ĐẠI HỌC QUỐC GIA TP HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC BÁCH KHOA

BÁO CÁO BÀI TẬP LỚN MÔN NHẬN DIỆN MẪU VÀ HỌC MÁY

Study the PCA tool of WEKA and apply it in feature extraction

---

GVHD:
PGS.TS. Dương Tuấn Anh

Sinh viên thực hiện:
Nguyễn Quốc Long - MSSV:1770023

TP.Hồ Chí Minh, 08/11/2018.

Mục lục

Bài luận tìm hiểu giải thuật Min-conficts

Nguyễn Quốc Long[1]

MSSV: 1770023

Tóm tắt nội dung. Tài liệu tìm hiểu về giải thật PCA và phần mền WEKA.

Từ khóa: giảm chiều dữ liệu, giải thuật PCA, WEKA

1 Giới thiệu

Giảm chiều dữ liệu (Dimensionality Reduction) là một trong kỹ thuật quan trọng của Học Máy (Machine Learning). Các feature vectors trong các bài toán thực tế có thể có số chiều rất lớn, tới vài nghìn. Ngoài ra, số lượng các điểm dữ liệu cũng thường rất lớn. Nếu thực hiện lưu trữ và tính toán trực tiếp trên dữ liệu có số chiều cao này thì sẽ gặp khó khăn cả về việc lưu trữ và tốc độ tính toán. Vì vậy, giảm chiều dữ liệu sẽ làm tăng tốc độ tính toán nên đây là bước quan trọng trong nhiều bà toán học máy (đây cũng được gọi là phương pháp nén dữ liệu).

Phân tích thành phần chính (Principal Component Analysis (PCA)) là một thuật toán Dimensionality Reduction dựa trên một mô hình tuyến tính. Phương pháp này dựa trên quan sát rằng dữ liệu thường không phân bố ngẫu nhiên trong không gian mà thường phân bố gần các đường/mặt đặc biệt nào đó. PCA xem xét một trường hợp đặc biệt khi các mặt đặc biệt có dạng tuyến tính là các không gian con (subspace).

2 Giới thiệu công cụ Weka

Weka (viết tắt của Waikato Environment for Knowledge Analysis) là một bộ phần mềm học máy được Đại học Waikato, New Zealand phát triển bằng Java. Weka là phần mềm tự do phát hành theo Giấy phép Công cộng GNU.

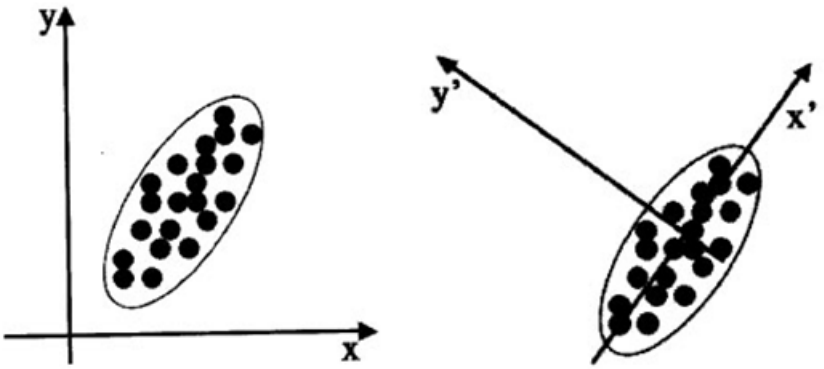
Để chạy chương trình, trước tiên người dùng cần tải Weka (64 bit) và tiến hành các thao tác cài đặt đơn giản. Sau khi hoàn tất quá trình cài đặt, Weka (64 bit) sẽ hiển thị bốn ứng dụng tích hợp cho phép người dùng truy cập, bao gồm "Explorer", "Experimenter", "KnowledgeFlow" và "Simple CLI".

Các tính năng chính của Weka (64 bit):
- Xem và phân tích tập tin dữ liệu ARFF.
- Thực hiện phân nhóm và hồi quy dữ liệu .
- Chỉnh sửa tùy thích hoặc lọc nội dung dữ liệu.
- Thay đổi thuộc tính và ảo hóa kết quả.
- Phân loại dữ liệu sẵn có theo quy tắc định trước.
- Tiến hành phân tích tổng lợi ích/chi phí.
- Lập biểu đồ dữ liệu.
- Khai thác kế hoạch Machine Learning.
- Lưu kết quả theo định dạng ARFF hoặc CSV hay cơ sở dữ liệu JDBC.
- Phân tích, kiểm thử tập tin dữ liệu.

3 Giới thiệu phương pháp PCA

PCA là một thủ tục toán học biến đổi một số thuộc tính tương quan thành một số lượng nhỏ hơn các thuộc tính không tương quan được gọi là các thành phần chính. PCA là một phương pháp của Giảm chiều dữ liệu (Dimensionality reduction)

Mục tiêu của PCA là xác định cơ sở có ý nghĩa nhất để thể hiện lại một tập dữ liệu



Hình 2.7: Hình hai trục tọa độ khác nhau. Tọa độ thứ hai được quay và biến đổi từ tọa độ thứ nhất bằng cách sử dụng PCA

4 Các bước rút trích dữ liệu đặc trưng bằng phương pháp PCA trong công cụ Weka

Các thông số có thể cấu hình cho giải thuật PCS của Weka

Tên class: weka.filters.unsupervised.attribute.PrincipalComponents.
Chương trình thực hiện phân tích thành phần chính và chuyển đổi dữ liệu. Suy giảm số chiều được thực hiện bằng cách chọn đủ các hàm riêng để chiếm một số phần trăm phương sai trong dữ liệu gốc - mặc định giá trị này là 0.95 (95Chương trình dựa vào code lựa chọn thuộc tính 'PrincipalComponents' của hai tác giả Mark Hall và Gabi Schmidberger.

1. centerData – Tìm điểm trung tâm của dữ liệu (thay vì chuẩn hóa). PCA sẽ được tính toán từ ma trận hiệp phương sai (chứ không phải từ khoảng cách tương quan)

2. debug – Nếu được đặt thành true, bộ lọc có thể xuất thông tin bổ sung cho console

3. varianceConvered – Giữ lại đủ các thuộc tính PC để chiếm tỉ lệ phương sai này

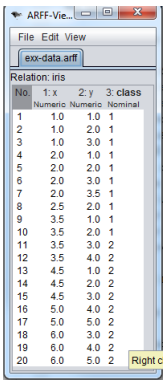
4. maximumAttribution – Số lượng thuộc tính PC tối đa được giữ lại

5. maximumAttributeNames – Số lượng thuộc tính tối đa bao giồn trong tên thuộc tính được chuyển đổi

6. doNotCheckCapabilities – Nếu được đặt giá trị, khả năng của bộ lọc sẽ không được kiểm tra trước khi nó được xây dựng. (Sử dụng thận trọng để giảm thời gian chạy)

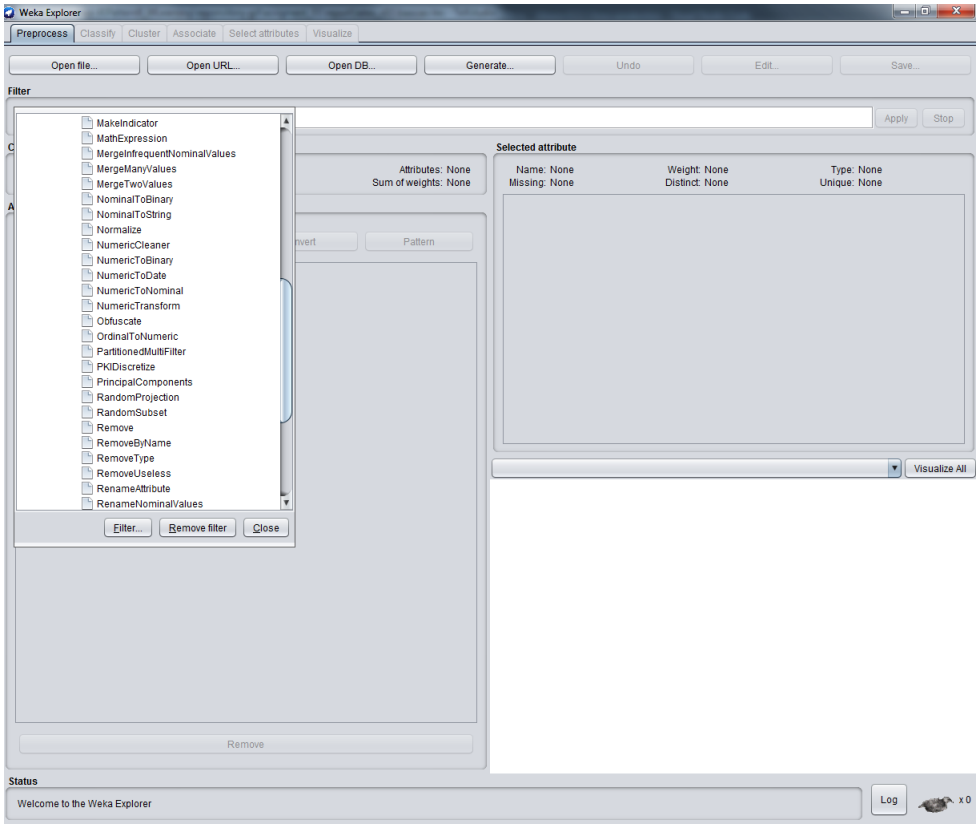
Dữ liệu đầu vào cho chương trình WEKA

```
% 1. Title: PCA Database
%
% 2. Sources:
%      (a) Creator: nqlong
%      (c) Date: Jan, 2019
%
@RELATION iris

@ATTRIBUTE x NUMERIC
@ATTRIBUTE y NUMERIC
@ATTRIBUTE class      {1,2}

@data
1,1,1
1,2,1
1,3,1
2,1,1
2,2,1
2,3,1
2,3.5,1
2.5,2,1
3.5,1,1
3.5,2,1
3.5,3,2
3.5,4,2
4.5,1,2
4.5,2,2
4.5,3,2
5,4,2
5,5,2
6,3,2
6,4,2
6,5,2
```



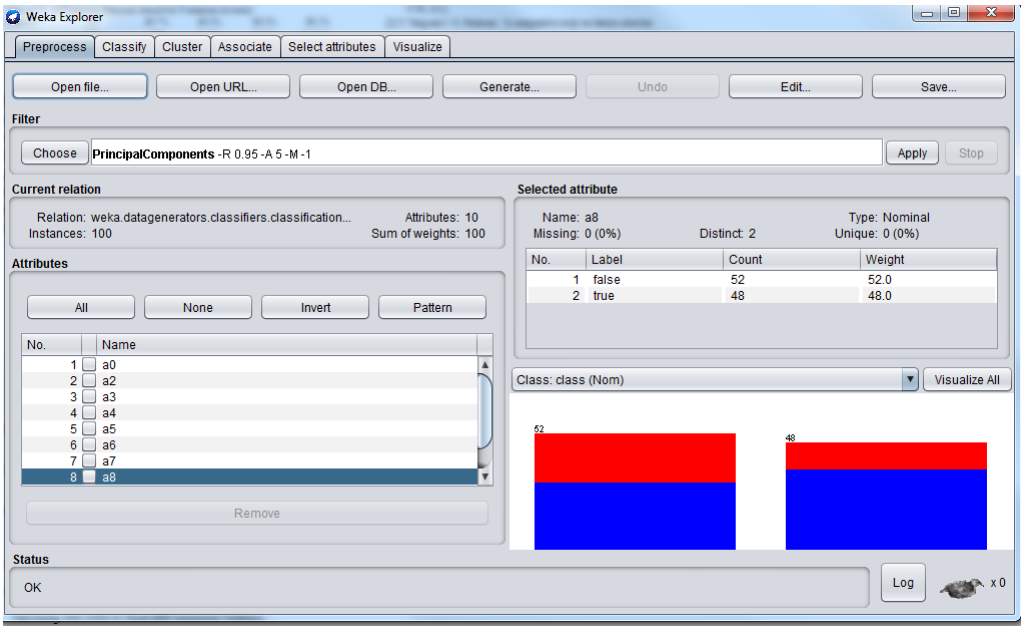
Hình 3: Dữ liệu đầu vào cho giải thuật PCA của Weka

Tiếp theo, ta thiết lập cấu hình cho phần mềm WEKA

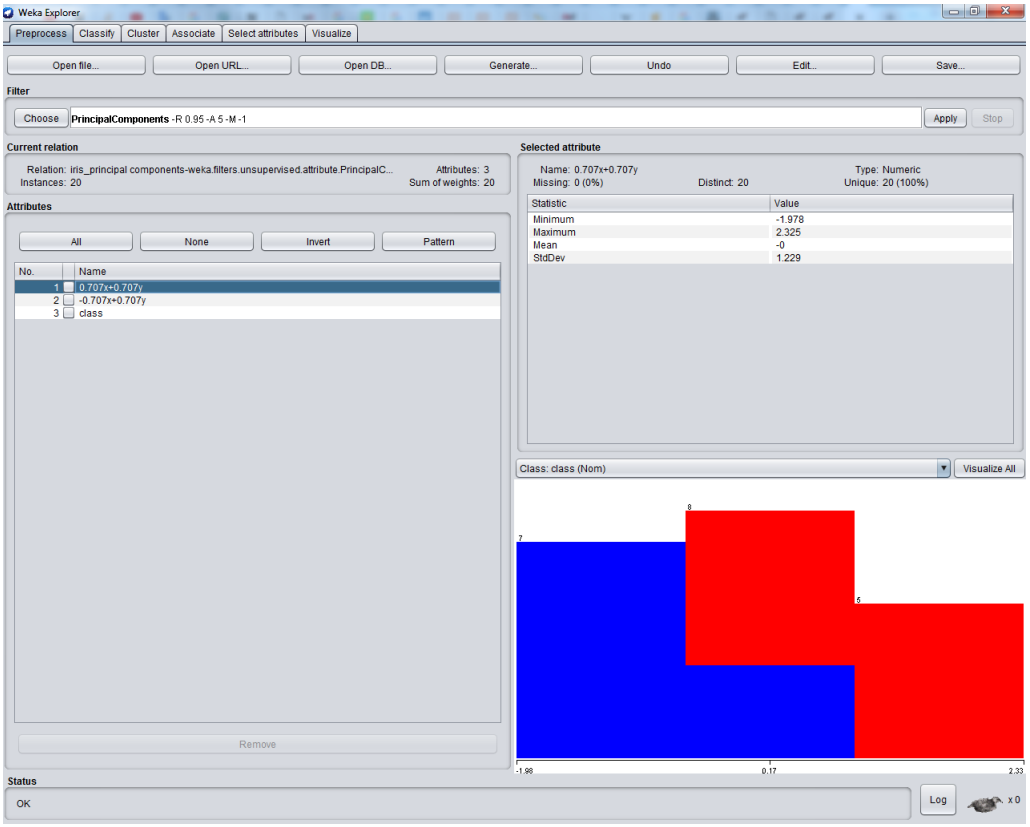


Hình 4: Lựa chọn giải thuật PCA

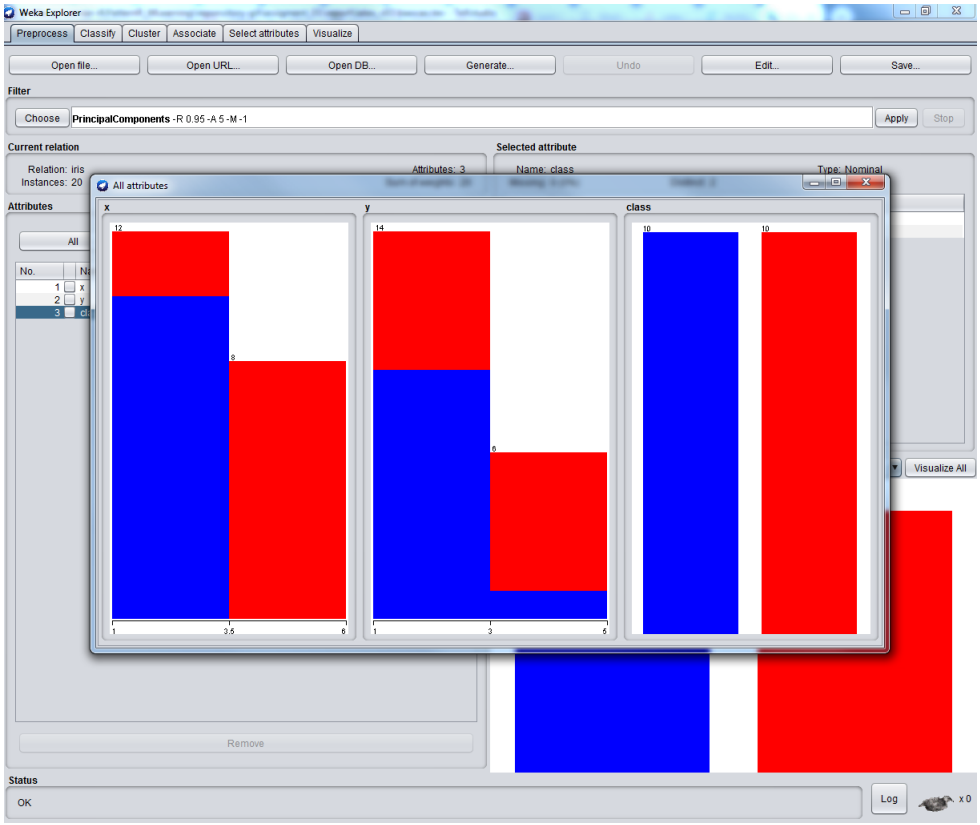
Tiếp theo sau đó, ta tải dữ liệu vào phần mềm Weka và áp dụng giải thuật PCA của Weka để tìm ra kết quả



Hình 5: Kết quả load dữ liệu mẫu



Hình 6: Kết quả của việc áp dụng giải thuật PCA (1)



Hình 7: Kết quả của việc áp dụng giải thuật PCA (2)

5 Kết Luận

Tài liệu được thực hiện với sự giúp đỡ tận tình của giáo viên hướng dẫn, Thầy TS. Dương Tuấn Anh
Trong quá trình nghiên cứu, thực hiện không thể tránh khỏi những thiếu sót, kính mong quý thầy cô và các bạn đóng góp thêm để tài liệu thêm hoàn thiện.
Xin chân thành cảm ơn.

6 Tài liệu Tham Khảo

http://en.wikipedia.org
https://machinelearningcoban.com/2017/06/15/pca/