Feature Extraction for Classification in Knowledge Discovery Systems

Mykola Pechenizkiy[1], Seppo Puuronen[1], and Alexey Tsymbal[2]

[1] University of Jyväskylä
Department of Computer Science and Information Systems,
P.O. Box 35, FIN-40351, University of Jyväskylä, Finland
{mpechen, sepi}@cs.jyu.fi
[2] Trinity College Dublin
Department of Computer Science
College Green, Dublin 2, Ireland
Alexey.Tsymbal@cs.tcd.ie

Abstract. Dimensionality reduction is a very important step in the data mining process. In this paper, we consider feature extraction for classification tasks as a technique to overcome problems occurring because of "the curse of dimensionality". We consider three different eigenvector-based feature extraction approaches for classification. The summary of obtained results concerning the accuracy of classification schemes is presented and the issue of search for the most appropriate feature extraction method for a given data set is considered. A decision support system to aid in the integration of the feature extraction and classification processes is proposed. The goals and requirements set for the decision support system and its basic structure are defined. The means of knowledge acquisition needed to build up the proposed system are considered.

1 Introduction

Data mining applies data analysis and discovery algorithms to discover information from vast amounts of data. A typical data-mining task is to predict an unknown value of some attribute of a new instance when the values of the other attributes of the new instance are known and a collection of instances with known values of all the attributes is given. In many applications, data, which is the subject of analysis and processing in data mining, is multidimensional, and presented by a number of features. The so-called "curse of dimensionality" pertinent to many learning algorithms, denotes the drastic raise of computational complexity and classification error with data having big amount of dimensions [2]. Hence, the dimensionality of the feature space is often tried to be reduced before classification is undertaken.

Feature extraction (FE) is one of the dimensionality reduction techniques. FE extracts a subset of new features from the original feature set by means of some

V. Palade, R.J. Howlett, and L.C. Jain (Eds.): KES 2003, LNAI 2773, pp. 526-532, 2003.

functional mapping keeping as much information in the data as possible [5]. Conventional Principal Component Analysis (PCA) is one of the most commonly used feature extraction techniques. PCA extracts the axes on which the data shows the highest variability [7]. There exist many variations of the PCA that use local and/or non-linear processing to improve dimensionality reduction, though they generally do not use class information [9].

In our research, beside the PCA, we consider also two eigenvector-based approaches that use the within- and between-class covariance matrices and thus do take into account the class information. We analyse them with respect to the task of classification with regard to the learning algorithm being used and to the dynamic integration of classifiers (DIC).

During the last years data mining has evolved from less sophisticated first-generation techniques to today's cutting-edge ones. Currently there is a growing need for next-generation data mining systems to manage knowledge discovery applications. These systems should be able to discover knowledge by combining several available techniques, and provide a more automatic environment, or an application envelope, surrounding this highly sophisticated data mining engine [4].

In this paper we consider a decision support system (DSS) approach that is based on the methodology used in expert systems (ES). The approach combines feature extraction techniques with different classification tasks. The main goal of such a system is to automate as far as possible the selection of the most suitable feature extraction approach for a certain classification task on a given data set according to a set of criteria.

In the next sections we consider the feature extraction process for classification and present the summary of achieved results. Then we consider a decision support system that integrates the feature extraction and classification processes, describing its goals, requirements, structure, and the ways of knowledge acquisition. As a summary the obtained preliminary results are discussed and the focus of further research is described.

2 Eigenvector-Based Feature Extraction

Generally, feature extraction for classification can be seen as a search process among all possible transformations of the feature set for the best one, which preserves class separability as much as possible in the space with the lowest possible dimensionality [5]. In other words we are interested in finding a projection **w**:

$$\mathbf{y} = \mathbf{w}^T \mathbf{x} \tag{1}$$

where **y** is a $p' \times 1$ transformed data point (presented using p' features), **w** is a $p \times p'$ transformation matrix, and **x** is a $p \times 1$ original data point (presented using p features).

In [10] it was shown that the conventional PCA transforms the original set of features into a smaller subset of linear combinations that account for the most of the variance of the original data set. Although it is the most popular feature extraction technique, it has a serious drawback, namely the conventional PCA gives high

weights to features with higher variabilities irrespective of whether they are useful for classification or not. This may give rise to the situation where the chosen principal component corresponds to the attribute with the highest variability but having no discriminating power.

A usual approach to overcome the above problem is to use some class separability criterion [1], e.g. the criteria defined in Fisher linear discriminant analysis and based on the family of functions of scatter matrices:

$$J(\mathbf{w}) = \frac{\mathbf{w}^T \mathbf{S}_B \mathbf{w}}{\mathbf{w}^T \mathbf{S}_W \mathbf{w}} \tag{2}$$

where $\mathbf{S}_B$ is the between-class covariance matrix that shows the scatter of the expected vectors around the mixture mean, and $\mathbf{S}_W$ is the within-class covariance, that shows the scatter of samples around their respective class expected vectors.

A number of other criteria were proposed in [5]. Both parametric and nonparametric approaches optimize the criterion (2) by using the *simultaneous diagonalization algorithm* [5].

In [11] we analyzed the task of eigenvector-based feature extraction for classification in general; a 3NN classifier was used as an example. The experiments were conducted on 21 data sets from the UCI machine learning repository [3]. The experimental results supported our expectations. Classification without feature extraction produced clearly the worst results. This shows the so-called "curse of dimensionality" with the considered data sets and the classifier supporting the necessity to apply some kind of feature extraction in that context. In the experiments, the conventional PCA was the worst feature extraction technique on average. The nonparametric technique was only slightly better than the parametric one on average. However, the nonparametric technique performed much better on categorical data.

Still, it is necessary to note that each feature extraction technique was significantly worse than all the other techniques at least on a single data set. Thus, among the tested techniques there does not exist "the overall best" one for classification with regard to all given data sets.

3 Managing Feature Extraction and Classification Processes

Currently, as far as we know, there is no feature extraction technique that would be the best for all data sets in the classification task. Thus the adaptive selection of the most suitable feature extraction technique for a given data set needs further research. Currently, there does not exist canonical knowledge, a perfect mathematical model, or any relevant tool to select the best extraction technique. Instead, a volume of accumulated empirical findings, some trends, and some dependencies have been discovered.

We consider a possibility to take benefit of the discovered knowledge by developing a decision support system based on the methodology of expert system design [6] in order to help to manage the data mining process. The main goal of the system is to recommend the best-suited feature extraction method and a classifier for a given data set. Achieving this goal produces a great benefit because it might be

possible to reach the performance of the *wrapper* type approach by using the *filter* approach. In the wrapper type approach the interaction between the feature selection process and the construction of the classification model is applied and the parameter tuning for every stage and for every method is needed. In the filter approach the evaluation process is independent from the learning algorithm and the methods, and their parameters' selection process is performed according to a certain set of criteria in advance. However, the additional goal of the prediction of model's output performance requires also further consideration.

The “heart” of the system is the *Knowledge Base* (KB) that contains a set of facts about the domain area and a set of rules in a symbolic form describing the logical references between a concrete classification problem and recommendations about the best-suited model for a given problem. The *Vocabulary* of KB contains the lists of terms that include feature extraction methods and their input parameters, classifiers and their input and output parameters, and three types of data set characteristics: simple measures such as the number of instances, the number of attributes, and the number of classes; statistical measures such as the departure from normality, correlation within attributes, the proportion of total variation explained by the first k canonical discriminants; and information-theoretic measures such as the noisiness of attributes, the number of irrelevant attributes, and the mutual information of class and attribute.

Filling in the knowledge base is among the most challenging tasks related to the development of the DSS. There are two potential sources of knowledge to be discovered for the proposed system. The first is the background theory of the feature extraction and classification methods, and the second is the set of field experiments. The theoretical knowledge can be formulated and represented by an expert in the area of specific feature extraction methods and classification schemes. Generally it is possible to categorise the facts and rules that will be present in the Knowledge Base. The categorisation can be done according to the way the knowledge has been obtained – has it been got from the analysis of experimental results or from the domain theory. Another categorisation criterion is the level of confidence of a rule. The expert may be sure in a certain fact or may just think or to hypothesize about another fact. In a similar way, a rule that has been just generated from the analysis of results by experimenting on artificially generated data sets but has been never verified on real-worlds data sets and a rule that has been verified on a number of real-world problems. In addition to the “trust“ criteria due to the categorisation of the rules it is possible to adapt the system to a concrete researcher's needs and preferences by giving higher weights to the rules that actually are the ones of the user.

4 Knowledge Acquisition from the Experiments

Generally, the knowledge base is a dynamic part of the decision support system that can be supplemented and updated through the knowledge acquisition and knowledge refinement processes [6].

Potential contribution of knowledge to be included into the KB might be found discovering a number of criteria from the experiments conducted on artificially generated data sets with pre-defined characteristics. The results of experiments can be

examined looking at the dependencies between the characteristics of a data set in general and the characteristics of every local partition of the instance space in particular. Further, the type and parameters of the feature extraction approach best suited for the data set will help to define a set of criteria that can be applied for the generation of rules of KB.

The results of our preliminary experiments support that approach. The artificially generated data sets were manipulated by changing the amount of irrelevant attributes, the level of noise in the relevant attributes, the ratio of correlation among the attributes, and the normality of the distributions of classes. In the experiments, supervised feature extraction (both the parametric and nonparametric approaches) performed better than the conventional PCA when noise was introduced to the data sets. The similar trend was found with the situation when artificial data sets contained missing values. The finding was supported by the results of experiments on the LED17, Monk-3 and Voting UCI data sets (Table 1) that are known as ones that contain irrelevant attributes, noise in the attributes and a plenty of missing values. Thus, this criterion can be included in the KB to be used to give preference to supervised methods when there exist noise or missing values in a data set. Nonparametric feature extraction essentially outperforms the parametric approach on the data sets, which include significant nonnormal class distributions and are not easy to learn. This initial knowledge about the nature of the parametric and nonparametric approaches and the results on artificial data sets were supported by the results of experiments on Monk-1 and Monk-2 UCI data sets (Table 1).

Table 1. Accuracy results of the experiments

Dataset	PCA	Par	NPar	Plain
LED17	.395	.493	.467	.378
MONK-1	.767	.687	.952	.758
MONK-2	.717	.654	.962	.504
MONK-3	.939	.990	.990	.843
Voting	.923	.949	.946	.921

5 Discussions

So far we have not found a simple correlation-based criterion to separate the situations when a feature extraction technique would be beneficial for the classification. Nevertheless, we found out that there exists a trend between the correlation ratio in a data set and the threshold level used in every feature extraction method to address the amount of variation in the data set explained by the selected extracted features. This finding helps in the selection of the initial threshold value as a start point in the search for the optimal threshold value. However, further research and experiments are required to check these findings.

One of our further goals is to make the knowledge acquisition process semiautomatic using the possibility of deriving new rules and updating the old ones based on the analysis of results obtained during the self-run experimenting. This process will include generating artificial data sets with known characteristics (simple,

statistical and information-theoretic measures); running the experiments on the generated artificial data sets; derivation of dependencies and definition of criteria from the obtained results and updating the knowledge base; validating the constructed theory with a set of experiments on real-world data sets, and reporting on the success or failure of certain rules.

We consider a decision tree learning algorithm as a mean of automatic rule extraction for the knowledge base. Decision tree learning is one of the most widely used inductive learning methods [12]. A decision tree is represented as a set of nodes and arcs. Each node contains a feature (an attribute) and each arc leaving the node is labelled with a particular value (or range of values) for that feature. Together, a node and the arcs leaving it represent a decision about the path an example follows when being classified by the tree. Given a set of training examples, a decision tree is induced in a "top-down" fashion by repeatedly dividing up the examples according to their values for a particular feature.

In this context, mentioned above data set characteristics and a classification model's outputs that include accuracy, sensitivity, specificity, time complexity and so on represent instance space. And the combination of a feature extraction method's and a classification model's names with their parameter values represent class labels. By means of analysing the tree branches it is possible to generate "if-then" rules for the knowledge base. A rule reflects certain relationship between meta-data-set characteristics and a combination of a feature extraction method and a classification model.

6 Conclusions

Feature extraction is one of the dimensionality reduction techniques that are often used to cope with the problems caused by the "curse of dimensionality". In this paper we considered three eigenvector-based feature extraction approaches, which were applied for different classification problems. We presented the summary of results that shows a high level of complexity in dependencies between the data set characteristics and the data mining process. There is no feature extraction method that would be the most suitable for all classification tasks. Due to the fact that there is no well-grounded strong theory that would help us to build up an automated system for such feature extraction method selection, a decision support system that would accumulate separate facts, trends, and dependencies between the data characteristics and output parameters of classification schemes performed in the spaces of extracted features was proposed.

We considered the goals of such a system, the basic ideas that define its structure and methodology of knowledge acquisition and validation. The Knowledge Base is the basis for the intelligence of the decision support system. That is why we recognised the problem of discovering rules from the experiments of an artificially generated data set with known predefined simple, statistical, and information-theoretic measures, and validation of those rules on benchmark data sets as a prior research focus in this area.

It should be noticed that the proposed approach has a serious limitation. Namely the drawbacks can be expressed in the terms of fragmentariness and incoherence (disconnectedness) of the components of knowledge to be produced. And we

definitely do not claim the completeness of our decision support system. Otherwise, certain constrains and assumptions to the domain area were considered, and the limited sets of feature extraction methods, classifiers and data set characteristics were considered in order to guarantee the desired level of confidence in the system when solving a bounded set of problems.

Acknowledgments

This research is partly supported by the COMAS Graduate School of the University of Jyväskylä, Finland and Science Foundation, Ireland. We would like to thank the UCI ML repository of databases, domain theories and data generators for the data sets, and the MLC++ library for the source code used in this study.

References

[1] Aivazyan, S.A.: Applied Statistics: Classification and Dimension Reduction. Finance and Statistics, Moscow, 1989.
[2] Bellman, R., Adaptive Control Processes: A Guided Tour, Princeton University Press, 1961.
[3] Blake, C.L., Merz, C.J. UCI Repository of Machine Learning Databases [http://www.ics.uci.edu/~mlearn/ MLRepository.html]. Dept. of Information and Computer Science, University of California, Irvine CA, 1998.
[4] Fayyad U.M. Data Mining and Knowledge Discovery: Making Sense Out of Data, *IEEE Expert*, Vol. 11, No. 5, Oct., 1996, pp. 20-25.
[5] Fukunaga, K. Introduction to Statistical Pattern Recognition. Academic Press, London, 1991.
[6] Jackson P. Introduction to Expert Systems, 3rd Edn. Harlow, England: Addison Wesley Longman, 1999.
[7] Jolliffe, I.T. Principal Component Analysis. Springer, New York, NY. 1986.
[8] Kohavi, R., Sommerfield, D., Dougherty, J. Data mining using MLC++: a machine learning library in C++. Tools with Artificial Intelligence, IEEE CS Press, 234-245, 1996.
[9] Liu H. Feature Extraction, Construction and Selection: A Data Mining Perspective, ISBN 0-7923-8196-3, 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] Tsymbal A., Puuronen S., Pechenizkiy M., Baumgarten M., Patterson D. Eigenvector-based feature extraction for classification, In: *Proc. 15th Int. FLAIRS Conference on Artificial Intelligence*, Pensacola, FL, USA, AAAI Press, 354-358, 2002.
[12] Quinlan, J.R. 1993. C4.5 Programs for Machine Learning. San Mateo CA: Morgan Kaufmann.