Application of the stochastic gradient method in the construction of the main components of PCA in the task diagnosis of multiple sclerosis in children

Mariusz Topolski[0000−0002−5213−6845]

Department of Systems and Computer Networks,
Faculty of Electronics, Wrocław University of Science and Technology,
Wybrzeże Wyspiańskiego 27, 50-370 Wrocław, Poland
email: mariusz.topolski@pwr.edu.pl

Abstract. Many different medical problems are characterized by quite large spatial dimensions, which causes the task of recognizing patterns to become troublesome. This is a well-known phenomenon called curse of dimensionality. These problems force the creation of various methods of reducing dimensionality. These methods are based on selection and extraction of features. The most commonly used method in literature, regarding the later, is the analysis of the main components of pca. The natural problem of this method is the possibility of applying it to linear space. It is a natural problem to develop the pca concept for cases of nonlinear feature spaces, optimization of feature selection for principal components and the inclusion of classes in the task of supervised learning. An important problem in the perspective of machine learning is not only a reduction of features and attributes but also separation of classes. The developed method was tested in two computer experiments using real data of multiple sclerosis in children. The discussed problem, even from the very nature of the data itself, is important because it can contribute to practical implementations in medical diagnostics. The purpose of the research is to develop a method of extracting features with the application of the stochastic gradient method in the task diagnosis of multiple sclerosis in children. This solution could contribute to the increasing quality of classification and thus may be the basis for building systems that support the medical diagnostics in recognition of multiple sclerosis in children.

Keywords: *Principal Components Analysis* · stochastic gradient · recognition of returns · multiple sclerosis.

1 Introduction

Nowadays machine learning techniques are being used in ever more fields, such as broadly understood medicine, neuroimaging, image classification and detection of network attacks. They produce huge amounts of data with many attributes.

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50423-6_3

Such a large dose of information, paradoxically, does not improve the quality of algorithms, and the data itself is expensive to acquire and store. This resulted in the need for methods to reduce the size of the data, without degrading (or even improving) the quality of classifiers. The reason why more information does not mean better classification is the so-called *curse of dimensionality*, described for the first time by Richard Bellman [1]. When adding dimensions to collections, the distances between specific points are constantly increasing. The number of objects needed for proper generalization is also increasing. It is estimated that in the case of linear classifiers this number increases linearly with dimensionality, and squarely in the case of quadratic algorithms. Even worse is the case of non-parametric classifiers, such as neural networks or those using radial base functions, where the number of objects needed for proper generalization increases exponentially [2]. Sometimes the problem of the curse of dimensionality is called *small n large p*" [4].

The curse of dimensionality results in the *Hughes phenomenon* [3]. For a fixed number of samples, recognition accuracy may first increase algorithms increase, but decreases when the number of attributes exceeds a certain optimal value. In addition to the distance between the samples, this is also caused by the noise in the data or insignificant features. *Selection and extraction* (reduction) features are used to reduce the dimensionality of the data. Feature selection is designed to select a subset of the features used for classification, while *feature extraction* is used to transform (e.g., linear) feature space.

2 Methods

Principal Component Analysis belongs to projection methods. The goal of projection methods is to find a mapping from original space with d dimensions for a new one ($k <= d$) space, to minimize information loss [5].

It is an unsupervised learning method, which means it doesn't need class labels. In the case of PCA, the new attributes are created in a way that maximises their variance. The algorithm aims to create new features (the so-called principal components) that will be uncorrelated (orthogonal) and ordered according to decreasing variance. In order for the algorithm to give correct results, the input data should be normalized first. The principal components are eigenvectors of the input attribute covariance matrix. Because the direction is important in them, these lengths are selected 1. Assuming that λ_i is the eigenvalue of the i^{th} eigenvector, after ordering the proportion of total variance is descending derived from the first k vectors can be calculated using the formula:

$$\frac{\lambda_1 + \lambda_2 + \ldots + \lambda_k}{\lambda_1 + \lambda_2 + \ldots + \lambda_k + \ldots + \lambda n} \tag{1}$$

If the original dimensions of the input data are strongly correlated with each other, we get a small number of eigenvectors with large eigenvalues. A large reduction in dimensions is then possible. However, if the dimensions are not strongly correlated, k will be similar to n and it is not possible to reduce the

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50423-6_3

dimensions without losing the initial part of the set variance [5]. If the number of attributes exceeds the number of objects, it is possible to reduce the dimensions to at most to the number of samples [6].

One of the disadvantages of PCA is that it uses a linear transformation, which makes it unsuitable for more complex spaces. The solution to this problem may be to develop a basic algorithm with the so-called *kernel trick*, getting KPCA (*Kernel Principal Component Analysis*).

In order to solve a non-linear problem, one would first have to transform the input space X as a certain highly-dimensional space F using the function $\phi(x)$, and then e.g. calculate the scalar product $< \phi(x), \phi(x') >$. However, it would be computationally complicated. Therefore, choose the $k(x, x') =< \phi(x), \phi(x') >$ for some transformation ϕ [7]. One of the models using this trick is e.g. SVM classifier.

Another idea for developing PCA is, for example, using class labels as in the development of Karhunen-Loève or carrying out selection of features in the space obtained by PCA [8]. In addition to using the standard PCA, new versions are often created to suit specific problems. One such variation of PCA method is *SuperPCA* [12]. It is used in the classification problem related to *hyperspectral imagining* [17]. The method combines PCA with a segmentation algorithm by means of super pixelization.

Another interesting development of PCA is the D*i*PCA (*Dynamic Inner PCA*), method, also used in process monitoring, but focusing on the aspect of data dynamics [13]. Its goal is to maximize covariance between components and their earlier values. It accomplishes this by extracting a model of dynamic hidden variables on which standard PCA is then performed.

When it comes to supervised methods, LDA is also still widely used. An example of the use of linear discriminant analysis is the already mentioned feature extraction for the task of cancer recognition based on microscopic tissue images [11]. A team from India used a different approach to diagnose lung cancer [14], that used computed tomography images as input. In the study, LDA was used to reduce the size of the data (*Optimal Deep Neural Network*). The results showed an improvement in quality compared to previously used classifiers.

Another proposed method is factor-rotation-modified CCPCA analysis. The authors [15] proposed factor rotation in terms of decision-making centroids. The method was used to assess the risk of *lymphocytic leukaemia*.

The article presents a new concept of GPCA for building main components in the pca method. For this purpose, the *stochastic-gradient-optimization* method was used [16].

In the case of GPCA properties and eigenvectors we are looking for a K matrix such that:

$$K_{i,j} = L(Z_i, Z_j), \tag{2}$$

where L is a function of the goal, Z is a standardized variable, k is e.g. the kernel:

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50423-6_3

$$L(Z_i, Z_j) = \sum_{i=1}^{n} \left(x_i - \omega^T Z_j\right)^2, \tag{3}$$

where: $L(Z_i, Z_j)$ is a overall error on the training set, ω^T is a gradient.

By minimizing the function $L(Z_i, Z_j)$ it starts with the selected start-up solution $\omega_0 = 0$. Then the gradient is determined at the point $\omega_{k-1}, \alpha_k \nabla_L (\omega_{k-1})$. The step along the negative gradient is determined one by one:

$$\omega_k = \omega_{k-1} - \alpha_k \nabla_L (\omega_{k-1}), \tag{4}$$

where α_k is the step length determined before the linear search. We calculate the gradient ∇_L using the difference:

$$\frac{\partial \left(Z_i - \omega^T Z_j\right)^2}{\partial \omega_j} = -2 \left(Z_i - \omega^T Z_j\right) Z_i j \tag{5}$$

Finally

$$\nabla_L (\omega) = -2 \left(x_i - \omega^T Z_j\right) Z_j. \tag{6}$$

The number of principal components can now be represented as a linear combination of original variables Z

$$G_{k_{ij}} = \sum_{i=1}^{k} \sum_{j=1}^{m} a_{k_{ij},j} Z_j, \tag{7}$$

where m is the number of primary variables in the training set, w is the number of main components, Z_j is the j-th standardized variable, $G_{k_{ij}}$ is the i-th main component, $a_{k_{ij},j}$ are factor loads.

The developed GPCA method can be used in non-linear feature spaces. Other kernel functions may be proposed depending on the class the problem. In the article we consider a linear case.

3 Experimental set-up

The aim of the research is to build a feature extraction method that will allow more accurate classification of children with multiple sclerosis. The problem is important because the prognosis for the development of the disease is an extremely difficult process. Often, only appropriately selected variables allow for accurate classification of children to certain risk groups. The developed method gives a chance to build a tool that will support the physician in diagnostics and thus can contribute to the correct diagnosis and treatment of children. Because multiple sclerosis does not give initial clear-cut symptoms, well-chosen variables and risk groups can improve the quality of classification. This goal has become the most important reason for undertaking research on the construction of the

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50423-6_3

extraction model, which will form the basis for classification using known algorithms. Similar studies have already been conducted and the developed CCPCA method [15] has found real application in the classification people with lymphocytic leukaemia. Particular attention was paid to the newly developed GPCA concept focusing on the optimization of factor rotation axes using the gradient method.

The real-world dataset was used in own research. Actual data relate to prognosis of multiple sclerosis in children. The data contained 230 instances and 20 features and two classes: 1 – poor prognosis, 2 – good prognosis. The number of respondents in the classes is 110, 120 instances. So we have balanced data.

In the experiments, several methods of extracting features known from the literature have been compared. Including: PCA (*Principal Component Analysis*) [5], KPCA (*Kernel Principal Component Analysis*) [7], CCPCA (Centroid Class Principal Component Analysis) [15], FA (*Factor Analysis*) [9], ICA (*Independent Component Analysis*) [10], GPCA (*Gradient Component Analysis*), which is the proposed proprietary method in this article.

Two experiments were performed in the tests, in which the accuracy score for three classifiers was verified in succession: SVM (*Support Vector Machine*, RF (*Random Forest*) and k-NN (k-*Nearest Neighbours*).

The *accuracy score metric* was used to assess the quality of the classification. Wilcoxon signed rank test at statistical significance level $\alpha = 0.05$, was used to assess the differences between accuracy for different methods and algorithms. A five-stratified cross-validation was used in all experiments.

4 Experimental evaluation

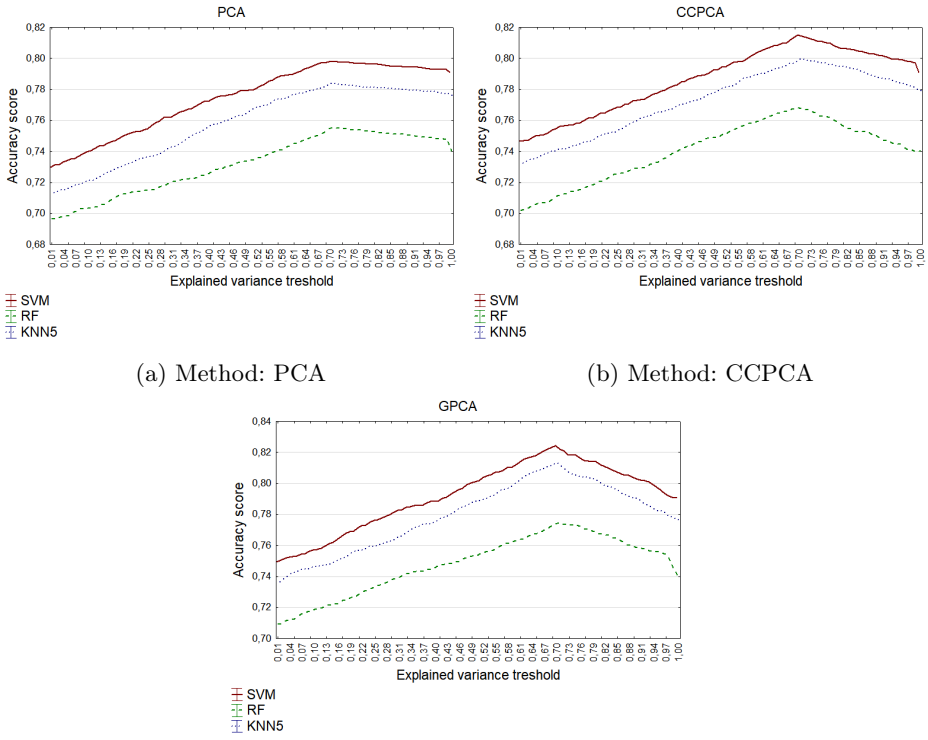
The conducted research was divided into two experiments. The results of the second experiment depend on the first experiment. In the first experiment, the number of principal components were determined experimentally for the PCA, CCPCA and GPCA methods, which explain the set threshold of total variance. Thanks to this approach, we control the selection of main components, and thus the number of features that will form the basis of the classification. The thresholds for which the best algorithm classifications were obtained were included in the second experiment.

4.1 Experiment 1 - Determining the quality of the classification depending on the threshold of total explained variance

Experiment 1 was carried out for three PCA, CCPCA and GPCA methods. The thresholds of explained total variance were adopted by 1 to 100. The study was conducted on three algorithms SVM, RF and k-NN. The results are presented in the chart Figure 1 and 2.

The results of the tests in Experiment 1 show that for each PCA, CCPCA and GPCA method there is a threshold of total variance at which the quality of all classifiers is the highest. As you can see, these thresholds are consistent and the

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50423-6_3



(a) Method: PCA

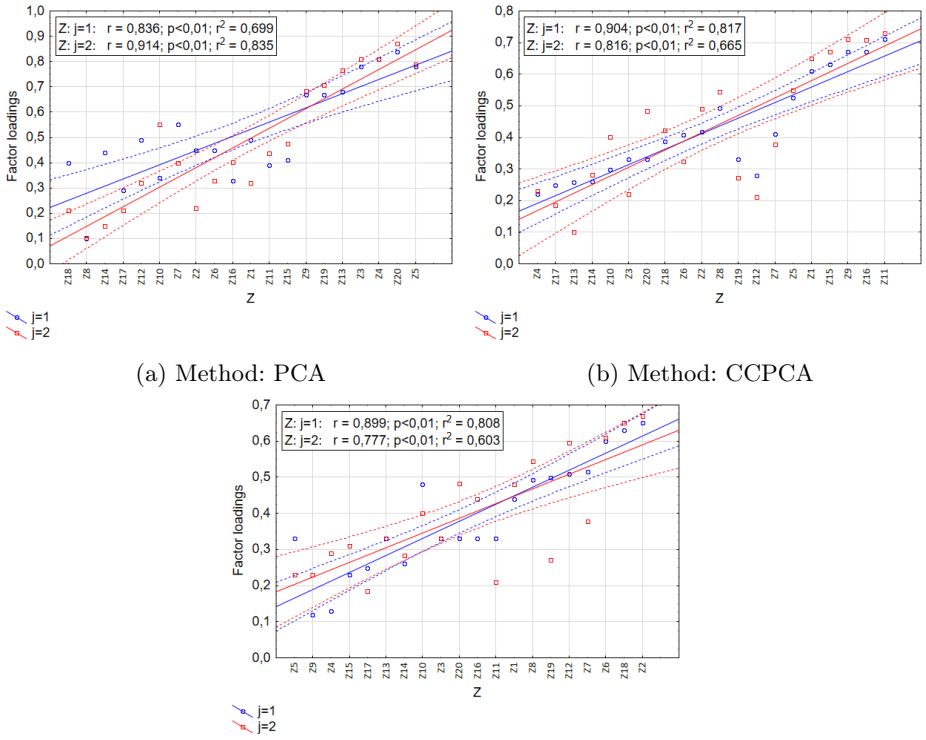
(b) Method: CCPCA

(c) Method: GPCA

Fig. 1: The plot of the dependence of the classification accuracy on the applied thresholds of the total explained variance for the methods of extracting the PCA, CCPCA and GPCA features on 230 teaching standards.

best results of correct classifications with each PCA method and classification algorithm are within 68-72 percent. It should be noted that for threshold 1 all features are taken for classification. In the case of 0.01, we have a situation where there is only one main component that combines are one to three attributes. For the 0.7 threshold, there are 3 main components. Also note that there is a slight data drift for different and near thresholds. However, as you can see, matching attributes to principal components is getting better. Therefore, there is a very interesting conclusion that as the total variance is threshold, the quality of matching attributes to these components increases. Figure 2 shows the results showing which features were assigned to a given principal component. The basis for classification of features into main components was the factor load value $\lambda > 0.6$. The results indicate that we will get a better fit for decision class 2 of the problem for component 1, and class 2 will be better classified by the set of features in components 2 and 3. Based on the GPCA method, the features

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50423-6_3



(a) Method: PCA

(b) Method: CCPCA

(c) Method: GPCA

Fig. 2: Plot of the relationship between the selection of object features for each of the three main components and the factor load values

Z7, *Z8*, *Z10*, *Z12*, *Z14* and *Z18* were rejected, which do not make a significant contribution to explaining object classes.

4.2 Experiment 2. Determining the quality of classification for various methods of feature extraction

The purpose of the experiment is to verify how proprietary CCPCA and GPCA algorithms perform in the task of extracting features against other methods, i.e. PCA, KPCA, FA and ICA. The goal was achieved by checking the quality of real data classification using three algorithms: SVM, RF and k-NN. Based on the results obtained in experiment 1, 70 percent of the total explained variance for the PCA, CCPCA and GPCA methods was selected for the training data set. The Accuracy score obtained and Wilcoxon signed rank test is shown in Table 1. The first measuring points with the names of the algorithms relate to the case without using the feature extraction method. The next results, i.e. PCA, CCPCA, GPCA, KPCA, FA and ICA relate to the classification for a given algorithm after the extraction of features by a given method.

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50423-6_3

Table 1: The results of the experiments for the binary case with application of *accurace-score* metrics. In the columns the algorithms are presented, where NO means lack of extraction of an object's features.

METHOD	SVM	RF	KNN
1 NO	0.791 –	0.740 –	0.750 –
2 PCA	0.798 1	0.755 1	0.770 1
3 CCPCA	0.823 1,2,5,6,7	0.769 1,2,5,6,7	0.828 1,2,5,6,7
4 GPCA	0.826 1,2,5,6,7	0.771 1,2,5,6,7	0.833 1,2,5,6,7
5 KPCA	0.810 1,2	0.764 1,2,7	0.802 1,2,7
6 FA	0.806 1,2	0.759 1,2	0.797 1,2
7 ICA	0.806 1,2	0.757 1,2	0.793 1,2

The first significant conclusion from the research is that after extraction with any of the methods, the quality of classification with each of the three algorithms increased statistically significantly ($p < 0.05$). In the task of feature extraction, the best results are obtained by using the GPCA and CCPCA methods. Classification quality after application of GPCA and CCPCA were statistically comparable. Methods KPCA and FA don't differ significantly from each other. Method ICA for algorithms RF and KNN gave better results than in the case of extraction with the ica method ICA.

5 Conclusions

The purpose of the work was to develop a feature extraction method based on updating the property matrix and eigenvector values. In this task, the stochastic gradients method was used, where the function of the goal was the regression function. The study was conducted on a balanced set describing prognosis of children with multiple sclerosis. In during the analysis, it was possible to create a model that gives promising results for such a task. Two experiments were carried out in the work. The first assumed estimation of the GPCA model parameters, i.e. the threshold of the greedy explained variance giving the best quality of classification, estimation of the belonging of variables to the main components. In experiment 2, the quality of SVM, RF and k-NN algorithm classification was tested for various methods of feature extraction. The obtained results showed that the best extraction method is GPCA and CCPCA . The method of stochastic gradients used in the task of minimizing the error in estimating the matrix of eigenvector values proved to be a good approach. The estimation of GPCA components was also carried out for each decision class. In this way, although

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50423-6_3

the same sets of characteristics for each class in each component were obtained, but different matching attributes of the teaching set, which in turn contributed to improving the quality of classification. The GPCA algorithm proved comparable to CCPCA method which was based on *Varimax* rotation normalized with respect to decision-making centroids. The elaborated method was, as already mentioned, tested on real data with MS disease in children. However, it can be used for other learning collections. In further research, the developed method will be tested on other learning sets, which will confirm the the ability to handle various types of data. The biggest problem that can be encountered in using the stochastic gradient approach is the algorithm step.

References

1. Bellman, R.E., 2015. Adaptive control processes: a guided tour (Vol. 2045). Princeton university press.
2. Jimenez, L.O. and Landgrebe, D.A., 1999. Hyperspectral data analysis and supervised feature reduction via projection pursuit. IEEE Transactions on Geoscience and Remote Sensing, 37(6), pp.2653-2667.
3. Hughes, G., 1968. On the mean accuracy of statistical pattern recognizers. IEEE transactions on information theory, 14(1), pp.55-63.
4. Fort, G. and Lambert-Lacroix, S., 2004. Classification using partial least squares with penalized logistic regression. Bioinformatics, 21(7), pp.1104-1111.
5. Alpaydin, E., 2009. Introduction to machine learning. MIT press.
6. Ringnér, M., 2008. What is principal component analysis?. Nature biotechnology, 26(3), p.303.
7. Schölkopf, B., 2001. The kernel trick for distances. In Advances in neural information processing systems (pp. 301-307).
8. Mao, K.Z., 2005. Identifying critical variables of principal components for unsupervised feature selection. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 35(2), pp.339-344.
9. Jain, P.M. and Shandliya, V.K., 2013. A survey paper on comparative study between principal component analysis (PCA) and exploratory factor analysis (EFA). International Journal of Computer Science and Applications, 6(2), pp.373-375.
10. Hyvärinen, A. and Oja, E., 2000. Independent component analysis: algorithms and applications. Neural networks, 13(4-5), pp.411-430.
11. Kaznowska, E., Depciuch, J., Łach, K., Kołodziej, M., Koziorowska, A., Vongsvivut, J., Zawlik, I., Cholewa, M. and Cebulski, J., 2018. The classification of lung cancers and their degree of malignancy by FTIR, PCA-LDA analysis, and a physics-based computational model. Talanta, 186, pp.337-345.
12. Jiang, J., Ma, J., Chen, C., Wang, Z., Cai, Z. and Wang, L., 2018. SuperPCA: A super pixelwise PCA approach for unsupervised feature extraction of hyperspectral imagery. IEEE Transactions on Geoscience and Remote Sensing, 56(8), pp.4581-4593.
13. Dong, Y. and Qin, S.J., 2018. A novel dynamic PCA algorithm for dynamic data modelling and process monitoring. Journal of Process Control, 67, pp.1-11.

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50423-6_3

14. Lakshmanaprabu, S.K., Mohanty, S.N., Shankar, K., Arunkumar, N. and Ramirez, G., 2019. Optimal deep learning model for classification of lung cancer on CT images. Future Generation Computer Systems, 92, pp.374-382.
15. Topolski, M. Topolska, K., 2019. Algorithm for Constructing a Classifier Team Using a Modified PCA (Principal Component Analysis) in the Task of Diagnosis of Acute Lymphocytic Leukaemia Type B-CLL, Hybrid Artificial Intelligent Systems Springer, pp.390-403.
16. Bootou, L., 2010. Large-Scale Machine Learning with Stochastic Gradient Descent, Proceedings of COMPSTAT'2010 pp 177-186
17. Krawczyk, B., Ksieniewicz, P., Woźniak, M. (2014, June). Hyperspectral image analysis based on colour channels and ensemble classifier. In International Conference on Hybrid Artificial Intelligence Systems (pp. 274-284). Springer, Cham.

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50423-6_3