ỨNG DỤNG PHÂN TÍCH DỮ LIỆU LỚN

Giảng viên: Nguyễn Tu Trung, Trần Mạnh Tuấn
BM HTTT, Khoa CNTT, Trường ĐH Thủy Lợi

Hà Nội, 2019

Nội dung

- Phân lớp văn bản với Naïve Bayes
- Phân cụm ảnh với MapReduce_K-Means
- Tra cứu thông tin từ internet

Phân lớp văn bản với Naïve Bayes

- Giới thiệu về phân lớp văn bản
- Các bước thực hiện:
  - Huấn luyện mô hình phân lớp văn bản
  - Phân lớp văn bản
- Ví dụ minh họa phân lớp văn bản
- Thực thi phân văn bản với ví dụ

Giới thiệu về phân lớp văn bản

- Dùng để dự đoán xem văn bản thuộc chủ đề nào, dựa trên nội dung văn bản
- Để áp dụng thuật toán Naïve Bayes vào phân loại văn bản, cần thực hiện các bước tiền xử lý và vector hoá các văn bản (trong tập huấn luyện hay văn bản dùng để phân lớp)
- Vector hoá các văn bản: đếm tần suất từ trong văn bản
- Nếu số chiều quá lớn, có thể thực hiện rút chọn các đặc trưng
- Vector đặc trưng được dùng để huấn luyện

Huấn luyện mô hình phân lớp văn bản

- Từ tập huấn luyện, ta rút trích:
  - Tập từ vựng (các đặc trưng) vocSet
  - Vector đặc trưng của từng văn bản
- Tính xác suất $P(C_i)$ và $P(x_k|C_i)$
  - $P(C_i) = \frac{totalDocC_i}{totalDocAll}$
  - $P(x_k|C_i) = \frac{|x_k|}{|TextC_i|}$ hoặc $P(x_k|C_i) = \frac{|x_k|+1}{n_{C_i}+|TextC_i|}$
  - $totalDocC_i$: số tài liệu của tập huấn luyện thuộc lớp C_i
  - $totalDocAll$: số tài liệu có trong tập huấn luyện
  - n_{C_i}: tổng số từ đôi một khác nhau của lớp C_i
  - $|x_k|$: tổng số từ x_k trong lớp C_i
  - $|TextC_i|$: tổng số từ vựng (không phân biệt đôi một) trong lớp C_i, $|TextC_i| = \sum_{x_k \in \text{vocSet}} |x_k|$

Phân lớp văn bản

- Cho văn bản doc^{new}
- Yêu cầu:
  - Dự báo văn bản doc^{new} thuộc lớp nào ?
- Các bước thực hiện:
  - Tính:
  - $F(doc^{new}, C_i) = P(C_i) * \prod_{x_k \in \text{vocSet}} (P(x_k|C_i) * |x_k|^{new})$
  - Ta có: $p(doc^{new}) = \max(F(doc^{new}, C_i))$, i=1,...,$n_C$

Ví dụ minh họa phân lớp văn bản

- Có tập tài liệu để huấn luyện sau khi đã vector hoá (sử dụng phương pháp đơn giản đếm số lần xuất hiện) và rút trích đặc trưng như sau:
  - Bộ từ vựng (đặc trưng) : var, bit, chip, log

Docs	Var	Bit	Chip	Log	Class
Doc1	42	25	7	56	Math
Doc2	10	28	45	2	Comp
Doc3	11	25	22	4	Comp
Doc4	33	40	8	48	Math
Doc5	28	32	9	60	Math
Doc6	8	22	30	1	Comp

Thực thi phân văn bản với ví dụ

- Có 2 lớp chủ đề:
  - C1 = “Comp”
  - C2 = “Math”
- B1: Huấn luyện chủ đề
- B2: Phân lớp chủ đề

Huấn luyện chủ đề

- Tính xác xuất các lớp C_i:
  - $P(C_1$ = "Comp") = 3/6 = 0.5
  - P(C2 = "Math") = 3/6 = 0.5
- B1: Huấn luyện chủ đề
  - Các xác suất $P(x_k|C_1)$
  - Các xác suất $P(x_k|C_2)$

Các xác suất $P(x_k|C_1)$

- Tổng số từ lớp C_1 = "Comp":
  - $|TextC_1| = (10 + 11 + 8) + (28 + 25 + 22) + (45 + 22 + 30) + (2 + 4 + 1) = 208$
- P(var|Comp) = (10 + 11 + 8) / 208 = 29/208
- P(bit|Comp) = (28 + 25 + 22) / 208 = 75/208
- P(chip|Comp) = (45 + 22 + 30) / 208 = 97/208
- P(log|Comp) = (2 + 4 + 1) / 208 = 7/208

Các xác suất $P(x_k|C_2)$

- Tổng số từ lớp C_2 = "Math":
  - $|TextC_2| = (42 + 33 + 28) + (25 + 40 + 32) + (7 + 8 + 9) + (56 + 48 + 60) = 388$
- P(var|Math) = (42 + 33 + 28) / 388 = 103/388
- P(bit|Math) = (25 + 40 + 32) / 388 = 97/388
- P(chip|Math) = (7 + 8 + 9) / 388 = 24/388
- P(log|Math) = (56 + 48 + 60) / 388 = 164/388

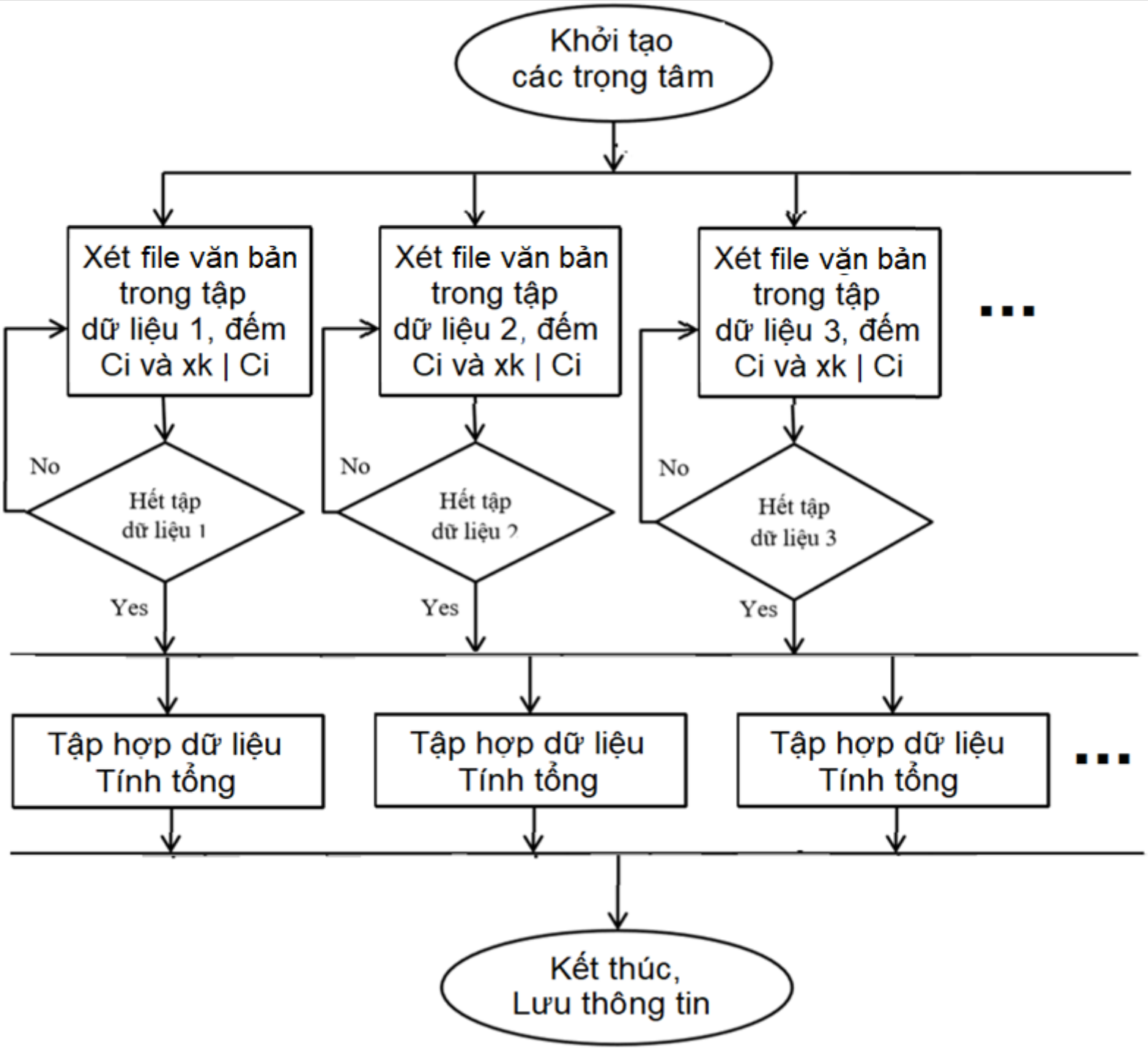
Phân lớp chủ đề

- Cho văn bản có vector đặc trưng Doc^{new} = (23, 40, 15, 50)
- $F(doc^{new}, C_1)$ = P(Math) * [P(var|Math) * 23 * P(bit|Math) * 40 * P(chip|Math) * 15 * P(log|Math) * 50] = 0.5 * [103/388 * 23 * 97/388 * 40 * 24/388 * 15 * 164/388 * 50] = 598.627
- $F(doc^{new}, C_2)$ = P(Comp) * [P(var|Comp) * 23 * P(bit|Comp) * 40 * P(chip|Comp) * 15 * P(log|Comp) * 50] = 0.5 * [29/208 * 23 * 75/208 * 40 * 97/208 * 15 * 7/208 * 50] = 272.204
- Kết luận: Văn bản Doc^{new} thuộc về lớp Math do $p(doc^{new}) = \max(F(doc^{new}, C_i))$ = 598,627

MapReduce hoá thuật toán Bayes

- Nhiệm vụ:
  - MapReduce hóa việc đếm số lần xuất hiện của các C_i và các $x_k|C_i$
- Ý tưởng:
  - Chia dữ liệu thành nhiều phần nhỏ
  - Đếm các số lần xuất hiện của từng biến C_i và $x_k|C_i$ trong hàm Map
  - Tập hợp kết quả và tính tổng theo từng biến trong hàm Reduce
  - Lưu thông tin số lần xuất hiện của từng biến C_i và $x_k|C_i$
  - Giai đoạn phân lớp: tính các xác suất $P(C_i)$ và $P(x_k|C_i)$ dựa trên dữ liệu về số lần xuất hiện của từng biến C_i và $x_k|C_i$ để tính các $F(X_{new}, C_i)$

Lưu đồ thuật toán MapReduce_Bayes



MapReduce hoá thuật toán Bayes

- Dữ liệu đầu vào:
  - Là danh sách các file văn bản (có thể lưu trên file txt)
  - Mỗi hàng là dữ liệu huấn luyện mô tả từng file văn bản, gồm tên lớp và tên file:
    - number D:\\Hadoop\\test\\input\\file1.txt
  - Được chuyển sang kiểu key/value làm đầu vào cho thuật toán
- Mô hình cơ bản của MapReduce:
  - map (keyIn, valIn) -> list (keyInt, valInt)
  - reduce (keyInt, list (valInt)) -> (keyOut, valOut)
- Áp dụng cho thuật toán Bayes:
  - Xây dựng hàm Map_TextBayes
  - Xây dựng hàm Reduce_TextBayes

Xây dựng hàm Map_TextBayes

- Đầu vào:
  - cặp key/value biểu diễn dữ liệu mô tả file văn bản
  - keyIn là giá trị bye offset của dòng
  - valIn là text biểu dữ liệu một file văn bản (number D:\\Hadoop\\test\\input\\file1.txt)
- Xử lý: Tính ValInt
  - Đếm 1 cho xuất hiện của C_i
  - Đếm 1 cho xuất hiện của $x_k|C_i$
- Đầu ra:
  - cặp key/value trung gian
  - keyInt là C_i hoặc $x_k|C_i$
  - valInt là giá trị 1

Xây dựng hàm Reduce_TextBayes

- Trước khi hàm reduce thực hiện
  - Kết quả của hàm map được trộn lại
  - Các cặp cùng keyInt được gom thành một nhóm
- Đầu vào:
  - keyInt được chuyển từ hàm map
  - list(valInt) là list các giá trị 1
- Xử lý:
  - Tính tổng các giá trị 1 trong list(valInt)
- Đầu ra:
  - keyOut là keyInt (C_i hoặc $x_k|C_i$)
  - valOut là tổng các giá trị 1 trong list(valInt)

Phân cụm ảnh với MapReduce_K-Means

- Phát biểu bài toán phân cụm ảnh
- Giải pháp phân cụm ảnh với MapReduce_K-Means

Phát biểu bài toán phân cụm ảnh

- Input: n điểm ảnh và số các cụm k
- Output: Các cụm C_i (i=1... k) (các cụm điểm ảnh) sao cho hàm tiêu chuẩn E đạt giá trị tối thiểu

Phân cụm ảnh với MapReduce_K-Means

- B1: Chuyển đổi dữ liệu
  - Chuyển đổi dữ liệu điểm ảnh thành list các hàng
  - Mỗi hàng là list giá trị là các thành phần của vector biểu diễn cho một điểm ảnh
- B2: Thực hiện phân cụm với MapReduce_K-Means
- B3: Chuyển đổi kết quả phân cụm của MapReduce_K-Means cho dữ liệu ảnh gốc

Tra cứu thông tin từ internet

- Ví dụ:
  - Tra cứu thông tin khách sạn, tham khảo: http://www.trivago.vn
  - Tra cứu thông tin sản phẩm điện máy
- Các bước xây dựng ứng dụng:
  - B1: Thu thập dữ liệu từ internet
  - B2: Lưu vào CSDL NoSQL
  - B3: Xây dựng ứng dụng tra cứu thông tin truy xuất dữ liệu từ CSDL NoSQL