PCA-based Feature Transformation for Classification: Issues in Medical Diagnostics

Mykola Pechenizkiy*, Alexey Tsymbal**, Seppo Puuronen*
*Department of Computer Science and Information Systems, University of Jyväskylä, Finland, e-mails mpechen@cs.jyu.fi, sepi@cs.jyu.fi
**Department of Computer Science, Trinity College Dublin, Ireland, e-mail Alexey.Tsymbal@cs.tcd.ie

Abstract

The goal of this paper is to propose, evaluate, and compare several data mining strategies that apply feature transformation for subsequent classification, and to consider their application to medical diagnostics. We (1) briefly consider the necessity of dimensionality reduction and discuss why feature transformation may work better than feature selection for some problems; (2) analyze experimentally whether extraction of new components and replacement of original features by them is better than storing the original features as well; (3) consider how important the use of class information is in the feature extraction process; and (4) discuss some interpretability issues regarding the extracted features.

1. Introduction

Current electronic data repositories, especially in medical domains, contain enormous amount of data including also currently unknown and potentially interesting patterns and relations that can be uncovered using knowledge discovery and data mining methods [3]. Commonly supervised machine learning is used, in which there exists a set of training instances (cases) represented by a vector of the values of features (attributes) and the values of the class label. An induction algorithm is used to learn a classifier, which maps the space of feature values into the set of class values. The classifier is later used to classify new instances with unknown classifications (class labels). Inductive learning systems were successfully applied in a number of medical domains, e.g. in localization of a primary tumor, prognostics of recurrence of breast cancer, diagnosis of thyroid diseases, and rheumatology [3].

However, researchers and practitioners realize that the effective use of these inductive learning systems requires data preprocessing before applying a learning algorithm. Especially this is important for multidimensional heterogeneous data, presented by a large number of features of different types. The so-called "curse of dimensionality" pertinent to many learning algorithms, denotes the drastic raise of computational complexity and classification error with data having large number of features. Hence, the dimensionality of the feature space is often reduced before classification is undertaken. There are a number of dimensionality reduction techniques, and according to the adopted reduction strategy they are usually divided into *feature selection* and *feature transformation* (also called feature discovery) approaches. The key difference between feature selection and feature transformation is that in the former only a subset of original features is selected while the latter is based on generation of completely new features. The variants of the latter are *feature extraction* and *feature construction*. Feature construction implies discovering missing information about relationships among the features

Proceedings of the 17th IEEE Symposium on Computer-Based Medical Systems (CBMS'04)
1063-7125/04 $ 20.00 © 2004 IEEE

by inferring or creating additional features while feature extraction discovers a new feature space having fewer dimensions through a functional mapping, keeping as much information in the data as possible [9].

For some problems a feature subset may be useful in one part of the instance space, and at the same time it may be useless or even misleading in another part of it. Therefore, it may be difficult or even impossible to remove irrelevant and/or redundant features from a data set and leave only useful ones by means of feature selection. Feature selection techniques that just assign weights to the individual features are also insensitive to interacted or correlated features. That is why the transformation of the given representation before weighting the features is often preferable.

In this paper we consider several approaches to PCA-based feature transformation for classification and discuss how important the use of class information is when transforming original and selecting extracted features. In Section 2 beside the brief discussion of PCA-based transformations, we consider how to decide whether a feature transformation is useful or not for a problem at hand. In Section 3 we address the problem whether extracted features should be used together with or instead of the original feature space to learn a classifier, and which and how many extracted features are useful for classification. In Section 4 we present the results of experiments with the feature transformation techniques on some problems of medical diagnostics and conclude with preliminary findings and consider the directions of further research. Finally, in Section 5 some interpretability issues with respect to the use of feature transformation are discussed.

2. PCA-based feature transformation for classification

Conventional Principal Component Analysis (PCA) is one of the most commonly used feature extraction techniques. It is based on extracting the axes on which data shows the highest variability [7]. Although PCA "spreads out" data in the new basis, and can be of great help in unsupervised learning, there is no guarantee that the new axes are consistent with the discriminatory features in a classification problem.

Another approach is to account class information during the feature extraction process. One technique is to use some class separability criterion from Fisher's linear discriminant analysis, based on a family of functions of scatter matrices: the within-class covariance, the between-class covariance, and the total covariance matrices. In [15] parametric and nonparametric eigenvector-based approaches that use the within- and between-class covariance matrices and thus do take into account the class information were analyzed and compared. In [14] these approaches were applied to dynamic integration of classifiers. Both parametric and nonparametric approaches use the simultaneous diagonalization algorithm [1] to optimize the relation between the within- and between-class covariance matrices. However, the difference between the approaches is in calculation of the between-class covariance matrix. The parametric approach accounts one mean per class and one total mean, and thus it potentially may extract at most the number of classes minus one features. The nonparametric method tries to increase the number of degrees of freedom in the between-class covariance matrix, measuring the between-class covariances on a local basis.

An important issue is how to decide whether a PCA-based feature transformation approach is appropriate for a certain problem or not. Since the main goal of PCA is to extract new uncorrelated features, it is logical to introduce some correlation-based criterion with a possibility to define a threshold value. One of such criteria is the Kaiser-Meyer-Olkin (KMO) criterion that accounts for both total and partial correlation:

$$KMO = \frac{\sum_i \sum_j r_{ij}^2}{\sum_i \sum_j r_{ij}^2 + \sum_i \sum_j a_{ij}^2}, \quad (1)$$

where $r_{ij} = r(x^{(i)}, x^{(j)})$ is the element of the correlation matrix $\mathbf{R}$ and a_{ij} are the elements of **A** (partial correlation matrix), and

$$a_{ij.X^{(i,j)}} = \frac{-R_{ij}}{\sqrt{R_{ii} R_{jj}}}, \quad (2)$$

where $a_{ij.X^{(i,j)}}$ is a partial correlation coefficient for $x^{(i)}$ and $x^{(j)}$, when the effect of all the other but i and j features denoted as $X^{(i,j)}$ is fixed (controlled), and R_{kl} is an algebraic complement for r_{kl} in the determinant of the correlation matrix $\mathbf{R}$.

It can be seen that if two features share a common factor with other features, their partial correlation a_{ij} will be small, indicating the unique variance they share. And then, if a_{ij} are close to zero (the features are measuring a common factor) KMO will be close to one, while if a_{ij} are close to one (the variables are not measuring a common factor) KMO will be close to zero.

Generally, it is recommended to apply PCA for a data set only if KMO is greater than 0.5. In [11] it was recommended to apply PCA for meta-learning tasks if KMO is greater than 0.6.

3. Which and how many principal components are useful for classification?

In this section we are interested in the problem of selecting the best subset of orthogonally transformed features for subsequent classification, i.e. we are not searching for the best transformation but rather try to find the best subset of transformed components, which allow achieving the best classification.

One common method is to introduce some threshold, e.g. variance accounted by a component to be selected. This results in selecting principal components, which correspond to the largest eigenvalues. The problem with this approach is that the magnitude of eigenvalue depends on data variance only and has nothing to do with class information. In [6] Jollife presents several real-life examples where principal components corresponding to the smallest eigenvalues are correlated with the output attribute. So, principal components important for classification may be excluded because they have small eigenvalues. In [10] another example of such a situation is shown. Nevertheless, criteria for selecting the most useful transformed features are often based on variance accounted by the features to be selected.

In [13], e.g. it is argued that the selection of all the components, the corresponding eigenvalues of which are significantly greater than one, would produce the similar results as if the number of components to select was defined according to the following formula:

$$\#eigenvalues > 1 + 2\sqrt{\frac{\#features - 1}{\#instaces - 1}}, \quad (3)$$

where the number of features and instances should be relatively large.

An alternative approach is to use a ranking procedure and select principal components that

have the highest correlations with the class attribute. Although this makes intuitive sense, there is criticism of such an approach. In [1] this alternative approach was shown to work slightly worse than using components with the largest eigenvalues in the prediction context.

The problem of selecting useful transformed features is not so important for the parametric class-conditional approach since: (1) it takes into account class information, and (2) it extracts only the number of classes minus one component(s).

Another important issue is deciding on whether selected transformed features should be used together with or instead of original features to learn a classifier. In [11] the use of selected transformed features as additional ones for a decision-tree learner, instance-based learner and Naïve Bayes learner is recommended. In [15] we show that the use of selected transformed features only for an instance-based learner can significantly increase its accuracy for many problems.

4. Experimental results

In this study we experimented with conventional PCA feature transformation and a parametric class-conditional approach. We compared the work of four different approaches that combine 3-nearest neighbour classifier (*3-NN*) with conventional PCA feature transformation or parametric class-conditional feature transformation and either use transformed features together with or instead of original ones. We also compare them with the work of *3-NN* without any feature transformation. We limit our study to data sets that have numerical features only. The non-parametric approach that in [15] was shown to have better results on data sets with categorical features than numerical ones is excluded from this study.

The experiments were conducted on five data sets from the UCI repository with different problems of medical diagnostics: Pima Indians *Diabetes*, *Heart* Disease, *Liver* Disorders, *Thyroid* Gland, and Wisconsin Breast *Cancer* (www.ics.uci.edu/~mlearn/MLRepository.html).

In Table 1, for each data set we present the number of instances, the KMO value, the number of features used by every approach and the corresponding accuracy.

Table 1. Experimental results

Dataset	Inst.	KMO	Accuracy of *3-NN* classifier and number of features used									
			Orig. space		PCA		Par		Orig.+PCA		Orig.+Par	
Diabetes	768	.549	.738	8	*.706*	8	.714	1	.734	11	.737	9
Heart	270	.533	.781	13	*.659*	12	**.825**	1	.788	17	.778	14
Liver	345	.551	.612	6	*.591*	5	.632	1	.594	8	**.644**	7
Thyroid	215	.568	.969	5	*.951*	4	.967	2	.970	6	.961	7
Cancer	569	.513	.968	30	*.935*	10	**.978**	1	.968	32	.971	31

For *Diabetes* and *Thyroid* data sets none of feature transformation techniques can improve the work of plain *3-NN* classifier. For *Heart* and *Cancer* data sets 3-NN achieves the highest accuracy results when the new features extracted by parametric approach are used instead of the original ones. And for *Liver* data set the best results are achieved when the feature extracted by the parametric approach is used together with the original ones. It can be seen from the table that KMO is not a relevant criterion (at least for the considered data sets) to decide whether a PCA-based feature transformation technique is worth being applied to a problem of medical diagnostics. Although for every data set KMO was higher than 0.5, the principal components, when used instead of the original features, resulted in the lower accuracy of *3-NN* classifier, and when used together with the original features, never improved the classification accuracy. Therefore, some additional measures beside KMO need to be

considered.

No theoretical analysis has been performed yet to answer whether transformed features need to be used together with or instead of original features. This is an interesting direction of further research.

5. Feature transformation and interpretability

Before arguing for and against feature transformation with respect to the interpretability issue, let us consider first what is commonly meant by interpretability.

Interpretability refers to whether a classifier is easy to understand. It is commonly accepted that rule-based classifiers like a decision tree and associative rules are very easy to interpret, and neural networks and other connectionist and "black-box" classifiers have low interpretability. *k*NN is considered to have very poor interpretability because the unstructured collection of training instances is far from readable, especially if there are many of them. While interpretability concerns a typical classifier generated by a learning algorithm, transparency (or comprehensibility) refers to whether the principle of the method of constructing a classifier is easy to understand (that is a users' subjective assessment). Therefore, for example, a *k*NN classifier is scarcely interpretable, but the method itself is transparent because it appeals to the intuition of humans who spontaneously reason from similar cases. Similarly, interpretability of Naive Bayes can be estimated as not very high, but the transparency of the method is good for example for physicians who find that probabilistic explanations replicate their way of diagnosing, i.e., by summing evidence for or against a disease [8].

However, when feature transformation is applied for rule-based approaches (e.g. association rules), interpretation of produced rules naturally becomes more difficult or even impossible. Since binarization of categorical attributes is required for PCA-based feature transformation, interpretation of results for a data set containing categorical features deteriorates drastically.

The common criticism of instance-based learning is that it does not provide explicit knowledge to the user like association rules do. However, each individual prediction can be explained transparently by uncovering those instances on which the decision is based. This is naturally useful for situations when the end user is rather familiar with previous medical cases than with some complex measures that are used to characterize them. However, a major problem of simple approaches like *k*NN is that the Euclidian distance will not necessarily be suited for finding intuitively similar cases, especially if irrelevant attributes are present. And feature extraction may be of great help in this context of explanations based on similar cases. In [15] we show how different feature extraction techniques "improve" the neighborhood of the instances with respect to their classes in the Euclidian space. The additional benefit of feature extraction is in the possibility of cases' visualization (and performing visual analysis) by projecting them onto 2D or 3D plots.

PCA-based feature extraction for classification can be treated as means of constructive induction. In [2] it is argued that constructive induction, when generating new features, can produce new (emerging) concepts which in turn may lead to a new understanding of a problem and can produce additional knowledge, including new concepts and their relationship to the primary concepts.

PCA-based feature transformations allow to summarize the information from a large number of features into a limited number of components, i.e. linear combinations of the original features. It is argued that, unfortunately, principal components are often difficult to interpret. Many methods of rotating components to improve their interpretability have been proposed. The *varimax* rotation is the most well known of the orthogonal rotation methods. Chapter 8 in

[5] contains a brief description of *varimax* along with several other rotation methods and relevant references.

It should be noticed also that the transformation formulae of principal components may provide useful information for interpretability of results. In order to achieve better interpretability, principal components can be replaced with suboptimal but better interpretable "simple components" [12]. The idea is to allow extracting less variability and there might be a small correlation between the components, but the approach might be advantageous for practical use since interpretability of components is better.

We would like to emphasise that the assessment of interpretability relies on the user's perception of the classifier and the assessment of an algorithm's practicality depends very much on a user's background, preferences and priorities. Most of the characteristics related to practicality can be described only by reporting users' subjective evaluations. Thus, the interpretability issues are still very disputable and difficult to evaluate, and therefore many conclusions on interpretability are relative and rather subjective.

Acknowledgments: This research is partly supported by the COMAS Graduate School of the University of Jyväskylä, Finland. This material is based upon works supported by the Science Foundation Ireland under Grant No. S.F.I.-02IN.1I111. We would like to thank the machine-learning library in java (WEKA) for the source code used in this study.

References

[1] Almoy, T. A simulation study on the comparison of prediction methods when only a few components are relevant. Computational Statistics and Data Analysis 21, 1996, pp. 87-107.

[2] Arciszewski, T., Wnek, J. and Michalski R.S., An Application of Constructive Induction to Engineering Design, Proceedings of the IJCAI-93 Workshop on AI in Design, France, 1993.

[3] Fayyad, U., G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, Advances in Knowledge Discovery and Data Mining, AAAI/ MIT Press, 1997.

[4] Hadi, A. S., Ling, R. F. Some cautionary notes on the use of principal components regression. The American Statistician. 52, 1998, pp. 15-19.

[5] Jackson, J. E. A User's Guide to Principal Components. Wiley & Sons, New York, 1991.

[6] Jollife, I. T. A note on the use of principal components in regression. Applied Statistics. 31, 1982, pp. 300-303.

[7] Jollife, I. T. Principal Component Analysis. Springer-Verlag, New York. 1986.

[8] Kononenko I. Inductive and Bayesian learning in medical diagnosis. Applied Artificial Intelligence 7(4), 1993, pp. 317-337.

[9] Liu H. Feature Extraction, Construction and Selection: A Data Mining Perspective, Kluwer Academic Publishers, 1998.

[10] Oza, N.C., Tumer, K. Dimensionality Reduction Through Classifier Ensembles. Technical Report NASA-ARC-IC-1999-124, Computational Sciences Division, NASA Ames Research Center, Moffett Field, CA, 1999.

[11] Popelínský L. Combining the Principal Components Method with Different Learning Algorithms. In Proc. of 12th European Conference on Machine Learning, Workshop on Integrating Aspects of Data Mining, Decision Support and Meta-Learning, Freiburg, 2001.

[12] Rousson, V., Gasser, T. Simple Component Analysis. Manuscript. 2003. Available from http://www.unizh.ch/biostat/Manuscripts/sca3.pdf

[13] Saporta G. Some simple rules for interpreting outputs of principal components and correspondence analysis. In: Bacelar-Nicolau H., Nicolau F.C., Janssen J.(eds.): Proceedings of ASMDA-99, University of Lisbon, 1999.

[14] Tsymbal A., Pechenizkiy M., Puuronen S., Patterson D.W., Dynamic integration of classifiers in the space of principal components, In: L.Kalinichenko, R.Manthey, B.Thalheim, U.Wloka (Eds.), Proc. Advances in Databases and Information Systems: 7th East-European Conf. ADBIS'03, Dresden, Germany, Lecture Notes in Computer Science, Vol. 2798, Springer-Verlag, 2003, pp. 278-292.

[15] Tsymbal A., Puuronen S., Pechenizkiy M., Baumgarten M., Patterson D. 2002. Eigenvector-based Feature Extraction for Classification. In: S.M. Haller, G. Simmons (Eds.), Proc. 15th Int. FLAIRS Conference on Artificial Intelligence, AAAI Press, pp. 354-358.