|x,y)

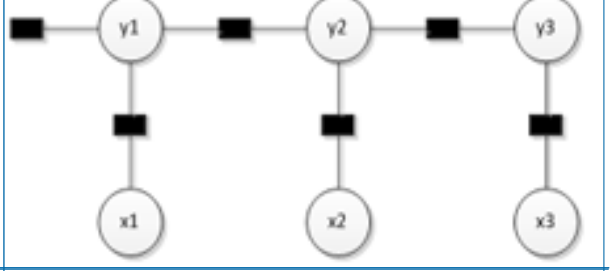

Рис. 3. Графическое представление факторизации модели Маркова

Условные случайные поля

Понятие условного случайного поля введено Лафферти в работе [4].

$$p(Y|X) = \frac{p(Y,X)}{P(X)} = \frac{p(Y,X)}{\sum_{Y'} p(Y',X)} = \frac{\prod_C \psi_c(Y,X)}{\sum_{Y'} \prod_C \psi_c(Y',X)},$$

где $\psi(Y,X)$ — некоторая возможная факторизация условной вероятности $p(Y,X)$ Таким образом, удалось выразить условную вероятность через факторизацию совместной вероятности:

$$p(Y|X) = \frac{1}{Z}\prod_C \psi_c(Y,X), \qquad Z = \sum_Y \prod_C \psi_c(Y,X).$$

Рассмотрим важное понятие в области условных случайных полей — это линейная цепь, структура которой показана на *рис. 4*.

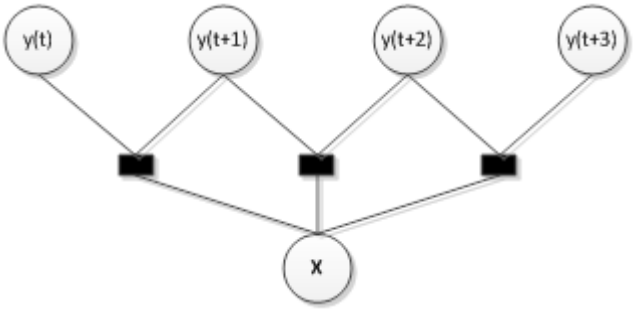

Рис. 4. Графическое представление факторизации линейной цепи условного случайного поля

Условное случайное поле линейной цепи можно записать в виде:

$$p(Y|X) = \frac{1}{Z(X)}\prod_{j=1}^{n} \psi_j(Y,X), \quad \text{где} \quad Z(X) = \sum_Y \prod_{j=1}^{n} \psi_j(Y,X)$$

и на основании (4) факторы имеет форму:

$$\psi_j(Y,X) = exp\left(\sum_{i=1}^{m} \lambda_i f_i(y_{j-1}, y_j, X, j)\right)$$

Предполагая, что длина последовательности наблюдения равна n+1, то для условной вероятности справедливо:

$$p(Y|X) = \frac{1}{Z(X)} exp\left(\sum_{j=1}^{n}\sum_{i=1}^{m} \lambda_i f_i(y_{j-1}, y_j, X, j)\right),$$

где

$$Z(X) = \sum_Y exp\left(\sum_{j=1}^{n}\sum_{i=1}^{m} \lambda_i f_i(y_{j-1}, y_j, X, j)\right).$$

В последнем выражении суммирование по *Y* означает суммирование по всевозможным последовательностям *Y*.

Процесс обучения линейной цепи состоит в максимизации логарифмического правдоподобия по всем путям, наблюдающимся в обучающих данных.

$$L = \sum_{(X,Y)\in Tr} \log p(Y|X) = \sum_{(X,Y)\in Tr} \log\left(\frac{exp\left(\sum_{j=1}^{n}\sum_{i=1}^{m}\lambda_i f_i(y_{j-1},y_j,X,j)\right)}{\sum_Y exp\left(\sum_{j=1}^{n}\sum_{i=1}^{m}\lambda_i f_i(y_{j-1},y_j,X,j)\right)}\right) - \sum_{i=1}^{m}\frac{\lambda_i^2}{2\sigma^2}$$

Последний член функционала позволяет ограничить решения лишь теми, которые находятся на гиперсфере радиусом $\sqrt{2\sigma^2}$.

Продифференцируем последний функционал по :λ_k:

$$\frac{\partial}{\partial\lambda_k}L = \sum_{(X,Y)\in Tr}\sum_{j=1}^{n} f_k(y_{j-1},y_j,X,j) - \sum_{(X,Y)\in Tr}\sum_{Y} p(Y|X)\sum_{j=1}^{n} f_k(y_{j-1},y_j,X,j) - \frac{\lambda_k}{\sigma^2} = 0$$

В предпоследнем слагаемом сначала проводится суммирование по всевозможным последовательностям *Y*, это означает, что второе суммирование проводится только по последовательностям *X*, которые входят в обучающую базу.

Последнее выражение равносильно:

$$\frac{\partial}{\partial\lambda_k}L = \hat{E}(f_k) - E(f_k) - \frac{\lambda_k}{\sigma^2} = 0 \qquad (5)$$

Вычисление параметра , как уже было сказано выше, связано с интегрированием значений функций и экспериментально полученных оценок распределений (1). Для вычисления используется известный алгоритм прямого и обратного хода [8], где прямую переменную можно записать в форме:

$$\alpha_j(s|X) = \sum_{s1\in T_j^{-1}(s)} \alpha_{j-1}(s1|X)\,\psi_j(X,s1,s),$$

и обратную переменную в форме:

$$\beta_j(s|X) = \sum_{s1\in T_j(s)} \beta_{j+1}(s1|X)\,\psi_j(X,s1,s),$$

где

$$\psi_j(X,s1,s) = exp\left(\sum_{j=1}^{n}\sum_{i=1}^{m}\lambda_i f_i(s,s1,X,j)\right),$$

приведут к новой форме записи:

$$E(f_k) = \sum_{(X,Y)\in Tr}\frac{1}{Z(X)}\sum_{j=1}^{n}\sum_{s\in S}\sum_{s1\in T_j(s)} f_i(s,s1,X,j)\alpha_j(s|X)\psi_j(X,s,s1)\beta_{j+1}(s1|X),$$

где $Z(X)=max\beta_0(X)$ — нормализующий фактор.

Таким образом, подставляя последнее выражение в (5), мы получим окончательную формулу для неизвестных коэффициентов λ_k:

Заключительная важная часть построения модели условного случайного поля состоит в том, чтобы найти наилучшую последовательность *Y* при заданной последовательности наблюдений *X*. Для решения этой проблемы используется известный алгоритм Витерби [8], который состоит в поиске максимума условной вероятности:

$$\delta_j(s|X) = \max_{y_1, y_2, \dots y_{j-1}} p(y_1, y_2, .., y_j = s|X),$$

по различным последовательностям имён классов.

Эту процедуру можно представить итеративно для значения текущей вероятности

$$\delta_{j+1}(s|X) = \max_{s1 \in S} \delta_j(s1|X)\psi_{j+1}(X, s, s1),$$

максимум которой в последний момент времени определит искомую условную вероятность наиболее вероятной последовательности классов:

$$p^* = \max_{s \in S} \delta_n(s|X)$$

и индекс класса, в котором заканчивается наиболее вероятная траектория:

$$y^* = \operatorname*{argmax}_{s \in S} \delta_n(s|X).$$

В заключение приведём пример [5], который покажет соответствие между параметрами скрытой модели Маркова и модели условных случайных полей.

Наша цель состоит в том, чтобы конвертировать параметры выражений стоящих в правой и левой части равенства:

$$\prod_{t=1}^{T} p(y_t|y_{t-1})p(x_t|y_t) = \prod_{t=1}^{T} exp\big(w_t f(t, y_t, y_{t-1}, X) + v_t g(t, y_t, x_t, X)\big) \quad (6)$$

Представим себе, что первая часть суммы в степени экспоненты соответствует переходной вероятности $p(y_t|y_{t-1})$. Введём функции индикаторы, с помощью которых функцию $f(t, y_t, y_{t-1}, X)$ запишем в форме:

$$f_{ij}(t, y_t, y_{t-1}, X) = I(y_t = i)I(y_{t-1} = j).$$

Индикаторная функция такова, что если условие внутри скобок — истинно, то она равна единице, и нулю в — противном случае.

Это представление позволяет соотнести переходную вероятность в левой части уравнения (6) и экспоненту в правой части уравнения (6) в каждый момент времени, в которые процесс переходит из состояния *j* в состояние *i:*

$$exp\big(w_{ij} f_{ij}\big) = p(y_t = i|y_{t-1} = j),$$

откуда следует:

$$w_{ij} = log\big(p(y_t = i|y_{t-1} = j)\big).$$

Аналогично можно представить с помощью коэффициентов экспонент параметра распределения вероятностей $p(x_t|y_t)$, если допустить, что оно является нормальным:

$$exp\left(v_t g(t, y_t = i, x_t, X)\right) = \frac{1}{\sqrt{2\pi\sigma}} exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\},$$

тогда

$$v_t g(t, y_t = i, x_t, X) = -\frac{1}{2\sigma^2}x^2 + \frac{\mu}{\sigma}x - \left(\frac{\mu^2}{2\sigma^2} + \log\left(\sqrt{\pi\sigma^2}\right)\right),$$

и функции можно представить в виде:

$$g_{0i} = I(y_t = i),$$
$$g_{1i} = I(y_t = i)\, x_t,$$
$$g_{2i} = I(y_t = i)\, x_t^2.$$

Как видно из приведённого примера, в простейших случаях весьма просто сопоставить параметры различных вероятностных моделей, но уже небольшое усложнение приводит к значительным вычислительным трудностям.

Мы можем это наблюдать уже в том случае, когда попытаемся представить распределение вероятностей $p(x_t|y_t)$, как смесь нормальных распределений.

Литература

1. *Jaynes E.T.*: Information Theory and Statistical Mechanics. In: Physical Review 106 (1957), May, No. 4, P. 620-630.
2. *Korn G.A., Korn T.M.* Mathematical Handbook for Scientists and Engineers: De_nitions, Theorems, and Formulas for Reference and Review. 2 Revised. New York: Dover Publications Inc., 2000.
3. *Bishop C.M.* Pattern Recognition and Machine Learning. Springer, 2006.
4. *Lafferty J.D.; McCallum A.; Pereira F. C. N.* Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In: Proceedings of the Eighteenth International Conference on Machine Learning (ICML 2001), Morgan Kaufmann Publishers, 2001, P. 282–289.
5. *Douglas L. V., C. Guestrin.* Conditional random fields for activity recognition // In Proceedings of the Sixth International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS 2007).
6. *Klinger R., Tomanek K.* Classical Probabilistic Models and Conditional Random Fields// Algorithm Engineering Report TR07–2-013 December, 2007.
7. *C. Sutton and A. McCallum.* An introduction to conditional random fields for relational learning. In L. Getoor and B. Taskar, editors, Introduction to Statistical Relational Learning. MIT Press, 2006.
8. *Jeff A. Bilmes A.* Gentle Tutorial of the EM algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models// ICSI-TR-97–021, April 1998.

Сведения об авторе

Леднов Дмитрий Анатольевич,
кандидат технических наук, старший научный сотрудник, научный консультант научно-технического департамента ООО «Стэл — Компьютерные Системы», основные научные интересы лежат в областях: моделей обработки данных, случайных процессов, распознавания речи и идентификации дикторов.