Data Mining Techniques and SAS as a tool for graphical presentation of Principal Components Analysis and Disjoint Cluster Analysis results

Emir Slanjankic*, Haris Balta*, Adil Joldic*, Alsa Cvitkovic*, Djenan Heric*, Emir Veledar**

*Faculty of Information Technologies,
University "Dzemal Bijedic"
Mostar, Bosnia and Herzegovina
{emir.s, haris.balta, adil, alsa.c, djenan.h}@fit.ba

**School of Medicine,
Emory University
Atlanta, USA
eveleda@emory.edu

***Abstract*—Complexity of data analysis in data mining often makes results difficult to interpret. This problem could be solved using various approaches. Principal Component Analysis (PCA) and Disjoint Cluster Analysis (DCA) are methods used for data reduction and summarization. In this paper, PCA and DCA were applied on dataset example containing information about students' courses and time necessary to pass related exams. The SAS software was used as a data mining tool for performing this analysis. Another approach for better interpretation is visualization of results. This means showing important attributes visually to aid informal users to interpret results.**

Keywords – principal component analysis, disjoint cluster analysis, data mining, data visualization, SAS.

I. INTRODUCTION

Data mining, sometimes referred to as knowledge discovery, is at the intersection of multiple research areas, including Machine Learning, Statistics, Pattern Recognition, Databases, and Visualization [1].

Data mining methods are used for analyzing complex data. The analysis may be complex because of data amount, relations between attributes and patterns, and aim of analysis. Different approaches, such as PCA and DCA may reduce this complexity. Visualization is another approach that can help user in better understanding and interpretation of results.

Principal Component Analysis (PCA) is a method used for composing a number of correlated variables into new variable called principal components, as in (1). The purpose of PCA is to reduce dimensionality by retaining only those characteristics of the dataset that contributes most of its variance. This method provides a very effective way to reduce dimensionality by performing a covariance analysis between components [2, 3]. The number of weak-associated components depends on several criteria (i.e. Kaiser-Guttmann rule, percentage of variance, the scree test). A linear combination of eigenvectors and original variables produces single principal component. A vector which produces a scalar multiple of the original vector when acting in a linear transformation is called eigenvector. The scalar is called the eigenvalue corresponding to this eigenvector. For example, principal component could be expressed as:

$$\xi_1 = C_1\alpha_1 + C_2\alpha_2 + \cdots + C_k\alpha_k + \cdots + C_n\alpha_n \qquad (1)$$

where α_i are the original variables, and coefficients are their eigenvectors.

On the other hand, Disjoint Cluster Analysis (DCA) groups observations into clusters suggested by the data, so that members of the same cluster have strong correlation, and members from different cluster have weak correlation [8]. When assigning observations to clusters, an arbitrary number of clusters is created. The observations are then being reassigned to other clusters, based on their similarity to other observations inside the cluster. Some techniques, such as Cubic Clustering Criterion (CCC), pseudo F statistics (PSF) and pseudo T^2 statistics (PST2) are used to determine optimal number of clusters [4].

When dealing with large amount of data, visualization of data analysis could be of great use for informal users. In most cases, information is better interpreted and more understandable when using graphical concepts and elements. This approach gives the ability to visually present lots of relationships and information in comprehensive way.

The above mentioned concepts and approaches will be used on sample dataset, which consists of data about first year students, courses and time necessary to prepare and pass the exams. This dataset was taken from database of Faculty of Information technologies (FIT), Mostar. The assumption is that more time needed for passing exam, the course is more difficult. Results of this analysis could be used for courses' arrangement over semesters, according to course complexity as mentioned above.

II. Data Preparation

According to the analyses, data preparation has the most influence. Researchers such as Pyle also defined it as the most important part of a data exploration process which leads to success [5]. Some authors claim that, in most cases, data preparation takes up to 80% of the time needed for the whole data mining process. Additionally other aspects of the data preparation directly depend on attribute selection (responsible for removing irrelevant and redundant information based on rough set theory), rule induction (calculation of decision rules) and visualization/testing aspect (checking the collected knowledge and presenting it in a form easily understandable by humans) [6, 7].

The analysis has been conducted on data from FIT database and exported to MS Excel table. The variables of initial and final dataset are shown in Table I.

TABLE I. Initial And Final Dataset

Initial Dataset	Final Dataset
StudentID (id)	StudentID (id)
Status (status)	Math (MM)
Sex (spol)	IntrodToEconomy (OEK)
StudyYear (godina_studija)	IntroToInformTechnol (uit)
RepeatsYear (obnavljac)	IntroToProgramm (upr)
Generation (generacija)	AlgorithDataStructure (asp)
EnrollmentDate (dat_upis)	ComputSystemArchit (aks)
LivesAt (gdje_stanuje)	English (ej)
CourseID (predmet)	Programming1 (pr1)
Mark (ocjena)	IntroToOperSystems (uos)
MarkDate (datum_polag)	

Data from initial dataset have been rearranged using MS Excel pivot table, in order to get necessary variables from many records. We calculated number of days necessary for course completion for each student as a distance between date of enrollment and date of exam passed. Initial dataset contained 1555 records, from which we extracted records for generations 2004-2007. During preparation, we excluded observations with missing values and kept only those records with all exams passed. The results in the final dataset contained 178 records. Then the final dataset has been exported in SAS software.

III. Methods

Using SAS® 9.1, we have created SAS dataset and applied SAS macros FACTOR and DISJCLUS on a given dataset. Macro is a set of procedures and statements which can be used repeatedly.

The FACTOR macro is a powerful SAS application for performing principal component and factor analysis on multivariate attributes. The SAS procedure, PROC FACTOR, is the main tool used in the macro as both PCA and EFA can be performed using PROC FACTOR [9, 10]. Other SAS procedures, such as CORR, GPLOT, BOXPLOT, and IML modules, are also incorporated in the FACTOR macro. The advantages of using the FACTOR macro are:

- The scatter plot matrix and simple descriptive statistics of all multivariate attributes and the significance of their correlations are reported.
- Test statistics and P values for testing multivariate skewness and kurtosis are reported.
- The quantile–quantile (Q–Q) plot for detecting deviation from multivariate normality and the outlier detection plot (PROC GPLOT) for detecting multivariate outliers are generated.
- Bi-plot displays (PROC GPLOT) showing the interrelationships between the principal components or factor scores and the correlations among the multiple attributes are produced for all combinations of selected principal components or factors.
- Options for saving the output tables and graphics in WORD, HTML, PDF, and TXT formats are available.

The DISJCLUS macro is another SAS application for exploring and visualizing multivariate data and for performing disjoint cluster analysis using the k-means algorithms. The SAS procedure FASTCLUS [11, 12] is the main tool used in the DISJCLUS macro. The FASTCLUS procedure is used to extract a user-specified number of clusters based on k-means cluster analysis. To verify the user-specified optimum cluster number, the cubic clustering criterion (CCC), pseudo F statistic (PSF), and pseudo T2 statistic (PST2) for a number of clusters ranging from 1 to 20 are generated using Ward's method of cluster analysis in PROC CLUSTER. To perform variable selection that significantly discriminates the clusters, the backward selection method in stepwise discriminant analysis using the STEPDISC procedure is used. Multiple canonical discriminant analysis based on the CANDISC is used to test the hypothesis that significant differences exist among the clusters.

SAS IML is also used to test the hypothesis of multivariate normality, which is a requirement for canonical discriminant analysis hypothesis testing.

PROC GPLOT is used to generate scatterplots by cluster, quantile–quantile plots for testing of multivariate normality, and diagnostic plots based on CCC, PSF, and PST2 for selecting the optimum cluster numbers and bi-plot display of canonical discriminant analysis. The BOXPLOT procedure is used to show the between-cluster differences for intra-cluster distances and canonical discriminant functions.

The advantages of using the DISJCLUS macro over the PROC FASTCLUS include:

- A scatterplot matrix of all multivariate attributes by cluster groups is displayed.
- Test statistics and P values for testing multivariate skewness and kurtosis after accounting for the variation among the cluster groups are reported.
- Quantile–quantile (Q–Q) plots for detecting deviation from multivariate normality and plots for detecting multivariate outliers after accounting for the variation among the cluster groups are produced.

- Graphical displays of CCC, PSF, and PST2 by cluster numbers ranging from 1 to 20 verify the user-specified number of clusters in the DCA as the optimum cluster solution.
- Significance of the variables used in DCA in discriminating the clusters is verified by performing a step-wise discriminant analysis.
- DISJCLUS macro offers options for performing a disjoint cluster analysis on standardized multi-attributes, as variables with large variances tend to have more influence on the resulting clusters than those with small variances; also, DCA based on principal components of highly correlated multi-attributes is also available in the DISJCLUS macro.
- The DISJCLUS macro offers options for detecting statistical significance among cluster groups by performing canonical discriminant analysis.
- The DISJCLUS macro offers options for displaying interrelationships between the canonical discriminant scores for cluster groups and correlations among the multi-attributes in bi-plot graphs.
- Options for saving the output tables and graphics in WORD, HTML, PDF, and TXT formats are available.

We applied macro DISJCLUS, initially with three clusters, exploratory graphs, and the courses as predictor variables (total of nine courses). In macro FACTOR two components were chosen, and PCA as a factor analysis method.

IV. Results

Applying FACTOR macros on sample dataset, we intend to find similarities between courses, according to complexity as number of days necessary to prepare and pass the exam. Optimal number of course components is expected as output from this macro. Using DISJCLUS macro should result in clusters of students according to the complexity as mentioned before.

For the given inputs in macro FACTOR, SAS suggested two components, as shown in Figure 1.

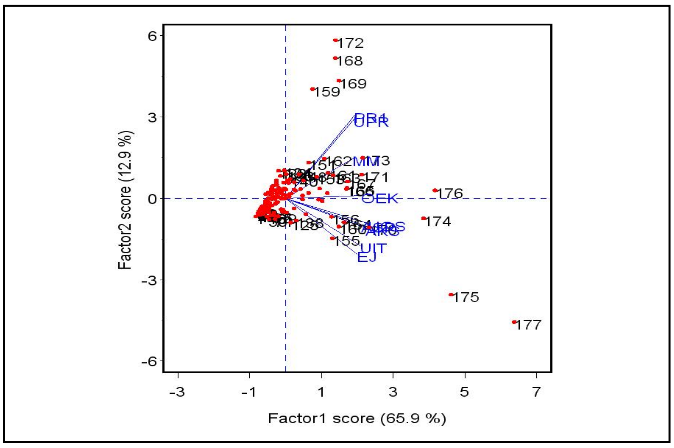

Figure 1. Components as a result of applying PCA (Factor 1, Factor 2)

The aim of PCA is to maintain maximum number of independent components which explain most of variance. According to this, these two components explain 78.8% of total dataset variance. All courses belong to both components, which is explained with eigenvectors shown in Table II.

TABLE II. Eigenvectors for Courses

	Eigenvectors	
	1	*2*
MM	0.30224	0.26873
OEK	0.34069	0.01778
UIT	0.33488	-0.31468
UPR	0.30631	0.52914
ASP	0.34202	-0.17863
AKS	0.36299	-0.19914
EJ	0.32140	-0.37268
PR1	0.30993	0.55607
UOS	0.37219	-0.16969

The eigenvectors represent projections of courses' vectors to the components, explaining participation of each course in single principal component.

Furthermore, there are similar courses, such as Introduction to Programming (UPR) and Programming 1 (PR1), and Architecture of Computer Systems (AKS), Introduction to Operating Systems (UOS) and Algorithms and Data Structures (ASP). Similar courses' vectors are close together, which is also shown in Figure 2.

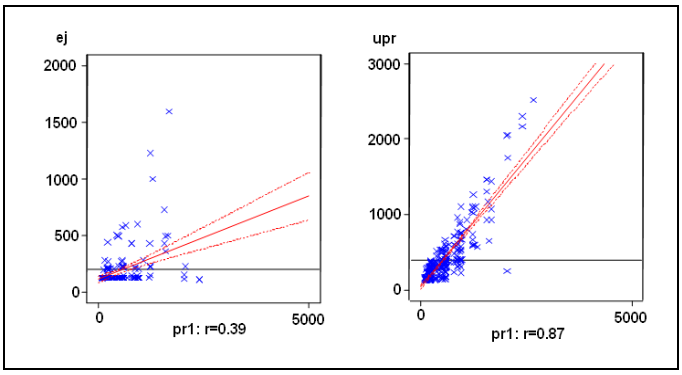

Figure 2. An example of uncorrelated (left) and correlated (right) course

In SAS, optimal number of components can be calculated using Kaiser-Guttmann's rule. The scree plot in Figure 3 helps in understanding this rule, because only components with eigenvalue greater than one are relevant for the analysis.

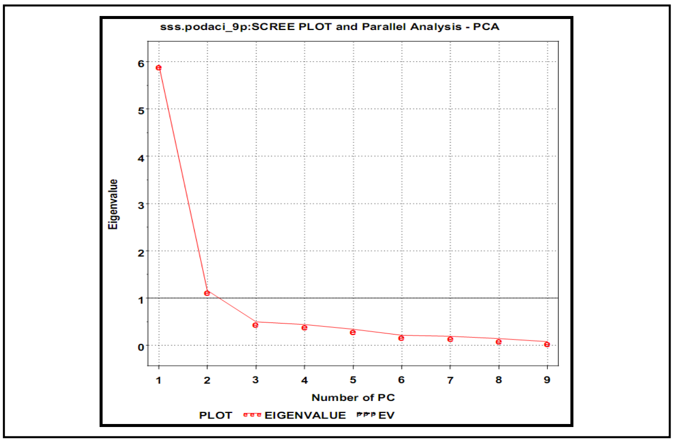

Figure 3. Scree plot; Kaiser-Guttmann rule choosing number of components

To determine optimal number of clusters, CCC, PSF and PST2 are used. The peaks of PSF and PST2 values suggest optimal number of clusters, as well as peaks of CCC values. Results of applying these techniques are presented in Figure 4.

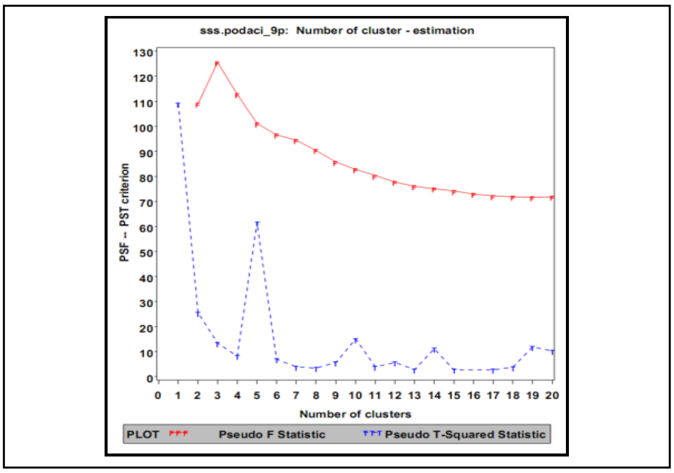

Figure 4. Determining number of clusters with PSF and PST2

Observing Figure 4, PSF suggests three clusters, while PST2 recommends either two or five number of clusters. According to CCC criterion, as presented in Figure 5, we should choose five or even six clusters.

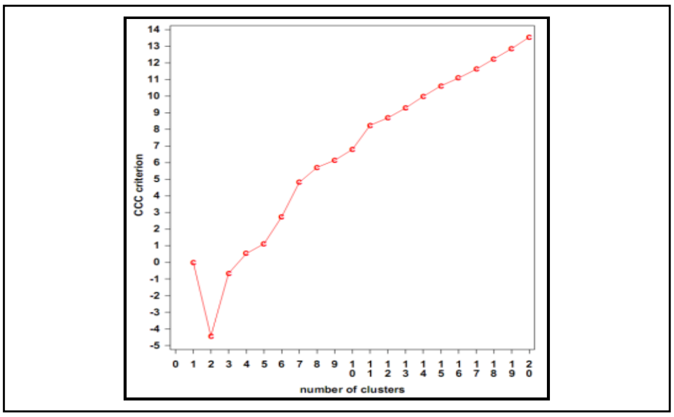

Figure 5. CCC criterion in determining number of clusters

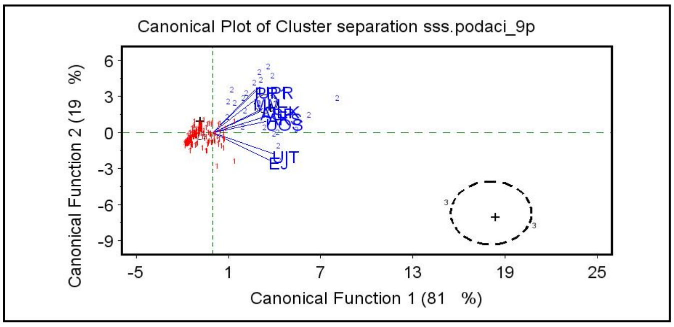

Figure 6. Cluster as a result of canonical discriminant analysis

Since PSF, PST2 and CCC are independent criteria, it is recommended to look for consensus among them. The difference between numbers of clusters might be a consequence of outliers' presence in dataset. In other words, there might be observations which significantly diverge from others, and such observations should sometimes be excluded from the dataset.

V. Discussion

The calculated components contain similar courses because of time spent for preparing and passing the exam. Those courses might cover similar topics; have the same number of hours; be taught by the same teacher; involve similar teaching methods etc. However, these aspects are not covered with this research.

Combining results of applied PSF, PST2, and CCC statistics, all observations are arranged into three clusters. These clusters correspond with groups of good, poor and extremely poor students, respectively. Small circles shown in Figure 6 represent clusters, with relatively small variance between members of clusters 1 and 2. Two members in cluster 3 are possibly outliers, because of relatively great difference between values of members from third cluster and members from the remaining clusters.

VI. Conclusion

Visualizing results helps to understand and interpret them. The graphical presentation is important especially for human users with relatively small pre-knowledge. Visualized results would ensure better view of students' efforts on preparing and passing exams, and results they achieve. This could be of significant importance for FIT, especially in planning strategy for next academic year.

This analysis shows that there are students with poor results in studying. These students should be provided with more attention and direct work to help them improve their results.

References

[1] R. Kohavi, “Data Mining and Visualization,” San Mateo, CA, USA, 2000.

[2] W. Muller, T. Nocke, H. Schumann, "Enhancing the visualization process with principal component analysis to support the exploration trends," ACM International Conference Proceeding Series; Vol. 243, 2004.

[3] A. Nahar, R. Daasch, S. Subramaniam, "Bum-in reduction using principal component analysis," Test Conference, Proceedings. ITC 2005. IEEE International Volume, Issue 8, 2005.

[4] SAS Institute Inc., “SAS Technical Report A-108: Cubic Clustering Criterion,” SAS online documentation, 1998.

[5] D. Pyle, “Data preparation for data mining,” San Francisco Morgan Kaufmann, 1999.

[6] Z. Pawlak, “Knowledge and Uncertainty: A Rough Set Approach,” SOFTEKS Workshop on Incompleteness and Uncertainty in Information Systems, 1993, p. 34-42.

[7] Z. Pawlak, J. Grzymala-Busse, R. Slowinski, W. Ziarko, “Rough sets,” Communications of the ACM, v.38 n.11, Nov. 1995, p.88-95.

[8] G. Fernandez, “Data Mining Using SAS® Applications,” Chapman & Hall/CRC, 2003.

[9] Comparison of the PRINCOMP and FACTOR. Available: http://support.sas.com/documentation/cdl_main/index.html

[10] The FACTOR Procedure. Available: http://support.sas.com/documentation/cdl_main/index.html

[11] Introduction to Clustering Procedures. Available: http://support.sas.com/documentation/cdl_main/index.html

[12] The FASTCLUS Procedure. Available: http://support.sas.com/documentation/cdl_main/index.html