Dimensionality Reduction for Classification Tasks Modelling

Seppo Puuronen
University of Jyväskylä, Department of Computer Science and Information Systems,
P.O. Box 35, FIN-40351, University of Jyväskylä, Finland
Tel: +358 14 2603028, Fax: +358 14 2603011, sepi@cs.jyu.fi, http://www.cs.jyu.fi/

David Patterson
University of Ulster, Northern Ireland Knowledge Engineering Laboratory,
Shore Road, Newtownabbey, Co. Antrim, BT37 0QB, University of Ulster, Northern Ireland, U.K.
Tel: +44 28 90368157, Fax: +4428 90366803, wd.patterson@ulster.ac.uk, http://www.nikel.ulst.ac.uk/

Alexey Tsymbal
University of Jyväskylä, Department of Computer Science and Information Systems,
P.O. Box 35, FIN-40351, University of Jyväskylä, Finland
Tel: +358 14 2602547, Fax: +358 14 2603011, alexey@cs.jyu.fi, http://www.cs.jyu.fi

Mykola Pechenizkiy
University of Jyväskylä, Department of Computer Science and Information Systems,
P.O. Box 35, FIN-40351, University of Jyväskylä, Finland
Tel: +358 14 2604626, Fax: +358 14 2603011, mpechen@cc.jyu.fi, http://www.cs.jyu.fi

KEY WORDS classification model, dimensionality reduction

ABSTRACT

Classification is a typical data-mining task where the value of some attribute for a new instance is predicted based on the given collection of instances for which all the attribute values are known. This paper shows the importance of the dimensionality reduction as a pre-processing step for building of classification models.

We consider both feature extraction and feature selection techniques as tools for dimensionality reduction with the emphasis on PCA-based feature extraction methods, namely conventional, parametric and nonparametric approaches. The research was focused on parametric and nonparametric eigenvector-based approaches that take into account the class information. The first approach optimises the ratio of between-class variance to within-class variance of the transformed data. The second approach is a nonparametric modification of the first one based on local calculation of the between-class covariance matrix. We compare the two approaches with each other, with conventional PCA, and with plain nearest neighbour classification without feature extraction. Among feature selection techniques the forward feature selection and backward feature elimination algorithms are considered and compared with each other and with the feature extraction algorithms being examined.

The experiments are conducted on 21 real-world benchmark data sets from the UCI machine-learning repository and on 10 artificial data sets from ELENA project. The experimental results support the expectations and show the advantages of the approaches that provide dimensionality reduction before building a classifier. The results show the importance of the use of class information in feature extraction for classification and inappropriateness of conventional PCA to feature extraction for classification in many cases. As well, the obtained results represent the general advantages and drawbacks of feature extraction and feature selection techniques. It is shown that there is no method that would be the best one for all problem and therefore the problem of searching for the dependencies between the characteristics of data set and the best suited methods for it is stated.