Local Dimensionality Reduction within Natural Clusters for Medical Data Analysis

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

Inductive learning systems have been successfully applied in a number of medical domains. Nevertheless, the effective use of these systems requires data preprocessing before applying a learning algorithm. Especially it is important for multidimensional heterogeneous data, presented by a large number of features of different types. Dimensionality reduction is one commonly applied approach. The goal of this paper is to study the impact of natural clustering on dimensionality reduction for classification. We compare several data mining strategies that apply dimensionality reduction by means of feature extraction or feature selection for subsequent classification. We show experimentally on microbiological data that local dimensionality reduction within natural clusters results in a better feature space for classification in comparison with the global search in terms of generalization accuracy.

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 found using knowledge discovery and data mining (DM) methods [1]. 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 [4].

However, researchers and practitioners realize that the effective use of these learning systems requires data preprocessing before they are applied. Especially it 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 on data having large number of features. Hence, the dimensionality of the feature space is often reduced before classification is undertaken. Generally, dimensionality reduction (DR) is only one effective approach to data reduction among others like instance selection or data selection [5]. We see the goal of DR in: (1) reducing the quantity of data with a focus on relevant data, and (2) improving the quality of data and/or its representation for a DM technique. Consequently, achievement of these goals results in a reduced amount of data, relevance of this reduced data to the domain and DM techniques being applied, and finally, improvement of the performance of these DM techniques.

There are a number of DR techniques, and according to the adopted reduction strategy they are usually divided into *feature selection* (FS) and *feature extraction* (FE) (also called feature discovery) approaches. The key difference between FS and FE is that in the former only a

1063-7125/05 $20.00 © 2005 IEEE

subset of original features is selected while the latter is based on generation of a completely new feature space through a functional mapping, keeping in fewer dimensions as much information in the data as possible [5]. Many FS techniques are usually insensitive to interacting or correlated features. That is why the transformation of the given representation before weighing the features is often preferable.

For some problem domains 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 for some datasets to remove irrelevant and/or redundant features and leave only the useful ones by means of global FS. However, if it is possible to find local homogeneous regions of heterogeneous data, then there are more chances to apply FS successfully (individually to each region). For FE the decision whether to proceed globally over the entire instance space or locally on different parts of the instance space is also one of the key issues. It can be seen that despite being globally high dimensional and sparse, data distributions in some domain areas are locally low dimensional and dense, e.g. in physical movement systems.

One possible approach for local FS or local FE would be clustering (partitioning) of the whole data set into smaller regions. Generally, different clustering techniques can be used for this purpose, e.g. the *k*-means or EM techniques [8]. However, in this paper we emphasize on a possibility to apply so-called *natural clustering* aimed to use contextual features for splitting whole heterogeneous data space into more homogeneous clusters. Usually, features that are not useful for classification alone but are useful in combination with other (context-sensitive) features are called *contextual* (or environmental) features [7].

In this paper we apply our natural clustering approach to the selected part of real clinical database trying to construct data models that would help in the prediction of antibiotic resistance and in understanding its development. The analysis of microbiological data included in antibiograms collected at different institutions over different periods of time is considered as one of the most important activities to restrain the spreading of antibiotic resistance and to avoid the negative consequences of this phenomenon [2].

In our experimental study we apply *k*-nearest neighbor classification (*k*NN) to build antibiotic sensitivity prediction models. We apply the principle of natural clustering, grouping the instances into partitions related to certain pathogen types. We apply three different wrapper-based sequential FS techniques and three PCA-based FE techniques globally and locally and analyze their impact on the performance of *k*NN classifier.

The paper is organized as follows. In Section 2 we briefly consider dimensionality reduction techniques used in the study. In Section 3 the data used throughout the experiments are described. In Section 4 we present the results of experiments with the FE and FS techniques applied globally for the whole data set and locally in clusters for further classification. Finally, in Section 5 we briefly conclude with a summary and present the directions of further research.

2. Dimensionality reduction techniques used in the study

Principal Component Analysis (PCA) is one of the most commonly used FE techniques. It is based on extracting the axes on which data shows the highest variability [3]. Although PCA "spreads out" the data in the new basis (new extracted axes), 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 FE process. One technique is to use some class separability criterion (e.g., 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. Parametric and nonparametric eigenvector-based approaches that use the within- and between-class covariance matrices thus taking into account class information have been analyzed and compared [6]. Both parametric and nonparametric approaches use the simultaneous diagonalization algorithm to optimize the relation between the within- and between-class covariance matrices. The difference between the approaches is in calculation of the between-class covariance matrix. The parametric approach accounts for one mean per class and one total mean, and therefore may extract at most *number_of_classes-1* 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.

Greedy hill climbing is one of the simplest search strategies that consider sequential changes to the current feature subset. Often, it is just the addition or deletion of a single feature from the subset. We selected the most commonly used sequential strategies for FS: forward feature selection (FFS), backward feature elimination (BFE), and bidirectional search (BS). The first strategy starts with no features and successively adds a new one. On the contrary, the second one begins with all the features and step-wisely deletes features one-by-one. Bidirectional search proceeds in the both forward and backward directions in turn.

The search algorithms that implement these strategies may consider all possible changes to the current subset and then select the best, or may simply choose the first change that improves the merit of the current feature subset. In either case, once a change is accepted, it is never reconsidered (that is why the name "greedy"). The FS process can stop adding/deleting features when none of the evaluated subsets improves the previous result or, alternatively, the search can continue to produce and evaluate new feature subsets while the result does not start to degrade. The evaluation of selected feature subset in our study was based on a *wrapper* paradigm that assumes interaction between the FS process and the classification model.

3. Data Used in the Experiments

The data used in our analysis were collected in the Hospital of N.N Burdenko, Institute of Neurosurgery using the Vitek-60 analyzer (developed by *bioMérieux*) over the years 1997-2003 and information systems "Microbiologist" (developed by the Medical Informatics Lab of the institute) and "Microbe" (developed by Russian company "MedProject-3").

Each instance of the data used in the analysis represents one *sensitivity test* and contains the following features: *pathogen* that is isolated during the microbe identification analysis, *antibiotic* that is used in the sensitivity test and the *result* of the sensitivity test (sensitive *S*, resistant *R*, or intermediate *I*), obtained from "Vitek" according to the guidelines of National Committee for Clinical Laboratory Standards (NCCLS). The information about sensitivity analysis is connected with a *patient*, his/her demographical data (*sex*, *age*) and hospitalization in the Institute (*main department*, whether the test was taken while patient was in *ICU*, *days spent* in the hospital before, etc.). Each instance of microbiological test in the database corresponds to a single specimen that may be blood, liquor, urine, etc. In this pilot study we focus on the analysis of meningitis cases only, and the specimen is liquor.

For the purposes of this exploratory analysis we picked up 4430 instances of sensitivity tests related to the meningitis cases of the period of January 2002 – July 2004. We introduced grouping features for pathogens and antibiotics so that 17 pathogens and 39 antibiotics were combined into 6 and 15 groups respectively. Thus, each instance had 28 features that included information corresponding to a single sensitivity test augmented with data concerning the type of the antibiotic used and the isolated pathogen, and clinical features of the patient and his/her demographics, and the microbiology test result as the class attribute.

The data is relatively high dimensional and heterogeneous; heterogeneity is presented by a number of contextual (environmental) features. Semantically, the *sensitivity* concept is related first of all to the *pathogen* and *antibiotic* concepts. For our study binary features that describe the pathogen grouping were selected as prior environmental features, and they were used for hierarchical natural clustering (the hierarchy was introduced by the grouping of the features). In our database, the whole data set can be divided into two nearly equal natural clusters: *gram+* and *gram–*. Then, the *gram+* cluster consists of the *staphylococcus* and *enterococus* clusters, and *gram–* cluster consists of the *enterobacteria* and *nonfermentes* clusters.

4. Experimental results

In our experimental studies we used an instance-based classifier (*k*NN), the FFS, BFE, and BS feature selection techniques available in the machine learning library with Java implementation "WEKA 3.4.2" [8]. We used the conventional PCA and the class-conditional parametric (Par) and nonparametric (NPar) FE techniques [6], which we implemented within the same library. We used *k*=7 and the inverse distance for *k*NN since these parameters were found to be the best combination for our data in our pilot studies. The threshold for variance covered was set to 95% for each FE technique and for NPar we used parameter settings *k*NN = 7 and *alpha* = 1.

The main results of experiments are given in Table 1. Each row of the table contains the name of the dataset (cluster), number of instances in it, accuracy of 7-*NN* classifier and the number of features used for each FS (FFS, BFE, and BS) and FE (PCA, Par, NPar) technique averaged over 30 test runs. The column *noFS* is related to the case when no FS was applied and *noFE* to the case when no FE was applied but all the categorical features were binarized.

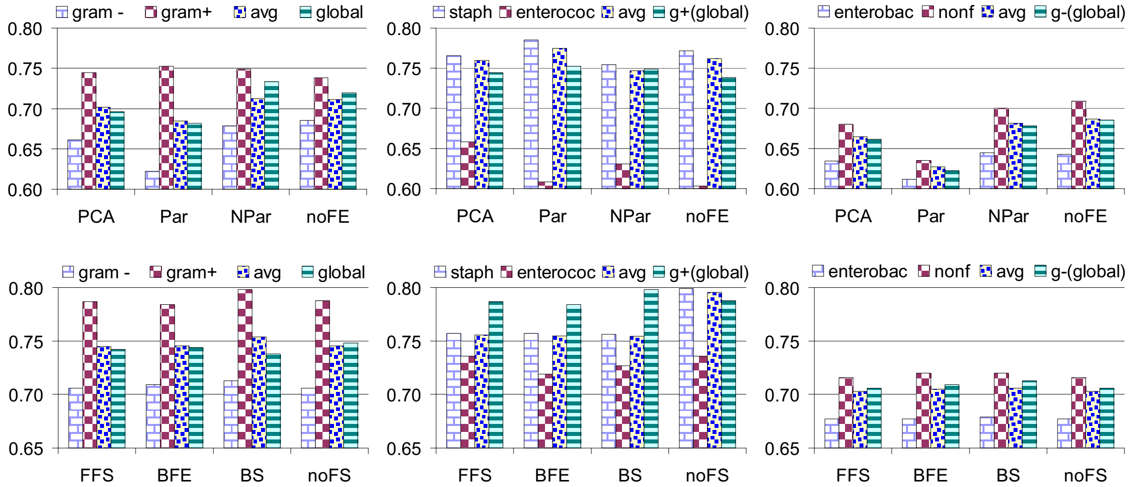
Table 1. Basic experimental results

Dataset	Inst	Accuracy of 7-*NN* classifier and number of features used															
		Feature Selection						noFS		Feature Extraction						noFE	
		FFS		BFE		BS				PCA		Par		NPar			
global	4430	.742	8	.744	8	.738	8	.748	28	.696	24	.682	1	.734	39.1	.719	44
gram– (g–)	2296	.706	6	.709	6	.713	6	.706	24	.662	31	.622	1	.678	31.5	.685	34
gram+ (g+)	2 134	.787	5	.784	5	.798	5	.788	24	.745	19	.752	1	.749	32.6	.738	35
eterobac	783	.677	4	.677	4	.679	4	.677	23	.635	16	.612	1	.644	28	.643	31
nonferm	1 513	.716	7	.72	8	.72	9	.716	23	.680	30	.635	1	.700	26.8	.709	31
staphiloc.	2013	.799	5	.757	5	.756	5	.799	23	.766	20	.785	1	.754	33.5	.772	37
enteroc.	121	.736	3	.719	3	.727	4	.736	23	.658	11	.608	1	.631	21.8	.603	28
best	4430	.730	5	.731	5	.733	5	.749	24	.709	19	.696	1	.711	33	.722	36

The first row corresponds to the (*global*) results on the whole data set. The last row corresponds to overall accuracy achieved with the most appropriate selection of sub data sets (clusters): *staphylococcus*, *enterococus*, and *gram–* (*best*). We can see that in many cases the number of features (original or transformed) selected is different in different clusters and this number depends on whether DR was applied globally or locally. This also supports our hypothesis about the heterogeneity of data.

In Figure 1 comparison of local and global results of 7-NN classifier for 7 different clusters (including the whole data set) are shown. Results show similar behavior of FS and FE across the 7 different clusters. Analyzing the histograms one by one we can see that the DR techniques for our data result in the best classification accuracy when applied locally to *staphylococcus*, *enterococcus*, and *gram–* clusters.

1063-7125/05 $20.00 © 2005 **IEEE**



a) whole dataset b) 'gram+' cluster c) 'gram–' cluster

Figure 1. Comparison of local vs. global 7-NN accuracy results for the a) whole data, b) 'gram+' cluster, c) 'gram–' cluster with and without applying FE (top part) and FS (bottom part).

Applying 7-NN with DR locally in the *gram+* and *gram–* clusters (see Figure 1a) does not outperform global accuracy. However, we can see that accuracy results for cluster *gram+* are much higher than for the *gram–* cluster. The FE methods were almost equally good for the *gram+* cluster. For the *gram–* cluster, Par was the worst and NPar was the best. But still 7-NN without any FE performed slightly better for *gram–*. The FS methods had no effect on 7-NN accuracy for the *gram–* cluster. And BS was the only FS method that increased 7-NN accuracy for *gram+*.

In Figure 1b we can see that applying 7-NN and DR techniques individually to the *staphylococcus* and *enterococcus* clusters significantly increases the overall accuracy. Local Par outperforms local 7-NN by 1.3% (avg Par vs. avg noFE), NPar decreases the accuracy of 7-NN by 2%, and PCA has no effect. Local FS decreases the performance of 7-NN by 3.5 - 4.5%. Relatively low accuracy for the *enterococcus* cluster does not decrease much the average accuracy since this cluster is rather small and contains only 5.7% of instances from *gram+* while *staphylococcus* contains 94.3%.

In Figure 1c we can see that dividing the cluster *gram–* further into *enterobacteria* and *nonfermentes* does not increase the accuracy of 7-NN both with and without local FE or FS.

Now, if we compare the FE and FS horizontal triples of histograms we can see that for our data the sequential strategies for FS have no success. Exceptionally, BS was successful when applied individually to *gram+* and *gram–* clusters. The FE methods have more diverse behaviour. So, that PCA is the best one for the *enterococcus* clusters while Par is the best for *staphylococcus* (Figure 1b upper). NPar showed the best accuracy for global FE on the whole data set. This leads to an idea of adaptive selection of FE method for each cluster, so that the use of PCA in one cluster, and Par or NPar in some other cluster may result in significantly higher overall accuracy.

We compare the impact of FS and FE on classification either globally or locally with the most appropriate selection of clusters: *staphylococcus*, *enterococus* (joined into averaged results of *gram+*), and *gram–*. Due to space limitations we do not present a separate figure but list here the main conclusions of this comparison: (1) Natural clustering is useful for our data only by means of FE with any (global or local) DR; (2) FE was useful (it improved generalization accuracy) both when applied globally and locally, while FS increases the accuracy of 7-NN only when applied locally to the *gram+* and *gram–* clusters. This fact indicates that our data is heterogeneous indeed; (3) FS applied locally on this data results in higher accuracy produced by 7-NN comparing to local FE. However we need to point out that this was due to the binarization of categorical features (that is required for FE). 7-NN produces almost 3% higher accuracy results for data presented by original categorical (not binarized) features. Perhaps, by analyzing possible reasons of why the accuracy of 7-NN on this data decreases after binarization, we can improve the overall situation in the FS-FE competition; (4) Par produced very poor results when applied globally, but performed surprisingly well in some of the clusters. NPar was quite stable across different clusters, and it was the best FE technique for the situation when DR was applied globally.

5. Conclusions and future directions

DR is an effective approach to data reduction aimed to focus on relevant features and to improve the quality of data representation for classification. We experimentally compared and showed the benefits of local and global DR by means of FS and FE. In this study we applied the *natural clustering* approach aimed to use contextual features for splitting a real-world clinical data set into more homogeneous clusters in order to construct local models that would help in the better prediction of antibiotic resistance.

The results of our experiments show that proper selection of a local DR technique can lead to significant increase of predictive accuracy comparing to the global 7NN with or without DR. The amount of features extracted or selected locally is always smaller than that in the global space that also shows the usefulness of natural clustering in coping with data heterogeneity.

Our future research efforts are going to be directed towards the comparison of a mixture of FE models for classification built on natural clusters and on clusters produced by traditional clustering techniques. We analyzed spatial contextual features related to categorization of different pathogens. We believe that natural clustering according to time features may give interesting results in different time contexts.

Another challenging goal is the adaptive selection of FE method for each cluster according to certain characteristics of the cluster. So, the appropriate use of PCA in one cluster, and Par or NPar in some other cluster may result in significantly higher overall accuracy.

Acknowledgments: This material is based upon works supported by the COMAS Graduate School of the University of Jyväskylä, Finland and the Science Foundation Ireland under Grant No. S.F.I.-02IN.1I111.

References

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

[2] R.P. Gaynes, Surveillance of nosocomial infections: a fundamental ingredient for quality. Infect Control Hosp Epidemiol 18(7), 1997, pp. 475– 478.

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

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

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

[6] A. Tsymbal, S. Puuronen, M. Pechenizkiy, M. Baumgarten, D. Patterson, Eigenvector-based feature extraction for classification. In Proc. 15th Int. FLAIRS Conf. on Artificial Intelligence, AAAI Press, 2002, pp. 354-358.

[7] P. Turney, The management of context-sensitive features: A review of strategies. In Proc Workshop on Learning in Context-Sensitive Domains at the 13th Int. Conf. on Machine Learning (ICML96), pp. 60-66.

[8] Witten, I., Frank E. Data Mining: Practical Machine Learning Tools with Java Implementations, Morgan Kaufmann, San Francisco, 2000.