УДК 004.056.53

ОПИС АЛГОРИТМУ ДЛЯ ПОБУДОВИ НАЇВНОГО БАЄСОВСЬКОГО КЛАСИФІКАТОРА

Р. М. Онищук І. М. Терещенко

[1] *Національний технічний університет України «Київський політехнічний інститут»*

Анотація

У даній статті розглядається в найближчому наближенні один з найбільш популярних алгоритмів класифікації тексту – Наївний Баєсовський класивікатор. Метою статті є розгляд теоретичних основ порушеного алгоритму, а так само опис алгоритму для подальшого використання на його практиці.

Ключові слова: класифікатор, ймовірність, апостеріорна та апріорна ймовірність

ВСТУП

З кожним днем об'єми інформації зростають з неймовірною швидкістю, поширення технологій і доступу до Інтернету привели до подвоєння обсягу інформації за останні 2 роки і за певними прогнозами об'єм інформації кожні 2 – 3 роки буде подвоюватись і зі зростом об'ємів інформації загострюється актуальність завдання їх автоматичної класифікації. Поміж багатьох методів класифікації інформації ми розглянемо Наївний Баєсовський класивікатор. В основі запропонованого методу лежить теорема Байеса - одна з основних теорем теорії ймовірностей, яка дозволяє визначити ймовірність якої-небудь події за умови, що сталась інша статистично взаємозалежна з ним подія. Іншими словами, теорема дозволяє розрахувати ймовірність того, що якась конкретна подія з'явилася внаслідок якоїсь конкретної події.

1. ТЕОРЕМА БАЄСА

Нехай X = {x_1,x_2, . . . , x_n} – вибірка, компоненти якої являють собою значення, на множині n атрибутів. Нехай H - деяка гіпотеза, як, наприклад, значення X належить до певного класу C. Для вирішення класифікаційних проблем, нашою метою є визначення P$H \mid X$, ймовірності, що гіпотеза H містить дані значення, тобто дані з вибірки X. Іншими словами, ми шукаємо ймовірність того, що зразок X належить класу C,з урахуванням того, що ми знаємо опис атрибутів X.

P(H|X) - апостеріорна ймовірність, ймовірність того, що значення H обумовлене певним атрибутом X. Наприклад, припустимо, що у нас є певна вибірка даних з атрибутами: вік і дохід, і об'єктом нашого спостереження, буде n-річний чоловік, з певною заробітною платою.

Припустимо, що H - гіпотеза, що наш клієнт купуватиме комп'ютер. Тоді P(H|X) - це умовна ймовірність, того, що клієнт X буде купувати комп'ютер, за умови, що ми знаємо вік і дохід клієнта.

На відміну від цього, P(H) - апріорна ймовірність H. Для нашого прикладу, це ймовірність того, що будь-який клієнт буде купувати комп'ютер, незалежно від віку, доходу, або іншої інформації. Апостеріорна ймовірність P(H|X) базується на більш докладній інформації (про клієнта), ніж апріорна ймовірність P (H), яка є незалежною від значень X.

Згідно з теоремою Байеса, ймовірність, котру ми хочемо,а саме P(H|X),можна виразити в термінах ймовірностей P(H), P(X|H) і P(X), як

$$P(H \mid X) = \frac{P(X \mid H)P(H)}{P(X)}, \qquad (1)$$

і ці ймовірності можуть бути оцінені з наведених даних.

2. ПОБУДОВА НАЇВНОГО БАЄСІВСЬКОГО КЛАСИФІКАТОРА

Наївний байесовский класифікатор працює таким чином:

1) Нехай T- навчальний набір виборок, кожен зі своїми значеннями класу. Існує k класів, C_1,C_2,...,C_k. Кожна вибірка представлена n-мірним вектором X = {x_1,x_2, . . . , x_n}, з відповідними зображеннями значень на множині з n атрибутів, A_1, A_2, . . . ,A_n.
2) Для даної вибірки X, класифікатор передбачить, чи належить X до классу з найвищою апостеріорною ймовірнісю, базуючись на значенях атрибутів X. X належатиме до класу C_i тоді і тільки тоді, коли виконується умова:

$$P(C_i|X) > P(C_j|X) for 1 \leq j \leq m, j \neq i. \qquad (2)$$

Таким чином, ми знаходимо клас, який максимізує ймовірність P($C_i \mid X$). Клас C_i, для якого

P(C_i | X) є максимальним будемо називати класом з максимальною апостеріорною гіпотезою. За теоремою Баєса:
$P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)}$.

3) Враховуючи той факт, що ймовірність P(X) – константна, тобто є однаковою для всіх класів, то тільки $P(X \mid C_i)P(C_i)$ має бути максимізована. Апріорна ймовірність класу $P(C_i)$ може бути оцінена, як $P(C_i) = freq(C_i, T)/|T|$,тобто частотою зустрічання C_i на множині T.
4) Ймовірності P($x_1 \mid C_i$),P($x_2 \mid C_i$),...,P($x_n \mid C_i$) можуть бути легко оцінені по навчальній вибірці, де x_k - це значення атрибуту A_k виборки X. Звідси,ми отримуємо два випадки:
   - Якщо значення A_k - дискретна(категорійована) змінна (тобто A_k прийматиме значення з певної фіксованої множини), то $P(x_k|C_i)$ це кількість зразків класу C_i в множинні значень навчальної вибірки T, що мають значення x_k для атрибуту A_k, розділене на частоту(C_i,T), кількості C_i-тих зразів класу C в T.
   - Якщо значення A_k безперервна (цифрова) змінна (тобто така, яка може приймати будь-яке значення в безперервному діапазоні), то зазвичай припускається, що змінні мають Гауссівський розподіл з середнім μ і стандартним відхиленням σ, і визначається формулою
     $g(x, \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma}} exp - \frac{(x-\mu)^2}{2\sigma^2}$,
     так що:
     $P(x_k|C_i) = g(x_k, \mu C_i, \sigma C_i)$.
     Ми повинні обчислити μC_i і σC_i , які являють собою середнє і стандартне відхилення значень атрибуту A_k для навчальних виборок класу C_i.
5) Класифікатор прогнозує, що клас C_i буде еталонному у виборці X, тоді і тільки тоді, коли ми отримаєм максимальне значення ймовірності $P(X|Ci)P(Ci)$. З чого можна зробити висновок, що у найпростішому вигляді, Наївний баєсовський класифікатор можна зобразити як
$P(X|C) = \text{argmax}_{j=1..n}\, P(C_j) \prod_{k=1}^{n} P(x_k|C_j)$

ПРОБЛЕМА "НУЛЬОВОЇ ЧАСТОТИ"

Лапласова корекція - спосіб боротьби з нульовою ймовірністю змінних, який відповідає на питання: що робити, якщо X має значення атрибуту xk, таке, що для нього не існує відповідного значення атрибуту в класі C? У цьому випадку $P(x_k|C_i) = 0$,що призводить до $P(X|C_i) = 0$, хоча $P(x_k|C_i)$ для всіх інших атрибутів в X може бути великим. Існує простий трюк, щоб уникнути цієї проблеми. Ми можемо припустити, що наш навчальний набір настільки великий, що додавши до кожної змінної, значення котрої нам треба підрахувати, одиницю привнесе мізерні зміни в розрахунок ймовірностей, але допоможе оминути отримання нульової ймовірності. Якщо у нас є q змінних, до якихми додаємо одиницю, то як результат потрібно додати q до відповідного знаменника при обрахунку ймовірностей . Отримана "підкоректована" за лапласом ймовірність практично не буде відрізнятись, зате буде уникнуто нульової ймовірності.

ВИСНОВКИ

Незважаючи на доволі дивне припущення взаємної незалежності атрибутів, наївний Байесівський класифікатор є одним з найефективніших класифікаторів на практиці, його класифікація може бути найточнішою серед інших подібних йому класифікаторів, навіть не зважаючи, на те, що ймовірність певних атрибутів може бути розрахована неточно. Нами був описаний алгоритм роботи як наївного баєсовського класифікатора, так і формули Баєса, на якій він базується. Хочеться відмітити що класифікатор можна дуже легко навчати, що в сукупності з його відносною простотою реалізації і високою ефективністю робить його одним з найпростіших в реалізації та одним з найефективніших у класифікації великих об'ємів даних.

ПЕРЕЛІК ВИКОРИСТАНИХ ДЖЕРЕЛ

1) M. Kantardzic, Data Mining - Concepts, Models, Methods, and Algorithms, IEEE Press, Wiley-Interscience, 2003, ISBN 0-471- 22852-4.
2) Jiawei Han and Micheline Kamber, Data Mining: Concepts and Techniques, Elsevier 2006, ISBN 1558609016. This part of the lecture notes is derived from chapter 6.4 of this book.
3) Kononenko, I. 1990. Comparison of inductive and naive Bayesian learning approaches to automatic knowledge acquisition. In Wielinga, B., ed., Current Trends in Knowledge Acquisition. IOS Press.
4) Pazzani, M. J. 1996. Search for dependencies in Bayesian classifiers. In Fisher, D., and Lenz, H. J., eds., Learning from Data: Artificial Intelligence and Statistics V. Springer Verlag.
5) Субботин С. В., Большаков Д. Ю. Применение байесовского классификатора для распознавания классов целей. // «Журнал Радиоэлектроники», 2006, № 4