Báo cáo phân tích: Nghiên cứu và so sánh thuật toán phân lớp SVM và Naive Bayes trên bộ dữ liệu Blood Transfusion Service Center

---

1. Giới thiệu

Bài toán phân lớp được áp dụng trên bộ dữ liệu của Trung tâm Dịch vụ Hiến máu tại Hsin-Chu City, Đài Loan, với mục tiêu phân loại khả năng hiến máu của từng cá nhân. Bộ dữ liệu gồm 748 mẫu, mỗi mẫu mô tả các đặc trưng liên quan đến việc hiến máu, bao gồm: số tháng kể từ lần hiến gần nhất (Recency), tổng số lần hiến máu (Frequency), tổng lượng máu đã hiến (Monetary), và số tháng kể từ lần hiến đầu tiên (Time). Biến mục tiêu là một nhãn nhị phân, đại diện cho việc cá nhân có hiến máu trong tháng 3 năm 2007 hay không (1 cho có, 0 cho không).

2. Phương pháp

Hai thuật toán phân lớp được lựa chọn là **Naive Bayes** và **Support Vector Machine (SVM)**. Đây là các thuật toán phân lớp phổ biến, trong đó:

- **Naive Bayes**: Là phương pháp xác suất đơn giản, hiệu quả khi giả định tính độc lập giữa các đặc trưng. Naive Bayes thích hợp với các bài toán có dữ liệu lớn và không yêu cầu chuẩn hóa dữ liệu.
- **SVM**: Là một thuật toán học máy có giám sát mạnh mẽ, hoạt động tốt trên các tập dữ liệu nhỏ và đa chiều. SVM nhạy cảm với các đặc trưng không đồng nhất về đơn vị đo, do đó dữ liệu đầu vào cần được chuẩn hóa.

3. Kết quả

Kết quả của hai mô hình trên tập kiểm tra (30% dữ liệu) được đánh giá dựa trên các chỉ số: ma trận nhầm lẫn, độ chính xác (accuracy), độ chính xác theo lớp (precision), độ nhạy (recall) và điểm F1.

3.1 Kết quả Naive Bayes

- **Ma trận nhầm lẫn**: Mô hình dự đoán đúng 160 trường hợp không hiến máu (lớp 0) và 8 trường hợp có hiến máu (lớp 1). Có 5 trường hợp không hiến máu bị dự đoán nhầm thành có hiến máu, và 52 trường hợp có hiến máu bị dự đoán nhầm thành không hiến máu.
- **Các chỉ số chính**:
    - Độ chính xác tổng thể đạt **75%**.
    - Độ chính xác cho lớp 1 (có hiến máu) là **0.62**, trong khi độ nhạy chỉ đạt **0.13**, thể hiện rằng Naive Bayes gặp khó khăn trong việc nhận diện lớp 1.

3.2 Kết quả SVM

- **Ma trận nhầm lẫn**: Mô hình dự đoán chính xác toàn bộ 165 trường hợp thuộc lớp 0, nhưng không thể phân loại được bất kỳ trường hợp nào thuộc lớp 1.

- **Các chỉ số chính**:
    - Độ chính xác tổng thể đạt **73%**.
    - Precision và recall của lớp 1 là **0**, cho thấy mô hình hoàn toàn không nhận diện được các mẫu thuộc lớp 1.

4. Thảo luận và So sánh

Qua các chỉ số hiệu suất, có thể thấy rằng **Naive Bayes** tỏ ra hiệu quả hơn trong việc phân loại hai lớp so với **SVM**. Naive Bayes đạt độ chính xác cao hơn và có khả năng nhận diện được một phần các mẫu thuộc lớp 1, mặc dù vẫn dự đoán sai đáng kể số lượng mẫu thuộc lớp này.

SVM, mặc dù có khả năng phân loại lớp 0 tốt hơn, lại hoàn toàn thất bại trong việc nhận diện lớp 1. Hiện tượng này có thể là do **bộ dữ liệu không cân bằng** – số lượng mẫu lớp 0 chiếm ưu thế (76%) so với lớp 1 (24%). Điều này có thể dẫn đến sự thiên lệch khi SVM ưu tiên dự đoán lớp 0.

5. Kết luận

Dựa trên phân tích và kết quả, có thể kết luận rằng:

- **Naive Bayes** là thuật toán phù hợp hơn cho bài toán này, đặc biệt là trong bối cảnh bộ dữ liệu không cân bằng. Mặc dù có tỷ lệ dự đoán sai đáng kể ở lớp 1, Naive Bayes vẫn thể hiện khả năng phân loại tốt hơn khi xét đến cả hai lớp.
- **SVM** gặp vấn đề trong việc nhận diện lớp 1 và cần được điều chỉnh hoặc sử dụng các kỹ thuật xử lý dữ liệu không cân bằng (như oversampling hoặc điều chỉnh trọng số mẫu) để cải thiện hiệu suất.

Trong tương lai, việc thử nghiệm các thuật toán khác như **Random Forest** hoặc **Logistic Regression**, cùng với các kỹ thuật xử lý dữ liệu không cân bằng, có thể sẽ mang lại hiệu quả phân loại cao hơn cho bài toán này.