CBIE 2018
CONGRESSO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO

Using Principal Component Analysis to support students' performance prediction and data analysis

Vinícius R. P. Borges, Stéfany L. Esteves, Patrícia De Nardi Araújo, Lucas C. Oliveira, **Maristela T. Holanda**

viniciusrpb@unb.br,lealesteves@sistemas.ufla.br,patynardi@sistemas.ufla.br,lucacharles@sistemas.ufla.br,mholanda@unb.br

Fortaleza - CE, 30 de outubro de 2018

UNIVERSIDADE FEDERAL DE LAVRAS

Outline

- Introduction/Motivation
- Proposed method
- Experimental results
- Conclusion

Introduction/Motivation

Introduction

- Educational Data Mining (EDM) has emerged with techniques and strategies to process, interpret and obtain useful and implicit knowledge on educational data

- Several EDM tasks have been explored by the EDM community [1]:
  - Students' performances prediction
  - Students' drop-out rates
  - Learning achievements

[1] BAKER, Ryan Shaun; INVENTADO, Paul Salvador. Educational data mining and learning analytics. In: Learning analytics. Springer, New York, NY, 2014. p. 61-75.

Problem description

- Educational datasets can present several attributes, denoting high dimensionality
- Generally, state-of-art EDM methods deal with high dimensional data by:
  - Removing manually data attributes
  - Automatically selecting the relevant attributes [2]
- Another possibility refers to the dimensionality reduction by attribute transformation
  - It is not well-explored in EDM tasks

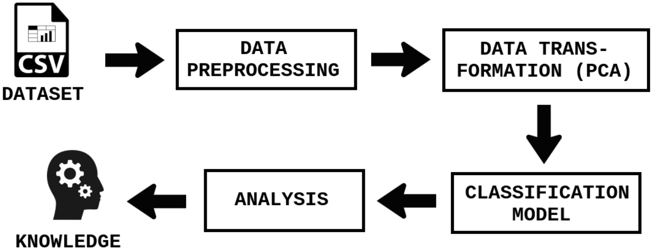
[2] BARADWAJ, Brijesh Kumar; PAL, Saurabh. Mining educational data to analyze students' performance. arXiv preprint arXiv:1201.3417, 2012.

Problem description

- A well-known technique for data transformation is Principal component analysis (PCA)
- PCA is a technique that can be simultaneously useful for:
  - Dimensionality reduction
  - Data analysis
- Succesfully applied in other knowledge domains

Proposed method

Proposed method



Dataset

- Data describes student achievements in secondary education of two Portuguese schools [3]
- Studens are described according to scholar, financial, social and personal attributes (33)
  - The first dataset (Dataset I) contains 649 students of the Portuguese subject
  - The second dataset (Dataset II) refers to final achievements of 394 students in the Math subject;

[3] CORTEZ, Paulo; SILVA, Alice Maria Gonçalves. Using data mining to predict secondary school student performance. 2008.

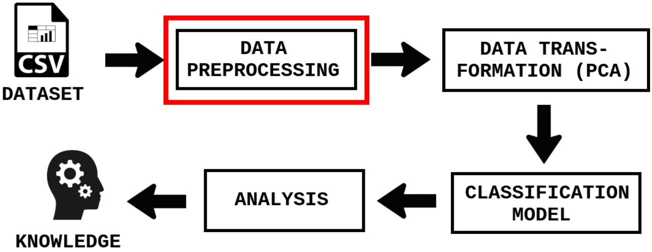
Dataset

- Scholar attributes
  - Midterms exams grades (G1 and G2)
  - Final exam grade (G3)
  - Past failures
  - Absences
- Personal attributes
  - Daily and weekly alcohol consumption
  - free time
  - romantic relationship
  - wants to take higher education (higher)

Dataset

- Familiar attributes
  - Mother and father education
  - Mother and father jobs
  - Student's guardian
- Other attributes
  - Travel time to school (in hours)
  - Address
  - Age, genre...

Proposed method



Preprocessing

- Transform categorical attributes to dummy variables [4]
- Obtain the categorical values $\{v_1, ..., v_k\}$ of an attribute A_i

Attribute i
A
B
B
C
A

[4] LEBART, Ludovic. Correspondence analysis. In: Data Science, Classification, and Related Methods: Proceedings of the Fifth Conference of the International Federation of Classification Societies (IFCS-96), Kobe, Japan, 1996. Springer Science & Business Media, 2013. p. 423.

Preprocessing

- Transform categorical attributes to dummy variables
- Create a new attribute $A_i = v_j$ for each categorical value v_j of A_i

Attribute i	**Attribute i =** A
A	1
B	0
B	0
C	0
A	1

Preprocessing

- Transform categorical attributes to dummy variables

Attribute i	Attribute i = *A*
A	1
B	0
B	0
C	0
A	1

- Transform categorical attributes to dummy variables

Attribute i	**Attribute i =** *A*	**Attribute i =** *B*
A	1	0
B	0	1
B	0	1
C	0	0
A	1	0

Preprocessing

- Transform categorical attributes to dummy variables

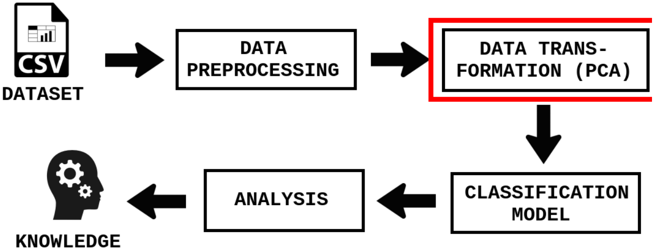
Attribute i	**Attribute i =** *A*	**Attribute i =** *B*
A	1	0
B	0	1
B	0	1
C	0	0
A	1	0

Preprocessing

- Final preprocessing result

Attribute i = A	**Attribute i =** B	**Attribute i =** C
1	0	0
0	1	0
0	1	0
0	0	1
1	0	0

Proposed method

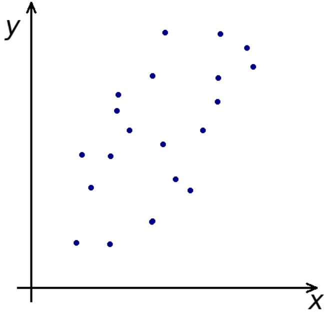


Data transformation

- Principal component analysis (PCA) [5]
- PCA performs a linear mapping of the data in a high dimensional space to a lower-dimensional space
  - so that data's variance is maximized

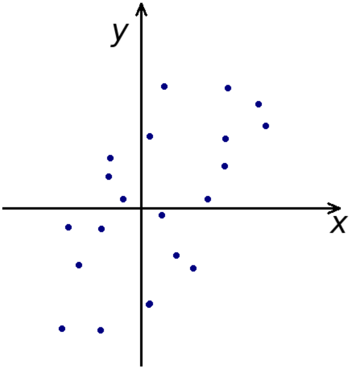
[5] JOLLIFFE, Ian. Principal component analysis. In: International encyclopedia of statistical science. Springer, Berlin, Heidelberg, 2011. p. 1094-1096.

- Consider the following dataset $\mathbf{X} = \{\mathbf{x}_1, ..., \mathbf{x}_N\}$, in which $\mathbf{x}_i = \{x_i, y_i\}$:



- 1. Center the data instances $\mathbf{x}_i$ in relation to the mean μ:

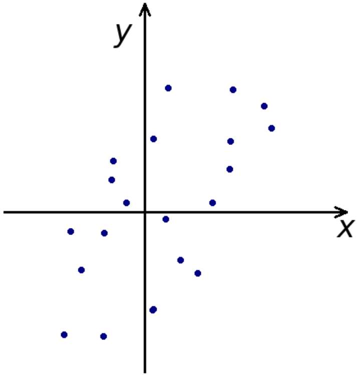
$$z_i = \mathbf{x}_i - \mu \tag{1}$$



Principal component analysis

- 2. Compute the covariance matrix Σ for centered data $Z = \{z_1, ..., z_N\}$:

$$\Sigma = Z^T Z \tag{2}$$



$$\Sigma = \begin{bmatrix} \sigma_x^2 & \sigma_x \sigma_y \\ \sigma_x \sigma_y & \sigma_y^2 \end{bmatrix}$$

- 3. From the decomposition of Compute the covariance matrix Σ such as

$$\Sigma = \mathbf{V} A \mathbf{V}^{-1} \tag{3}$$

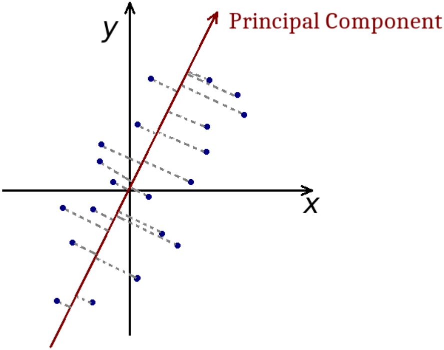
- obtain:
  - the eigenvalues $\lambda = \{\lambda_1, ..., \lambda_D\}$
  - the eigenvectors $\mathbf{V} = \{\mathbf{v_1}, ..., \mathbf{v_D}\}$
- The eigenvalues on the diagonal of A correspond the columns in $\mathbf{V}$

Principal component analysis

- Sort the eigenvalues in descending order;
- Select the k eigenvectors associated to the k largest eigenvalues
  - k is the number of dimensions of low-dimensional space (reduced space)
- Each eigenvector is associated to a principal component (PC)

Principal component analysis

- Generation of a principal component (the eigenvector associated to the higher eigenvalue):

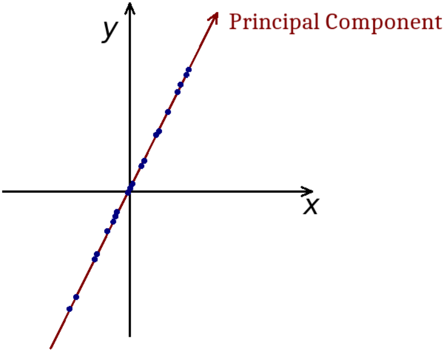


Principal component analysis

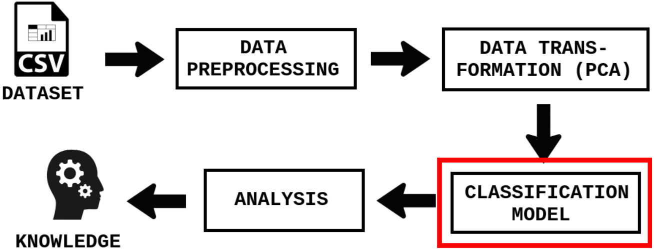
- Transforming original data $\mathbf{x}_i = \{x_{i,1}, ..., x_{i,D}\}$ to the principal component values

$$PC_l = c_{l,1}x_1 + c_{1,2}x_2 + ... + c_{1,D}x_D \tag{4}$$

in which $c_{l,j}$ is a coefficient of PC_l



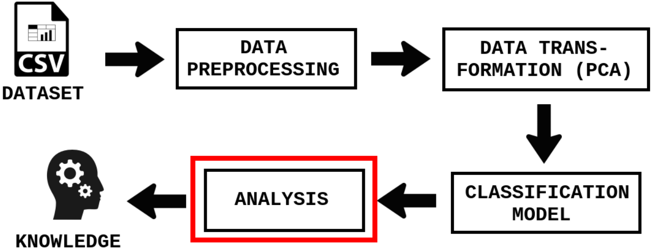
Proposed method



Classification

- The prediction of students' performances is deal as a classification problem
  - predicts if a student is approved or fail at the end of the scholar year
- Support vector machines (SVM)
  - SVM was set using a Radial Basis Function (RBF)
  - RBF width set as 2.0
- Naive Bayes
  - Probabilistic classification model on the Bayes theorem

Proposed method



Experiments and results

Experiments and results

- Experiments were conducted uing the Weka 3.9.1 environment
  - Holdout cross-validation
  - 66% of data instances are used for training, while 34% are used for test
- F-Score is the evaluation measure:

$$F_1 = \frac{2 \times TP}{2 \times TP + FP + FN}. \tag{5}$$

- F1-Scores obtained by the SVM classifier

Dataset	High	2 PCs	5 PCs	10 PCs
Dataset I (Portuguese)	0.776	**0.893**	0.773	0.776
Dataset II (Math)	0.511	**0.790**	0.511	0.511

Experiments and results

- F1-Scores obtained by the Naive-Bayes classifier

Dataset	High	2 PC's	5 PC's	10 PC's
Dataset I (Portuguese)	0.930	**0.992**	0.883	0.895
Dataset II (Math)	0.849	**0.917**	0.909	0.915

- Coefficients of the two principal components for the Portuguese subject
- PC1: positive correlation and negative correlation

PC1	PC2
Midterm exam 1 (-0.3)	Daily alcohol cons.="1" (0.357)
Midterm exam 2 (-0.294)	Weekly alcohol cons.="1" (0.279)
Mother educ. = "4" (-0.259)	sex="M" (-0.265)
wants higher educ.="no" (0.222)	Weekly alcohol cons.="5" (-0.219)
Mother educ. = "1" (0.221)	Weekly alcohol cons.="4" (-0.216)

Experiments and results

- Top-5 coefficients of the two principal components for the Portuguese subject
- PC2: positive correlation and negative correlation

PC1	PC2
Midterm exam 1 (-0.3)	Daily alcohol cons. =“1” (0.357)
Midterm exam 2 (-0.294)	Weekly alcohol cons.=“1” (0.279)
Mother educ. = “4” (-0.259)	sex=“M” (-0.265)
wants higher educ. =“no” (0.222)	Weekly alcohol cons.=“5” (-0.219)
Mother educ. = “1” (0.221)	Weekly alcohol cons.=“4” (-0.216)

Discussion (Educational perspective)

- The midterm exams and the higher educational degree of mother:
  - Vary together in PC1, so higher PC1 values are associated to lower values of such attributes
  - Mother education is related to the students' performances on midterms

- Mother education also influences students when deciding to take higher education

- Frequent alcohol consumption is more related to male students
  - Such attributes vary together in PC2
  - Higher values for PC2 are associated to lower alcohol consumption

- Top-5 coefficients of the two principal components for the Math subject
- PC1: positive correlation and negative correlation

PC1	PC2
absences (-0.998)	Midterm exam 2 (-0.752)
age (-0.029)	Midterm exam 1 (-0.649)
Midterm exam 2 (0.024)	failures (0.058)
Weekly alcohol cons. (-0.023)	go out (0.04)
Midterm exam 1 (0.021)	absences (-0.034)

Experiments and results

- Top-5 coefficients of the two principal components for the Math subject
- PC2: positive correlation and negative correlation

PC1	PC2
absences (-0.998)	Midterm exam 2 (-0.752)
age (-0.029)	Midterm exam 1 (-0.649)
Midterm exam 2 (0.024)	failures (0.058)
Weekly alcohol cons. (-0.023)	goout (0.04)
Midterm exam 1 (0.021)	absences (-0.034)

Discussion (Educational perspective)

- The first component (PC1) is strongly affected by the students' absences
  - Higher values in PC1 denotes lower absences values
- The student's absences, age and weekly alcohol consumption vary together
- The midterms exams are strongly correlated to the second principal component
  - Higher values in PC2 are related to lower values of midterm exams

Conclusion

Conclusion

- Method based on PCA for students' performance prediction tasks:
  - Retained or improved the prediction F-Scores when compared to the high dimensional spaces
  - PCA results provided information for data analysis
- Limitations
  - Choose the number of principal components (k) to consider in the low dimensional space
- Future works
  - Use Brazilian educational datasets
  - Consider other classification models (Neural networks, decision trees) and visualization techniques

Using Principal Component Analysis to support students' performance prediction and data analysis

Vinícius R. P. Borges, Stéfany L. Esteves, Patrícia De Nardi Araújo, Lucas C. Oliveira, **Maristela T. Holanda**

viniciusrpb@unb.br, lealesteves@sistemas.ufla.br, patynardi@sistemas.ufla.br, lucacharles@sistemas.ufla.br, mholanda@unb.br

Fortaleza - CE, 30 de outubro de 2018