**ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH**

**TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN**

**KHOA CÔNG NGHỆ THÔNG TIN**

**BỘ MÔN KHAI THÁC DỮ LIỆU VÀ ỨNG DỤNG**

**BÁO CÁO :**

**THỰC NGHIỆM VỀ CÁC MÔ HÌNH PHÂN LỚP TRÊN MỘT TẬP DỮ LIỆU CHO TRƯỚC**

**ĐỀ TÀI:**

**SO SÁNH MỘT SỐ MÔ HÌNH PHÂN LỚP TRÊN TẬP DỮ LIỆU CHO TRƯỚC**

**Người thực hiện: Nguyễn Minh Quang**

**Vũ Thành Tâm**

**Phạm Khắc Quyền**

**Giáo viên: TS. Nguyễn Hoàng Khai**

**HCM,10-12-2016**

1. **Giới thiệu:**

* Bài toán mô tả dự đoán số lượt share của các trang tin tức trên mạng. Dựa vào các tiêu chí đánh giá để dự đoán số lượt share bài của các trang tin tức như số từ trong tiêu đề là bao nhiêu, số từ trong nội dung là bao nhiêu, có bao nhiêu ảnh video trong tin tức đó,phân loại theo nội dung dữ liệu về các mảng như giải trí, kinh doanh, truyền thông, xã hội…và từ các tiêu chí đó mà người dùng có thể share bài viết của trang tin tức với số lượt share là bao nhiêu.
* Để thực hiện hóa được mục đích như vậy, thì nhóm chúng tôi đã lấy dữ liệu của các trang tin tức, và sẽ phân tích và xử lý dữ liệu để đưa ra kết luận về sự ảnh hưởng của các tiêu chí, thuộc tính nào để cho trang tin tức có nhiều lượt share bằng các mô hình phân lơp dữ liệu, sau đó so sánh kết quả của các mô hình đó, và xem ưu điểm và nhược điểm của các mô hình khi thực hiện phân lớp dữ liệu mẫu trên

1. **Tổng quan về tập dữ liệu:**

Dữ liệu của chúng tôi sẽ có 39644 trường dữ liệu và với mỗi trường dữ liệu sẽ có 61 thuộc tính trong đó có 58 thuộc tính dự đoán, 2 thuộc tính không dự đoán và 1 thuộc tính làm đích. Chúng tôi phân tích thấy đa số kiểu dữ liệu của các thuộc tính là kiểu numeric nhưng có thuộc tính Share thì thuộc tính bản chất là số nhưng được tham chiếu qua các giá trị chữ.

* Chúng tôi sử dụng các bộ phân lớp như NaiveBayes(Bayes),Cây quyết định(J48), ZeoR(rules),K-NN(lazy)
* Về tập dữ liệu, thì chúng tôi không sử dụng một phép tiền xử lý nào, và sau khi có tìm hiểu các mô hình phân lớp thì chúng tôi nghĩ Phương pháp dùng rules dùng thuật toán K-NN là có thể đạt kết quả nhanh và chính xác nhất. Lí do vì sao chúng tôi lại đánh giá K-NN có thể đạt hiệu quả tốt vì điểm chính của thuật toán là tính khoảng cách giữa tất cả các mẫu trong dữ liệu, xử lý tốt với dữ liệu nhiễu.
* Môi trường để chúng tôi thực hiện bài báo cáo này là chúng tôi dùng công cụ Weka phiên bản 3.8
* Trong thực tế, ta cần áp dụng nhiều thuật toán Machine learning để chọn ra được mô hình phù hợp nhất cho bài toán của mình. Vấn đề đặt ra, làm thế nào để đánh giá và chọn ra các mô hình. Ngoài thuật toán học máy, sự thực thi của mô hình có thể phụ thuộc vào các yếu tố khác như sự phân bố của các lớp, chi phí phân loại sai, kích thước của tập huấn luyện và tập thử nghiệm, độ đo thực thi. Trong bài viết này, ta sẽ đánh giá thực thi: tập trung vào khả năng dự đoán của mô hình hơn là tốc độ phân loại hay xây dựng mô hình, khả năng co giãn.
* Vì vậy, phương pháp thử nghiệm 5-folds cross validation với cách chia ra 5 tập con có cùng kích thước. Tại mỗi vòng lặp sử dụng một tập con là tập thử nghiệm, còn các tập còn lại là tập huấn luyện

1. **Quy trình phân tích:**
2. **Bộ phân lớp NaiveBayes:**

Cho X, C là các biến bất kỳ (rời rạc, liên tục, cấu trúc). Mục tiêu của ta là dự đoán C từ X. Từ mô hình Bayes ta có thể lượng giá các tham số của P(X|C), P(C) trực tiếp từ tập huấn luyện. Sau đó, ta sử dụng định lý Bayes để tính P(C|X=x).

**Độc lập điều kiện (conditional independence):** X độc lập điều kiện với Y khi cho Z nếu phân bố xác suất trên X độc lập với các giá trị của Y khi cho các giá trị của Z. Ta thường viết P(X|Y,Z) = P(X|Z). Ví dụ, P(Sấm sét|Mưa,Chớp) = P(Sấm sét|Chớp).

Giả sử D là tập huấn luyện gồm các mẫu biểu diễn dưới dạng X = <x_1, ..., x_n>. C_{i,D} là tập các mẫu của D thuộc lớp C_i ( i = \{1, ...,m\}). Các thuộc tính x_1, ...,x_n độc lập điều kiện đôi một với nhau khi cho lớp C.

1. **Cây quyết định:**

Cây quyết định (Decision Tree) là một cây phân cấp có cấu trúc được dùng để phân lớp các đối tượng dựa vào dãy các luật (series of rules). Các thuộc tính của đối tượng (ngoại trừ thuộc tính phân lớp – Category attribute) có thể thuộc các kiểu dữ liệu khác nhau (Binary, Nominal, ordinal, quantitative values) trong khi đó thuộc tính phân lớp phải có kiểu dữ liệu là Binary hoặc Ordinal.

Tóm lại, cho dữ liệu về các đối tượng gồm các thuộc tính cùng với lớp (classes) của nó, cây quyết định sẽ sinh ra các luật để dự đoán lớp của các đối tượng chưa biết (unseen data)

1. **K-NN:**

K-Nearest Neighbors algorithm (K-NN) được sử dụng rất phổ biến trong lĩnh vực Data Mining. K-NN là phương pháp để phân lớp các đối tượng dựa vào khoảng cách gần nhất giữa đối tượng cần xếp lớp (Query point) và tất cả các đối tượng trong Training Data.

Một đối tượng được phân lớp dựa vào K láng giềng của nó. K là số nguyên dương được xác định trước khi thực hiện thuật toán. Người ta thường dùng khoảng cách Euclidean để tính khoảng cách giữa các đối tượng.

1. **ZeroR:**
2. **Kết quả thực nghiệm:**
3. **Về độ chính xác:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Phân lớp chính xác | Phân lớp không chính xác | Precision | Recall | F-Measure | Time build |
| ZeroR | 17.1905% | 82.8095% | 0.03 | 0.172 | 0.05 | 0.02 seconds |
| NaiveBayes | 17.6975% | 82.3025% | 0.164 | 0.177 | 0.152 | 0.37 seconds |
| Cây quyết định(J48) | 15.0414% | 84.9586% | 0.150 | 0.150 | 0.150 | 60.52 seconds |
| K-NN | 14.7538% | 85.2462% | 0.147 | 0.148 | 0.147 | 0.01 seconds |
|  |  |  |  |  |  |  |

1. **Phân tích kết quả:**

* Như đã dự đoán kết quả, bộ phân lớp K-NN có thời gian chạy nhanh nhất trong các bộ phân lớp mà chúng tôi thử nghiệm. Tại sao lại có kết quả khá chênh lệch về 2 nhóm, một nhóm có thời gian chạy rất nhanh như ZeroR, NaiveBayes và K-NN và có thuật toán chạy khá chậm như J48
* Điểm khá chung của các thuật toán chạy nhanh trong tập dữ liệu này là các tin tức đều độc lập với nhau, chạy nhanh hơn vì không sử dụng việc kết hợp các từ để đưa ra phán đoán chủ đề

1. **Kết luận:**

* Qua bài phân tích và chạy thử nghiệm với các mô hình phân lớp với cùng một tập dữ liệu thì các mô hình như K-NN và NaiveBayes là tốt nhất,

\*Phân công công việc của các thành viên

|  |  |
| --- | --- |
| **Nguyễn Minh Quang-1412427** | **Tìm hiểu các mô hình và chạy thực nghiệm** |
| **Vũ Thành Tâm-1412484** | **Giới thiệu bài báo cáo** |
| **Phạm Khắc Quyền-1412446** | **Phân tích tập dữ liệu** |