Research Article

Agnieszka Wosiak*

Principal Component Analysis based on data characteristics for dimensionality reduction of ECG recordings in arrhythmia classification

<https://doi.org/10.1515/phys-2019-0050>

Received Jun 17, 2019; accepted Jul 02, 2019

Abstract: Due to the growing problem of heart diseases, the computer improvement of their diagnostics becomes of great importance. One of the most common heart diseases is cardiac arrhythmia. It is usually diagnosed by measuring the heart activity using electrocardiograph (ECG) and collecting the data as multidimensional medical datasets. However, their storage, analysis and knowledge extraction become highly complex issues. Feature reduction not only enables saving storage and computing resources, but it primarily makes the process of data interpretation more comprehensive. In the paper the new igPCA (in-group Principal Component Analysis) method for feature reduction is proposed. We assume that the set of attributes can be split into subgroups of similar characteristic and then subjected to principal component analysis. The presented method transforms the feature space into a lower dimension and gives the insight into intrinsic structure of data. The method has been verified by experiments done on a dataset of ECG recordings. The obtained effects have been evaluated regarding the number of kept features and classification accuracy of arrhythmia types. Experiment results showed the advantage of the presented method compared to base PCA approach.

Keywords: dimensionality reduction, data characteristics based feature extraction, decomposition, medical data analysis, group structure, ECG recordings

PACS: 87.85.-d, 02.50.Sk

1 Introduction

The progress in health technologies' development and the growing capabilities of diagnostic equipment cause the process of medical analysis and diagnosis highly challenging due to large and multidimensional datasets. Automated knowledge extraction from massive data and the medical inference based on data analysis pose highly complex issues. This is mainly due to the limitations imposed by the performance of computer systems, but also because of the methodological problems inherent in multidimensional data analysis. Therefore it is often called "the curse of dimensionality" [1–3].

Reduction of large datasets can be performed by reducing the number of analyzed parameters (dimensions) or by decreasing the number of analyzed cases. The dimensionality reduction can be carried out through statistical methods, primarily Principal Component Analysis (PCA) [4] or by using feature selection techniques [5, 6]. Dataset cardinality reduction can be achieved by sampling, grouping or instance selection methods [7].

In this research, we propose a modification to the application of PCA method called igPCA (in-group Principal Component Analysis). It introduces the preprocessing phase that arranges the related features into groups of similar distribution. We compare the performance of the considered algorithm in arrhythmia classification with accuracy results attained for an original set of features and for a dataset passed through unchanged PCA. We applied our method to reduce data derived from ECG signals to improve storage and inference process in solving arrhythmia classification problem. In the research we use a reference "ARRHYTHMIA" dataset, derived from the UCI repository [8]. However, the proposed method is to be applied for a real dataset of similar structure.

The remainder of this paper is organized as follows. Section 2 (Background) describes the problem of feature cardinality reduction in terms of data intrinsic characteristic. It also presents literature review introducing feature selection techniques based on data characteristics. Next

*Corresponding Author: Agnieszka Wosiak: Institute of Information Technology, Lodz University of Technology, ul. Wólczńska 215, 90-924 Lodz, Poland; Email: agnieszka.wosiak@p.lodz.pl

section 3 (Method overview) presents the proposed procedure for in-group principal component analysis. Section 4 describes the medical problem of arrhythmia and a dataset resulting from ECG recordings. In Section 5 (Experimental results and discussion) we describe the studies that were conducted. We introduce data characteristic and discuss the results. Finally, in Section 6 (Conclusions) we summarize our research and describe further works.

2 Background

2.1 Problem statement

The development of computer technologies and their wide use in medicine give a significant increase in medical data repositories. The process of extracting information from these huge datasets, which is essential for a given treatment, becomes more and more complex, and requires analysis for different types of data [9]. Even though collecting greater number of data may contribute to a comprehensive diagnosis, it requires more resources related to storage, processing and increased computing costs [10, 11]. For this reason, a wide range of methods for data complexity reduction are considered. They refer to two basic approaches: instance reduction and feature reduction. The difference between them is shown in Figure 1 [7].

Figure 1: Data reduction approaches

In literature, the problem of data multidimensionality is often referred to as "the curse of dimensionality" [1]. There is no doubt that a larger number of parameters describing cases may lead to more comprehensive analysis. However, the multidimensional nature of data may hinder, and sometimes even prevent, their proper interpretation, and classical approaches to the analysis of these data may not be sufficient [12]. Moreover, it is not very likely that all variables are independent, and that the collinearity may

lead to instability in the solution space and, as a consequence, inconsistent results. Hence, to eliminate the negative impact of multidimensionality on data analysis, key attributes should be identified and the dimensionality ought to be reduced in such a way as not to lose the knowledge that could be obtained from the data. Experts continue to play the leading role in the selection of parameters necessary to make a diagnosis, but more and more often the process of discovering medical knowledge is carried out using the IT techniques.

There is usually no need to reduce cases with regard to medical datasets. However, if data streams derived, for example, from ECG or EEG, are considered [13], correct (normal) values or the most common values can be disregarded. The problem of a large number of features describing medical cases (referred to as "*multidimensionality of data*"), is much more common. Multidimensional data are characterized by a very large number of parameters, far exceeding the number of cases.

Two main approaches for dimensionality reduction can be distinguished:

- feature extraction (feature transformation),
- feature selection.

The process of feature selection (FS) includes identification of relevant attributes from the dataset's characteristics and removal of the majority of irrelevant or redundant parameters [14]. The selected subset of features should be chosen so that it still reflects the relations and characteristics of the entire set of instances [15, 16]. The goal of feature selection is to build a subset of m features $F_S = x_{i_1}, x_{i_2}, \dots, x_{i_m}$ from the original set of n attributes $F = x_1, x_2, \dots, x_n$ with $m < n$ that optimizes an objective function $J(F)$. It can be expressed by (1).

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \rightarrow \begin{bmatrix} x_{i_1} \\ x_{i_2} \\ \vdots \\ x_{i_m} \end{bmatrix}, \quad (1)$$

$$x_{i_1}, x_{i_2}, x_{i_m} = \arg \max_{m, i_m} [J(x_1, x_2, \dots, x_n)]$$

The application of space transformation techniques introduces another possibility for dimensionality reduction. The result constitutes an entirely new set of features of a smaller cardinality by combining the original attributes. The goal of feature extraction is to find for a feature space $x_i \in R_n$ a mapping $y = g(x) : R_n \rightarrow R_m$ with $m < n$ such that the transformed feature vector $y_i \in R_m$ preserves (most of) the information or structure in R_n . It

can be expressed by (2).

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \rightarrow \begin{bmatrix} y_{i_1} \\ y_{i_2} \\ \vdots \\ y_{i_m} \end{bmatrix} \quad g = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \quad (2)$$

Numerous approaches of space transformation were proposed to meet different criteria and specific requirements of domain applications. Nonetheless, the most common approaches are based on the linear methods and include factor analysis and principal component analysis (PCA) [17–19].

Principal component analysis is a standard multivariate data analysis technique. It reduces the number of space dimensions, preserving dataset's variation and intrinsic dependencies [20]. It goes towards explaining the correlations between variables using a smaller set of linear combinations of these variables, which are referred to as the main components. The principal components analysis for dimensionality reduction results from the fact that the total variability of the data set consisting of m variables can often be kept for a smaller set of k other variables, which are constituted by linear combinations of primary variables, i.e. k main components carry almost as much information as the original m variables. Consequently, the procedure involves building the first few data components that hold the majority of a dataset's variation, instead of investigating thousands of original variables. Furthermore, principal component analysis may bring benefits also for datasets with an average or a low number of features [21].

Feature space reduction is usually performed without considering group structure. However, possessing the underlying group characteristic may bring additional benefits, making use of structural information about the features and discovering meaningful subsets of features. Moreover, there are some domain applications, where parameters tend to repeat distribution models, which makes incorporating group structure into the process of feature selection even more well-founded.

2.2 Related works

Feature selection methods have been an active field of the studies for decades. However, most of them address selecting features at the individual feature level, whereas considering group structure may contribute to the better performance of the subsequent analysis.

Nonetheless, there are many interesting investigations based on group feature selection and the Lasso approach [22]. Yuan and Lin in [23] proposed the group Lasso model that selects grouped variables and improves prediction in regression problems. In [24] Meier, van de Geer and Bühlmann extended the group Lasso to logistic regression models, which are suitable for high-dimensional data. Jacob, Obozinski and Vert in turn extended the group Lasso to sparsity patterns, which are unions of overlapping groups and help recover sparse connected patterns in a graph [25].

Li et al. in [26] perform parameter selection at the group and individual feature levels simultaneously with streaming features. Their GFSSF algorithm identifies relevant features from important groups and selects variables with sparsity at both the group and individual feature levels.

In [27] Murkute and Borkar presented group feature selection method at group level to execute feature selection. Their objective was to execute the feature selection within the group and between groups of features in order to select discriminative features and remove redundant features to obtain optimal subset. The method called EGVS (efficient group variable selection) comprises of two stages: within group variable selection that select discriminative features within the group (each feature is evaluated individually) and between group selection, when all the features are re-evaluated so as to remove redundancy.

All presented surveys confirmed efficiency of grouping features in terms of data reduction. However, the authors focused on feature selection, whereas there is still lack of a solution considering transformation techniques, which is a subject of the proposed method.

3 Method overview

The proposed in-group feature extraction method (igPCA) is based on principal component analysis incorporating diversity in distribution of various parameters. The steps can be presented as follows:

1. The process starts with data preparation, which aims at adjusting original datasets to analysis needs.
2. Feature grouping based on statistical analysis of distribution is carried out and groups of similar characteristics are distinguished:

- (a) The first group consists of binary features (B).

- (b) The rest of the features are divided into five groups based on skewness of the frequency distributions and ranging:

- i. a highly positive distribution (1),
- ii. a moderately positive distribution (2),
- iii. a symmetric distribution (3),
- iv. a moderately negative distribution (4),
- v. a highly negative distribution (5).

Differences in skewness for distributions are illustrated in Figure 2.

The process of splitting starts with designating a center of each group. It is the feature that distribution is:

- i. the most positively skewed,
- ii. the most negatively skewed,
- iii. the closest to being symmetrical,
- iv. of skewness value in the middle between the most positively distributed one and 0,
- v. of skewness value in the middle between the most negatively distributed one and 0,

The rest of features are assigned to the closest group.

3. The principal component analysis is performed for each of six separate groups of features and results in six sets of new features:

- (a) F_{SB} - for binary features,
- (b) F_{S1} - for highly positive distributed features,
- (c) F_{S2} - for moderately positive distributed features,
- (d) F_{S3} - for symmetrically distributed features,
- (e) F_{S4} - for moderately negative distributed features,
- (f) F_{S5} - for highly negative distributed features.

4. The final result set of new features is a sum denoted as (3):

$$F_{igPCA} = F_{SB} \cup F_{S1} \cup F_{S2} \cup F_{S3} \cup F_{S4} \cup F_{S5} \quad (3)$$

Figure 2: Skewness of frequency distribution

threatening. Therefore correct identification of arrhythmia is essential for further medical treatment [28].

Cardiac arrhythmia is usually diagnosed by measuring the heart activity using electrocardiograph (ECG) and then analyzing the recorded data. Parameter values comes in the form of ECG waveforms and can be used along with other information on the patient, including his age and medical history. The spectral characteristics and time domain features of the ECG may be combined to improve arrhythmia classification [29, 30]. However, it may be difficult for a medical staff to find dependencies and irregularities in complex and long ECG recordings. Therefore, new solutions for automating arrhythmia diagnosis are considered.

The dataset for experimental studies derives from UCI Machine Learning Repository [8] and can be referenced as “ARRHYTHMIA” dataset. This dataset contains medical recordings of 452 patients. Each sample is described by 279 attributes. They include 4 parameters referencing general patient’s data: his age, sex, height and weight. The rest of features relates to ECG recordings, e.g. an average width (in msec.) of linear Q wave, an average width (in msec.) of linear R wave, an average width (in msec.) of linear S wave, the number of intrinsic deflections and an existence of ragged R wave. Every observation is described by a label designating one of 16 different classes. The first class corresponds to normal ECG recording with no arrhythmia. Recordings labeled by numbers 2–15 are classified as different types of cardiac arrhythmia, whereas class 16 refers to unlabeled patients. The details of the dataset were also introduced in [31].

4 Data description

Due to the growing problem of heart diseases, the computer improvement of their diagnostics becomes of great importance. One of the most common heart diseases is cardiac arrhythmia. It refers to a medical condition, when a heart beats irregularly, which may be harmless or even life

5 Experimental results and discussion

The aim of the experiments was to examine the performance of proposed in-group Principal Component Analysis for space reduction of ECG data to support its classification. The experiments were conducted on the reference "ARRHYTHMIA" dataset described in Section 4.

The investigation included following stages:

1. Data preparation with exclusion of parameters with uniform values (or parameters for which the number of non-zero values was below 10)
2. Separating features with binary values and putting them together in one group (B).
3. Grouping the rest features into 5 groups of similar distribution according to the methodology introduced in Section 3.
4. Classification based on all features using kNN with $k = 5$.
5. Standardization of features and performing PCA in all the features.
6. Choosing the number of principal components to be retained by explained variance ratio and a scree plot.
7. Classification based on retained number of principal components using kNN with $k = 5$.
8. Performing PCA and carrying out the step 6. for each of six groups.
9. Classification based on a set of features resulting from the sum of principal components derived from the step 8.
10. Comparing results of classification.

All steps of the experimental procedure were implemented in Python language [32]. The key stages and their results are described in the following subsections.

5.1 Data preparation

The original "ARRHYTHMIA" dataset contained parameters of uniform values (usually zeros) or attributes where the number of rows with non-zero values was below 10 (< 2% of all recordings). These features were removed and the final number of attributes that were subjected for further analysis equaled 192.

5.2 Grouping features

The whole set of features was divided into 6 groups as described in Table 1. The first column represents group name, the second - number of features that were assigned to that group, the third and the fourth columns contain the type of distribution and the range of skewness values. The last column shows a frequency distribution plot of the center of that group as indicated in the method description (Section 3).

5.3 Principal component analysis

The key element of the principal component analysis is determination of the number of principal components to be kept for further analysis. Different stopping criteria can be applied [33]. In our approach scree plot [34] was used. Moreover, according to Kaiser rule, all components with

Figure 3: Scree plot of PCA

Figure 4: Scree plot of igPCA in the Group 1

Figure 5: Scree plot of igPCA in the Group 2

Table 1: Characteristics of separated groups

Group name	No of features included	Type of distribution	Skewness range	Distribution of the center feature not applicable
Group B	14	binary	not applicable	
Group 1	42	highly positive	2.5001 – 15.3435
Group 2	39	moderately positive	0.4110 – 2.4916
Group 3	46	close to symmetrical	-0.3375 – 0.3932
Group 4	32	moderately negative	-2.5776 – -0.4483
Group 5	19	highly negative	-10.2626 – -2.6664

Figure 6: Scree plot of igPCA in the Group 3

Figure 7: Scree plot of igPCA in the Group 4

eigenvalues under 1.0 were dropped [35]. As a result we decided to retain:

Figure 8: Scree plot for PCA performed in the Group 5

Table 2: Classification accuracies

Dataset	No of attributes	Accuracy $\pm$ SD	p-value
Original	192	0.72 ± 0.10	N/A
After PCA	12	0.75 ± 0.07	< 0.001
After igPCA	10	0.79 ± 0.10	< 0.001

- 13 principal components after PCA performed for the whole dataset (Figure 3.),
- no principal component for binary features (Group B),
- 3 principal components for every other group (see Figures 4–8)

5.4 Classification results

The accuracy of classification based on a set of features resulting from the sum of principal components as proposed in our method igPCA, described in Section 3, was compared to classification accuracies obtained for:

- the original set of features,
- dataset described by principal components of analysis performed on original dataset.

The experiments have been conducted using 10-fold cross-validation.

To verify the experimental results, the statistical Mann–Whitney U test was applied. The outcomes were regarded as statistically significant if p -value < 0.05 was observed.

The results of comparison are presented in Table 2. The first column represents datasets, the second number of features for the original dataset or principal components for transformed datasets, and the third average accuracy results of ten repeated classification processes.

The last column of the Table 2 shows the statistical significance of differences when compared to the original dataset. One can notice that the feature transformation process significantly improves classification regardless of

PCA approach. However, it must be emphasized that igPCA keeps less principal components than PCA and allows to draw additional conclusions, pointing out the importance of different groups of attributes. In terms of arrhythmia, the binary features appeared to have no impact on the final structure of principal components. Moreover, it is also noticeable that the remainder of groups equally contribute to the final results.

6 Conclusions

In this paper, we evaluate the possibility to apply in-group principal component analysis to improve handling ECG recordings and their analysis in terms of arrhythmia classification.

Cardiac arrhythmia is one of the most common heart diseases. Its diagnosis demands measuring the heart activity using electrocardiograph (ECG) and collecting the data as multidimensional medical datasets. Their storage, analysis and knowledge extraction become a highly complex issues. Feature reduction not only enables saving storage and computing resources, but primarily it makes the process of data interpretation more comprehensive.

We proposed a new igPCA (in-group Principal Component Analysis) method for feature reduction. We assumed that the set of attributes can be split into subgroups of similar characteristic and then subjected to principal component analysis. The method has been verified by experiments done on a dataset of ECG recordings. The obtained effects have been evaluated regarding the number of kept features and accuracy of classification of arrhythmia types. Experiment results confirmed by statistical verification showed the advantage of the presented approach compared to base principal component analysis. The igPCA outperformed base PCA approach in terms of classification accuracy and number of principal components required to perform analysis. Moreover, it enables revealing insight into intrinsic data structure.

Each medical study has its own design, measurement characteristics and various assumptions about data structure. Therefore, there is no universal statistical method that deals with different datasets, and new investigations should be performed. Further studies will involve other medical problems, e.g. juvenile growth restriction disorder in children where the problem of feature extraction was encountered [6] They will also focus on investigating the impact of amounts of missing data on the validity of an imputation technique. Some other methods for dealing with grouped feature selection will be considered.

References

- [1] Bellman R.E., *Adaptive control processes: a guided tour*, Princeton University Press, 2015.
- [2] Chen L., Curse of dimensionality, In: Liu L., Özsu M.T. (Eds.), *Encyclopedia of Database Systems*, Springer US, 2009.
- [3] Keogh E., Mueen A., Curse of dimensionality, In: Sammut C., Webb G.I. (Eds.), *Encyclopedia of Machine Learning*, Springer Science & Business Media, 2011.
- [4] Abdi H., Williams L.J., *Principal component analysis*, Wiley Interdisciplinary Reviews: Computational Statistics, 2010, 2(4), 433-459.
- [5] Wosiak A., Zakrzewska D., *Feature Selection for Classification Incorporating Less Meaningful Attributes in Medical Diagnostics*, In: M. Ganzha, L. Maciaszek, M. Paprzycki (Eds.), *Proceedings of the 2014 Federated Conference on Computer Science and Information Systems (7-10 September 2014, Warsaw, Poland)*, *Annals of Computer Science and Information Systems*, 2014, 235-240.
- [6] Wosiak A., Zakrzewska D., *Integrating Correlation-Based Feature Selection and Clustering for Improved Cardiovascular Disease Diagnosis, Complexity*, 2018, DOI: 10.1155/2018/2520706.
- [7] Byczkowska-Lipińska L., Wosiak A., *Instance Selection Techniques in Reduction of Data Streams Derived from Medical Devices*, *Przegląd Elektrotechniczny*, 2017, 93, 115-118.
- [8] Lichman M., *UCI Machine Learning Repository*, 2017, <http://archive.ics.uci.edu/ml>
- [9] Wojciechowski A., Staniucha R., *Mouth features extraction for emotion classification*, In: M. Ganzha, L. Maciaszek, M. Paprzycki (Eds.), *Proceedings of the 2016 Federated Conference on Computer Science and Information Systems (11-14 September 2016, Gdańsk, Poland)*, *Annals of Computer Science and Information Systems*, 2016, 1685-1692.
- [10] Liu G., Kong L., Gopalakrishnan V., *A Partitioning Based Adaptive Method for Robust Removal of Irrelevant Features from High-dimensional Biomedical Datasets*, In: *AMIA Summits on Translational Science Proceedings*, American Medical Informatics Association, 2012, 52-61.
- [11] Pyle D., *Data preparation for data mining*, Morgan Kaufmann, 1999.
- [12] Donoho D.L., *High-dimensional data analysis: The curses and blessings of dimensionality*, *AMS Math Challenges Lecture*, 2000, 1-33.
- [13] Szajerman D., Napieralski P., Lecointe J.P., *Joint analysis of simultaneous EEG and eye tracking data for video images*, *COMPEL - The international journal for computation and mathematics in electrical and electronic engineering*, 2018, 37(5), 1870-1884.
- [14] Hall M.A., *Correlation-based feature selection for machine learning*, The University of Waikato, 1999.
- [15] Guyon I., Elisseeff A., *An introduction to variable and feature selection*, *Journal of Machine Learning Research*, 2003, 3, 1157-1182.
- [16] Chandrashekar G., Sahin F., *A survey on feature selection methods*, *Computers & Electrical Engineering*, 2014, 40(1), 16-28.
- [17] Kim J.O., Mueller C.W., *Factor analysis: Statistical methods and practical issues*, Sage, 1978.
- [18] Duntelman G. H., *Principal components analysis*, Sage, 1989.
- [19] Smith L.I., *A tutorial on principal components analysis*, Cornell University, USA, 2002.
- [20] Groth D., Hartmann S., Klie S., Selbig J., *Principal components analysis*, *Computational Toxicology: Volume II*, 2013, 527-547.
- [21] Bressan M., Vitria J., *On the selection and classification of independent features*, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2003, 25(10), 1312-1317.
- [22] Tibshirani R., *Regression shrinkage and selection via the lasso*, *Journal of the Royal Statistical Society: Series B (Methodological)*, 1996, 58(1), 267-288.
- [23] Yuan M., Lin Y., *Model selection and estimation in regression with grouped variables*, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2006, 68(1), 49-67.
- [24] Meier L., Van De Geer S., Bühlmann, P., *The group lasso for logistic regression*, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2008, 70(1), 53-71.
- [25] Jacob L., Obozinski G., Vert J. P., *Group lasso with overlap and graph lasso*, In: *Proceedings of the 26th Annual International Conference on Machine Learning*, ACM, 2009, 433-440.
- [26] Li H., Wu X., Li Z., Ding W., *Group feature selection with streaming features*, In: *IEEE 13th International Conference on Data Mining*, 2013, 1109-1114.
- [27] Murkute N., Borkar, P., *Effective Method of Feature Selection on Features Possessing Group Structure*, *International Journal of Computer Science and Information Technologies*, 2016, 7(3), 1111-1115.
- [28] Gupta V., Srinivasan S., Kudli S.S., *Prediction and Classification of Cardiac Arrhythmia*, Stanford University, 2014.
- [29] Lipinski P., Yatsymirskyy M., *Efficient 1D and 2D Daubechies wavelet transforms with application to signal processing*, In: *International Conference on Adaptive and Natural Computing Algorithms*, Springer, Berlin, Heidelberg, 2007, 391-398.
- [30] Clifford G.D., Azuaje F., Mcsharry P., *ECG statistics, noise, artifacts, and missing data*, *Advanced methods and tools for ECG data analysis*, 2006, 55-100.
- [31] Guvenir H.A., Acar B., Demiroz G., Cekin A., *A supervised machine learning algorithm for arrhythmia analysis*, *Computers in Cardiology*, IEEE, 1997, 433-436.
- [32] van Rossum G., *Python tutorial*, Technical Report CS-R9526, Centrum voor Wiskunde en Informatica (CWI), Amsterdam, 1995.
- [33] Jackson D. A., *Stopping rules in principal components analysis: a comparison of heuristic and statistical approaches*, *Ecology*, 1993, 74(8), 2204-2214.
- [34] Cattell R.B., *The Scree Test For The Number Of Factors*, *Multivariate Behavioral Research*, 1966, 1(2), 245-276.
- [35] Raïche G., Walls T.A., Magis D., Riopel M., Blais J.G., *Non-graphical solutions for Cattell's scree test*, *Methodology*, 2013, 9(1), 23-29.