УДК 519.8

Г. А. Степанкова, І. В. Баклан

НАВЧАННЯ ПРИХОВАНИХ МАРКІВСЬКИХ МОДЕЛЕЙ

Вступ

Приховані марківські моделі (ПММ) останнім часом сталі дуже популярні. Перші відомості про приховані марківські моделі опублікував Баум з колегами ще наприкінці 1960-х років. На початку 70-х Бейкер і Джелінек з колегами в IBM застосували ПММ до розпізнавання мови. З середини 1980-х років ПММ застосовуються при аналізі біологічних послідовностей, зокрема ДНК. Проте, широкого поширення ПММ набули зовсім недавно.

Свою популярність ПММ отримали завдяки тому, що, по-перше, математична структура ПММ дуже багата і дозволяє вирішувати математичні проблеми різних областей науки. По-друге, грамотно спроектована модель дає на практиці добрі результати роботи. У цій статті ми розглянемо алгоритми навчання ПММ та дамо їм оцінку.

Постановка проблеми

ПММ є одним із способів отримання математичної моделі деякого спостережуваного сигналу. ПММ відносяться до класу стохастичних моделей. Стохастичні моделі намагаються охарактеризувати тільки статистичні властивості сигналу, не володіючи інформацією про його специфічні властивості. Настроєну ПММ можна розглядати як джерело деякого випадкового сигналу з цілком певними характеристиками. Також, для настроєної ПММ є можливість підрахувати ймовірність генерації тестового сигналу даною моделлю.

У додатку до завдання розпізнавання, представивши вектор ознак об'єкту у вигляді сигналу (набору послідовних спостережень), можна змоделювати клас об'єктів за допомогою ПММ. Ймовірність належності тестового об'єкту класу, заданому ПММ оцінюється ймовірністю генерації сигналу, відповідного до його вектору ознак. Навчання прихованих харківських моделей – полягає в модифікації її параметрів для того, щоб добитися максимальної ймовірності генерації сигналів, відповідних до векторів тренувального набору.

Визначимо ПММ як $\lambda = (A, B, \pi)$, де A – матриця ймовірності переходу, B – матриця ймовірності спостережуваних об'єктів, π - вектор ймовірності початкових станів. Задача навчання ПММ полягає в тому, щоб підібрати параметри моделі λ так, щоб вона правильно розпізнавала початкові дані $O = O^1, O^2, \ldots, O^K$. Необхідно вибрати один із способів навчання [1]:

1. розпізнати послідовності спостережень, порівняти результати розпізнавання $Q^{*1}, Q^{*2}, \ldots, Q^{*K}$ з правильними відповідями $Q^1, Q^2, \ldots, Q^K$, обчислити в деякому сенсі середню помилку та мінімізувати її, модифікуючи λ;
2. розпізнати послідовність спостережень і максимізувати функцію правдоподібності спостереження послідовності O, в припущенні, що послідовність прихованих станів знайдена правильно. Тобто максимізувати $\prod_{k=1}^{K} P(O^k | Q^{*k})$, модифікуючи λ;
3. максимізувати функцію правдоподібності від спостережень, тобто максимізувати $P(O) = \prod_{k=1}^{K} P(O^k)$, модифікуючи λ.

Перший спосіб навчання – це навчання з вчителем. Його можна проводити, наприклад, методом градієнтного спуску, і він всім гарний, окрім своєї трудомісткості. Решта два способи – це навчання без вчителя, хоча в другому способі можна використовувати вчителя. Найчастіше застосовується третій спосіб навчання (іноді з подальшим донавчанням іншими способами), оскільки для нього відомий швидкий алгоритм.

Це алгоритм, в загальній ситуації званий максимізація очікування (МО, EM – expectation maximization) або, конкретно для ПММ, алгоритмом Баума-Велша. Даний алгоритм є ітеративним і сходиться до локального максимуму правдоподібності. Окрім цього методу вирішення можливо так само використовувати і інші методи оптимізації для пошуку максимуму функції правдоподібності. Проте, в [2] відмічено, що алгоритм Гауса-Ньютона має тенденцію до незбіжності і залежить від хорошого початкового вибору параметрів моделі λ. Там же окрім алгоритму Баум-Велша розглянутий метод

градієнтного спуску Болді-Чавіна. У [3] проводиться аналогія між МО алгоритмом і градієнтними методами.

Згідно [4] для ефективного обчислення ймовірності спостереження послідовності O, яка породжена послідовністю Q, рівною

$$P(Q|\lambda) = P(q_1, q_2, \dots, q_T) = \pi_{q_1} \prod_{t=1}^{T-1} a_{q_t q_{t+1}} \quad (1)$$

використовують алгоритм прямого-зворотного проходу (Forward-Backward). Для цього алгоритму існують дві модифікації рівноцінні за обчислювальними витратами – алгоритм прямого ходу і алгоритм зворотного ходу. Ці алгоритми розрізняються вибором ведучої змінною, прямою або зворотною, яка предпочтительней у кожному конкретному випадку.

Алгоритм прямого ходу: Введемо пряму змінну $\alpha_t(i)$, яку визначимо для заданої моделі λ як значення ймовірності того, що до моменту часу t спостерігалася послідовність $o_1 o_2 \dots o_t$ і у момент t система знаходиться в стані S_i:

$$\alpha_t(i) = P(o_1 o_2 \dots o_t, q_t = S_i|\lambda) \quad (2)$$

Значення прямої змінною обчислюються за наступним алгоритмом:

1. Ініціалізація: $\alpha_1(i) = \pi_i b_i(o_1)$, $1 \le i \le N$.
2. Для всіх $t = 1, 2, \dots, T-1$; $1 \le i \le N$; маємо $\alpha_{t+1}(j) = \left[\sum_{i=1}^{N} \alpha_t(i) a_{ij}\right] b_j(o_{t+1})$.
3. Обчислюємо шукану ймовірність: $P(O|\lambda) = \sum_{i=1}^{N} \alpha_T(i)$.

Алгоритм зворотного ходу: Введемо зворотну змінну $\beta_t(i)$, яку визначимо як умовну ймовірність спостереження послідовності, починаючи з моменту $t+1$ до T за умови, що у момент часу t система знаходиться у стані S_i:

$$\beta_t(i) = P(o_{t+1} o_{t+2} \dots o_T | q_t = S_i, \lambda).$$

Значення зворотної змінної знаходяться так:

1. Ініціалізація: $\beta_T(i) = 1$, $1 \le i \le N$.

2. Для всіх $t = T-1, T-2, \ldots, 1$; $1 \le i \le N$; маємо $\beta_t(i) = \sum_{j=1}^{N} a_{ij} b_j(o_{t+1}) \beta_{t+1}(j)$.

3. Обчислюємо вірогідність: $P(O|\lambda) = \sum_{i=1}^{N} \pi_i b_i(o_1) \beta_1(i)$.

Одним з способів відновлення послідовності прихованих станів є алгоритм Вітербі. Він складається з прямого і зворотного проходів. Введемо наступну змінну

$$\delta_t(i) = \max_{q_1 q_2 \ldots q_{t-1}} P(q_t = S_i | q_1 q_2 \ldots q_{t-1}, o_1 o_2 \ldots o_t, \lambda), \quad (3)$$

що має сенс максимальної ймовірності того, що при заданих спостереженнях до моменту t послідовність станів завершиться у момент часу t в стані S_i, а також змінну $\psi_t(j)$ для зберігання аргументів, що максимізували $\delta_t(j)$. Алгоритм:

1. Ініціалізація: $\delta_1(i) = \pi_i b_i(o_1), 1 \le i \le N$, $\psi_1(i) = 0$.
2. Індуктивний перехід: $\delta_t(j) = \max_{1 \le i \le N} [\delta_{t-1}(i) a_{ij}] b_j(o_t), 1 \le j \le N$, $2 \le t \le T$, $\psi_t(j) = arg \max_{1 \le i \le N} [\delta_{t-1}(i) a_{ij}]$.
3. Останов: $P^* = \max_{1 \le i \le N} [\delta_T(i)]$ – найбільша ймовірність спостереження послідовності $o_1 o_2 \ldots o_T$, яка досягається при проходженні деякої оптимальної послідовності станів $Q^* = (q_1^*, \ldots q_T^*)$, для якої до даного моменту відомий тільки останній стан: $q_T^* = arg \max_{1 \le i \le N} [\delta_T(i)]$.
4. Зворотний перехід. Відновлення оптимальної послідовності станів: $q_t^* = \psi_{t+1}(q_{t+1}^*)$, $t = T-1, T-2, \ldots, 1$.

На етапі навчання ПММ необхідно провести настройку параметрів моделі $\lambda = (A, B, \pi)$ так, щоб максимізувати $P(O|\lambda)$. Для вирішення даного завдання можна використовувати метод Баум-Велша. У [5], [2], [6], [7], [8], [1] достатньо детально приведено виведення формул для цього методу. Фактично необхідно підібрати послідовність прихованих станів Q до послідовності спостережень O [9], тобто вирішити завдання з данними, яких не вистачає. Цю проблему вирішують за допомогою МО алгоритму, який орієнтований на пошук максимуму функції правдоподібності за допомогою

 ISSN 1562-9945

параметрів неспостережуваної функції розподілу для множини спостережень, де дані не повні або є пропуски. Для моделі сумішей загальна схема МО алгоритму приведена в [5] [10]. У такій моделі дані, яких не вистачає – це змінні, які вказують, з якого компоненту суміші витягнутий елемент даних.

Введемо змінну $\xi_t(i,j) = P(q_t = S_i, q_{t+1} = S_j | O, \lambda)$, яка є ймовірністю того, що при заданій послідовності спостережень O система в моменти часу t і $t+1$ знаходитиметься відповідно в станах S_i і S_j. Використовуючи пряму і зворотну змінні, визначені вище, можна записати:

$$\xi_t(i,j) = \frac{\alpha_t(i) a_{ij} b_j(o_{t+1}) \beta_{t+1}(j)}{P(O|\lambda)} = \frac{\alpha_t(i) a_{ij} b_j(o_{t+1}) \beta_{t+1}(j)}{\sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_t(i) a_{ij} b_j(o_{t+1}) \beta_{t+1}(j)}.$$

Введемо наступну змінну, що є апостеріорною ймовірністю того, що при заданій послідовності спостережень O система у момент часу t знаходитиметься у стані S_i: $\gamma_t(i) = \sum_{j=1}^{N} \xi_t(i,j)$.

Введені величини мають наступні властивості: $\sum_{t=1}^{T-1} \gamma_t(i)$ - очікуване число переходів із стану S_i; $\sum_{t=1}^{T-1} \xi_t(i,j)$ - очікуване число переходів із стану S_i в стан S_j.

На основі цих властивостей отримуємо формули переоцінки параметрів марківської моделі:

$$\pi_i^{\theta} = \gamma_t(i), \quad a_{ij}^{\theta} = \frac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\sum_{t=1}^{T-1} \gamma_t(i)}, \quad b_i^{\theta}(k) = \frac{\sum_{t=1, o_t=k}^{T-1} \gamma_t(j)}{\sum_{t=1}^{T-1} \gamma_t(j)} \qquad (4)$$

В процесі застосування цих формул можуть бути тільки два випадки:

1. $\lambda = \lambda^{\theta}$ – точка екстремуму,
2. $P(O|\lambda^{\theta}) > P(O|\lambda)$, тобто правдоподібність появи даної послідовності спостережень для моделі з переоціненими параметрами вище, ніж для початкової моделі.

Безпосередньо алгоритм Баум-Велша складається з 3 основних етапів:

1. Алгоритм прямого-зворотного проходу.
2. Переоцінка параметрів моделі □ (4).

3. Доки не досягнутий поріг збіжності: $|P(O|\lambda)^{\text{ітерація}-1} - P(O|\lambda)^{\text{ітерація}}| > \varepsilon$, повторювати кроки 1 та 2.

Початкові значення параметрів A та π моделі можна задавати довільно, враховуючи імовірнісні нормування. Алгоритм навчання завжди сходиться і при цьому майже завжди – до точки локального максимуму правдоподібності $P(O|\lambda)$. Проте залишаються питання, на які немає універсальної відповіді: з якою швидкістю сходиться алгоритм навчання і чи завжди таке навчання забезпечує хороше розпізнавання.

Висновки

Ми розглянули теоретичні підходи до навчання прихованих харківських моделей та алгоритми навчання ПММ. Був зроблений деякий огляд літератури по даній тематиці.

Також був отриманий наступний результат: окрім визначення оптимальної послідовності прихованих станів алгоритм Вітерці збільшує значення функції правдоподібності при навчанні ПММ. Підключення на останньому етапі навчання алгоритму Баума-Вельша призводить до кінцевого виводу значення функції правдоподібності на максимальне значення.

Нами ведуться роботи по впровадженню апарату прихованих марківських моделей сумісно з гібридними нейронними мережами для задач прогнозування на фінансово-економічних часових рядах. Цій темі буде присвячений окремий доклад.

ЛИТЕРАТУРА

1. Мерков А. Б., Основные методы, применяемые для распознавания рукописного текста. http://www.recognition.mccme.ru/pub/RecognitionLab.html/methods.html
2. Clote P., Formerly, Backofen R. Computational Molecular Biology, An Introduction. England, John Wiley & Sons Ltd, 2000 – 306 pp.
3. Lei Xu, Jordan M. I., On Convergence Properties of the EM Algorithm for Gaussian Mixtures// Neutral Computation, 8, 129-151, 1996

4. Rabiner L. R. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition// Proceedings of the IEEE, 1989, vol. 77, no. 2, p. 257-285.
5. Bilmes J. A. A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov models// Technical Report 97-021, International Computer Science Institute Berkley CA, 1998.
6. Kapadia S. Discriminative Training of Hidden Markov Models// Downing College, Dissertation submitted to the University of Cambridge for the degree of Doctor of Philosophy, 1998
7. Kinscher J., Trebbe H. The Munster Tagging Project Mathematical Background// Arbeitsbereich Linguistik University of Miinster, 1995
8. Luettin J. Visual Speech And Speaker Recognition// Department of Computer Science, University of Sheffield, Dissertation submitted to the University of Sheffield for the degree of Doctor of Philosophy, 1997
9. Форсайт Д. А., Понс Д. Компьютерное зрение. Современный подход – 928 с.
10. Айвазян С. А., Бухштабер В. М., Енюков И. С., Мешалкин Л. Д. Прикладная статистика. Классификация и снижение размерности. М.: Финансы и стати-стика, 1989 – 607 с.
11. Баклан І.В., Рифа В.М., Гибридні моделі в статистичних методах розпізнавання образів// Вестник ХГТУ №3(19), 2003 – с. 26-28.