

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

1	Giới thiệu	3
2	Giới thiệu công cụ Weka	4
3	Giới thiệu phương pháp PCA	4
4	Các bước rút trích dữ liệu đặc trưng bằng phương pháp PCA trong công cụ Weka	4
5	Kết Luận	9
6	Tài liệu Tham Khảo	9

Bài luận tìm hiểu giải thuật Min-conflicts

Nguyễn Quốc Long¹

MSSV: 1770023

Tóm tắt nội dung. Tài liệu tìm hiểu về giải thuật PCA và phần mềm 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 bài 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

Change of basis - Is there another basis which is a liner combination of the original basis, that best re-express our data set

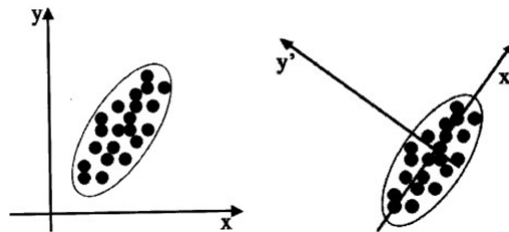


Figure 2.7: Two different sets of coordinate axes. The second consists of a rotation and translation of the first and was found

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 NAME `weka.filters.unsupervised.attribute.PrincipalComponents`

SYNOPSIS Performs a principal components analysis and transformation of the data. Dimensionality reduction is accomplished by choosing enough eigenvectors to account for some percentage of the variance in the original data – default 0.95 (95Based on code of the attribute selection scheme 'PrincipalComponents' by Mark Hall and Gabi Schmidberger.

OPTIONS `centerData` – Center (rather than standardize) the data. PCA will be computed from the covariance (rather than correlation) matrix

`debug` – If set to true, filter may output additional info to the console.

`varianceCovered` – Retain enough PC attributes to account for this proportion of variance.

`maximumAttributes` – The maximum number of PC attributes to retain.

`maximumAttributeName` – The maximum number of attributes to include in transformed attribute names.

doNotCheckCapabilities – If set, the filter’s capabilities are not checked before it is built. (Use with caution to reduce runtime.)

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}

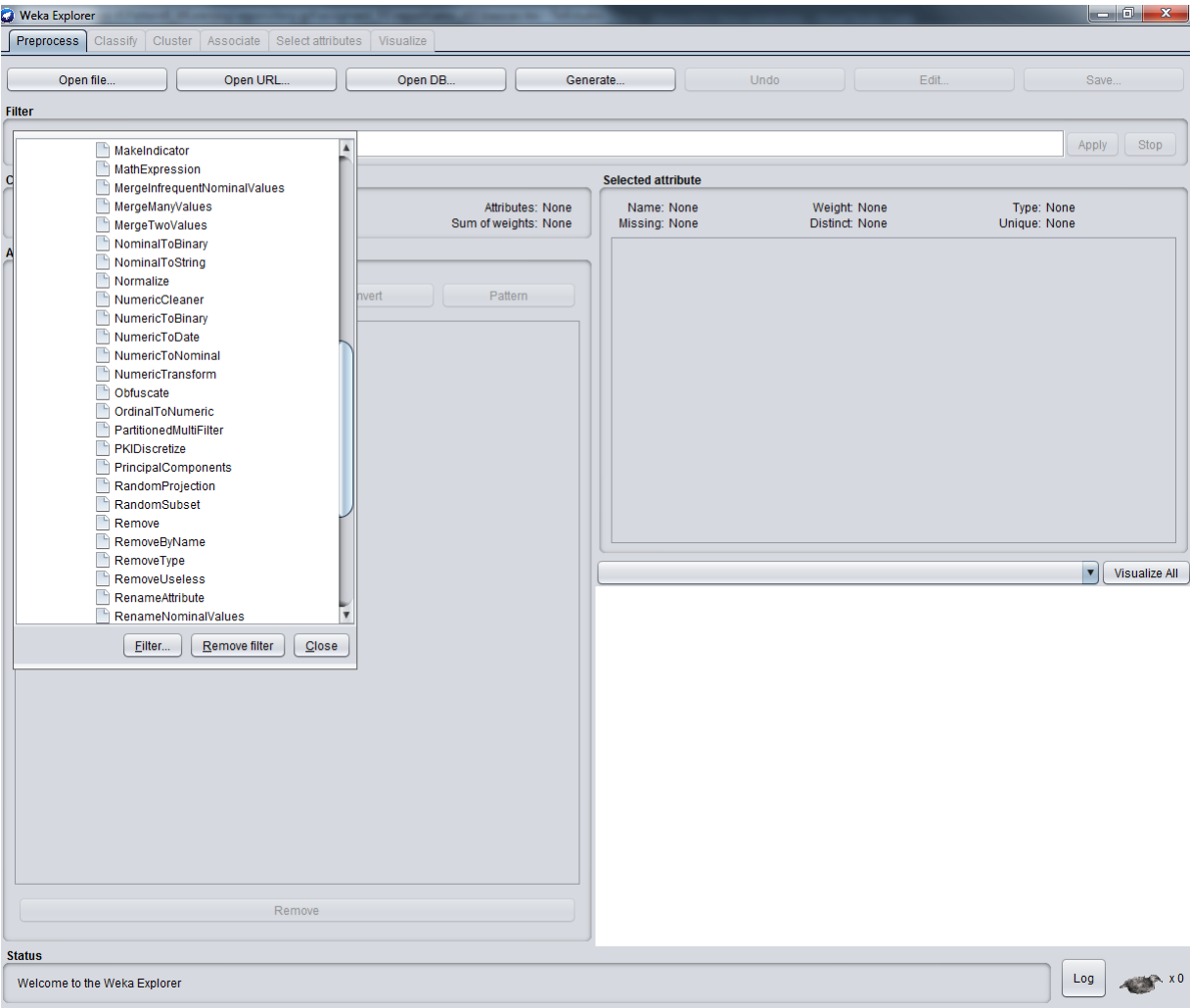
@data
1,1,1
1,2,1
1,3,1
2,1,1
2,2,1
2,3,1
2,3.5,1
2.5,2,1
3.5,1,1
3.5,2,1
3.5,3,2
3.5,4,2
4.5,1,2
4.5,2,2
4.5,3,2
5,4,2
```

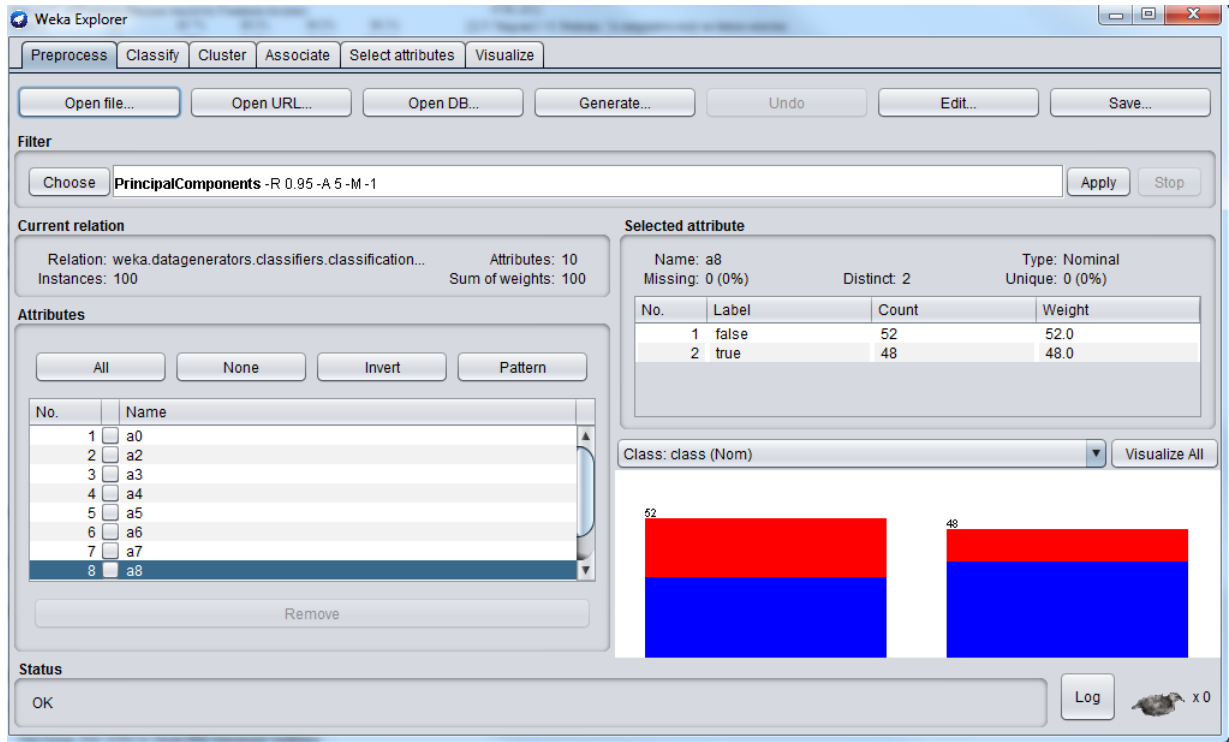
```
5,5,2
6,3,2
6,4,2
6,5,2
```

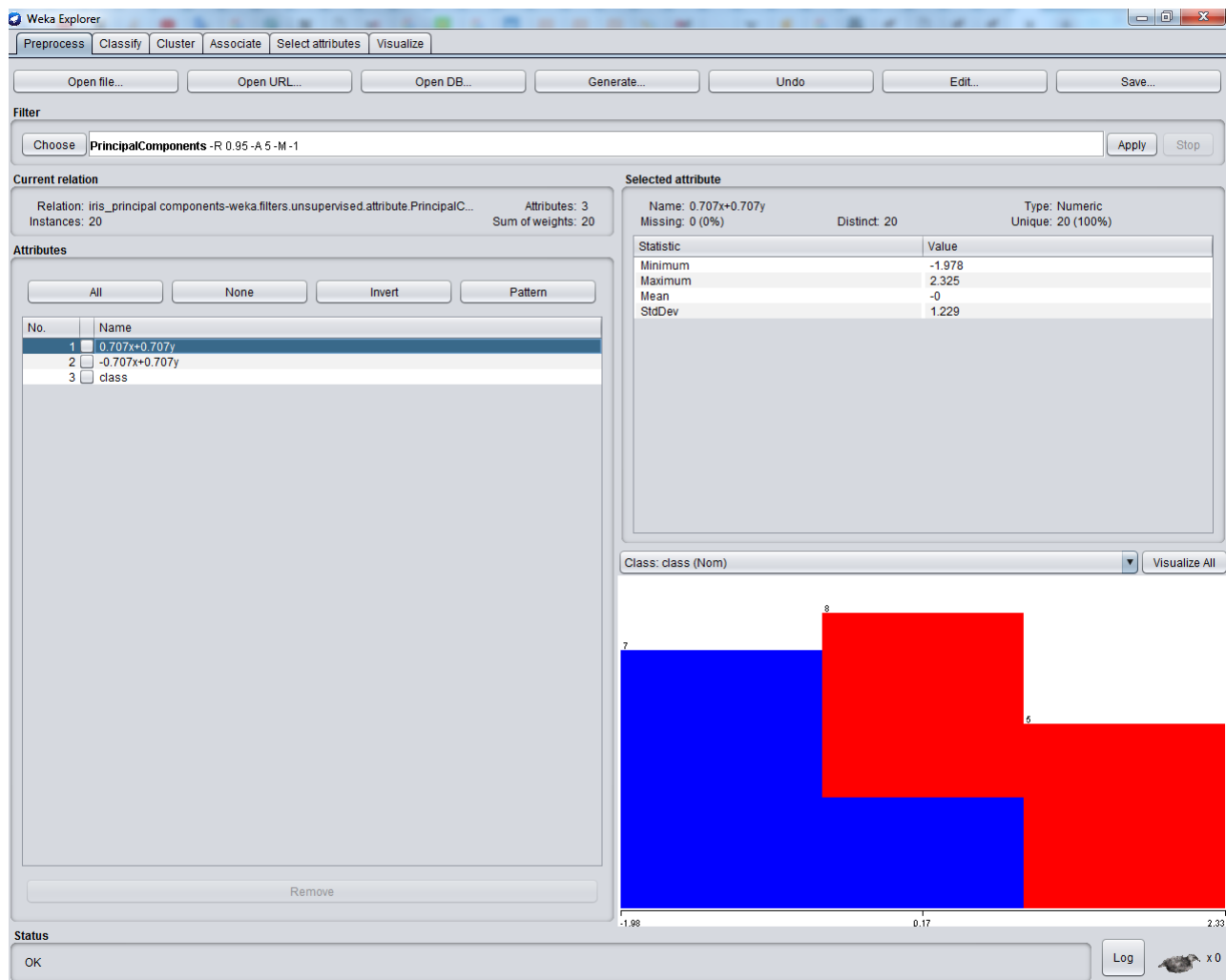
No.	1: x	2: y	3: class
	Numeric	Numeric	Nominal
1	1.0	1.0	1
2	1.0	2.0	1
3	1.0	3.0	1
4	2.0	1.0	1
5	2.0	2.0	1
6	2.0	3.0	1
7	2.0	3.5	1
8	2.5	2.0	1
9	3.5	1.0	1
10	3.5	2.0	1
11	3.5	3.0	2
12	3.5	4.0	2
13	4.5	1.0	2
14	4.5	2.0	2
15	4.5	3.0	2
16	5.0	4.0	2
17	5.0	5.0	2
18	6.0	3.0	2
19	6.0	4.0	2
20	6.0	5.0	2

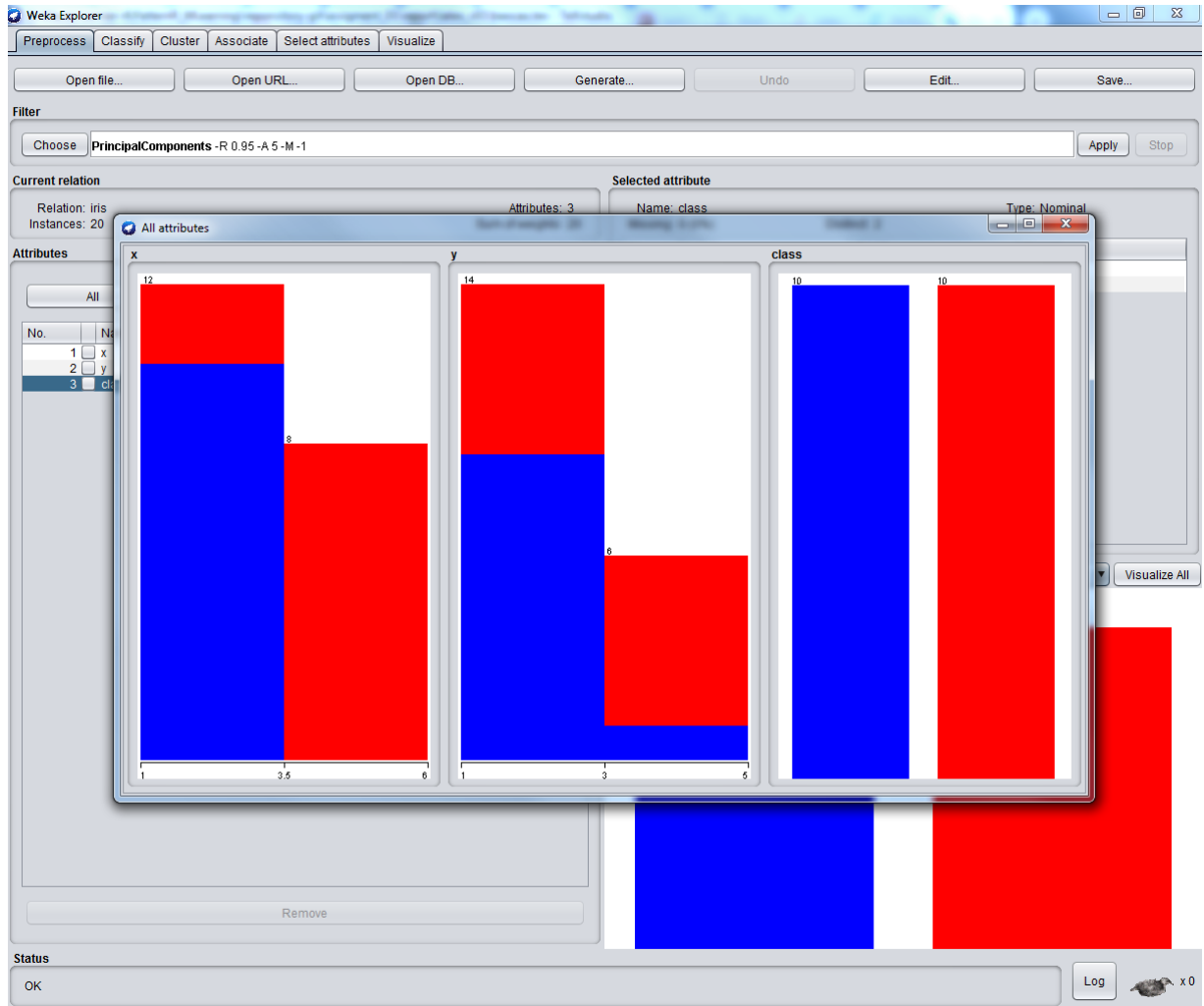
$$\begin{bmatrix} x_{11} & x_{12} & x_{13} & \dots & x_{1n} \\ x_{21} & x_{22} & x_{23} & \dots & x_{2n} \\ \dots & \dots & \dots & \dots & \dots \\ x_{d1} & x_{d2} & x_{d3} & \dots & x_{dn} \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & x_{13} & \dots & x_{1n} \\ x_{21} & x_{22} & x_{23} & \dots & x_{2n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{d1} & x_{d2} & x_{d3} & \dots & x_{dn} \end{bmatrix}$$

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









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/>