Feature Space Optimization

Dana Klimešová[1,2]

[1]Czech University of Life Sciences Prague, Faculty of Economics and Management, Kamýcká 129, 165 21 Praha 6 - Suchdol
klimesova@pef.czu.cz
[2]Institute of Information Theory and Automation of the ASCR,Pod vodárenskou věží 4, 182 00 Prague 8, Czech Republic

Annotation. The contribution deals with the problem of feature selection strategies for the case of high dimensional feature space. In this paper an effective method of feature selection is proposed where two different measures are considered and time and storage complexity is acceptable.

Key words: feature selection, local context, computational complexity, feature significance

1 Introduction

In the last decade the content of context is considered as an inherent part of feature space modelling. The context understanding of objects and phenomena produces extensive feature space that is not so suitable for the following classification or space modelling. On the other hand the decision quality depends on the quality and the amount of information which is in disposal. The large number of features is produced by the methods dealing with the neighbourhood properties evaluation; others are based on the statistical approaches of the first or second order to express object relationships.

From this point of view the effective feature selection methods are still needed in contextual modelling. Roughly speaking there are two main approaches to optimize (in sense to decrease the dimensionality) the feature space [5], [1]:

- transformation of features to achieve effective, non-correlated and evaluated new set of features,
- selection of original features using searching, testing and evaluation of given features with regard to the training data.

2 Goal and Methods

Common goal of both approaches is to reduce the number of features, resp. reduce the dimensionality of feature space $\boldsymbol{W}^n$.

The methods of transformation are based on linear projection $\boldsymbol{T}$ that transforms the data to a new coordinate system and keeps the subspace with largest variance.

$$\boldsymbol{T}: \ \boldsymbol{W}^n \rightarrow \boldsymbol{W}^m, \text{ where } m \leq n. \tag{1}$$

Well known is *PCT, discrete Karhunen-Loeve Decomposition* and others. These methods are used when we are dealing with smaller number of features [3].

2.1 Selection strategy

As mentioned upper the aim is to select significant features from the original set of features ***Y*** as optimal subset ***X*** that contain *d* features where

$$\boldsymbol{X} = \{x_j \mid j = 1,2,\ldots,d \quad x_j \in \boldsymbol{Y}, \quad d \le n\} \tag{2}$$

$$\boldsymbol{Y} = \{y_j \mid j = 1,2,\ldots,n\}. \tag{3}$$

As an optimal the subset is understood that is the best in the sense of appointed criterial function with respect to any other *d*-subset from ***Y***. The two ways exist to solve this task {8]:

- total searching where the whole state space is processed and time complexity is exponential and
- selection of features where the criterial function is evaluated feature by feature without impression of others features. It is also the main disadvantage of this approach.

The following method makes possible to select the best features and through weighting matrix seeks to account the impression of before selected features [9], [4].

3 Results and Discussion

In case we have a large number of features and the training data of given classes ω_i are in disposal the following approach can be applied.

We suppose to have a set of normalized features

$$\boldsymbol{Y} = \{y_j \mid j = 1,2,\ldots,n\}.$$

The distance Q between glasses ω_i , ω_k , for the feature y_j, where $i, k =$ 1, 2, …R and $j = 1, 2, \ldots n$ can be defined as follows:

$$Q_{ikj} = \left[\left(\mu_{ji} - \mu_{jk}\right) / \left(\sigma_{ji} + \sigma_{jk}\right)\right]^2 \tag{4}$$

For all y_j, where μ_{ji} and σ_{ji} denote the mean value and standard deviation of class ω_i in feature y_j.

The measure C_j cumulates the distances of classes considered for given feature j and this measure is used only to select the first feature

$$C_j = \sum_{i=1}^{R} \sum_{\substack{k=1 \\ i \ne k}}^{R} Q_{ikj}\,, \tag{5}$$

and the first feature is selected to fulfil the equation (6).

$$y^{*} = \max_{j} C_{j} \qquad \textbf{(6)}$$

It means that y^{*} is the most significant feature with greatest contribution to the separability of classes.

Consequently we define the matrix $B_{\max}$ as

$$B_{\max} = \lfloor b_{ik} \rfloor_{RxR} = \lfloor Q_{ik*} \rfloor_{RxR} \quad , \qquad \textbf{(7)}$$

and instead of asterisk the selected feature y^{*} is fill in. Matrix elements are the distances of classes for the selected feature. The following feature is selected using measure D_j

$$D_{j} = \sum_{i=1}^{R} \sum_{\substack{k=1 \\ i \neq k}}^{R} \left(\frac{Q_{ikj}}{b_{ik}} \right)^{2} , \qquad j = 1, \ldots, n \qquad \textbf{(8)}$$

and

$$y^{s} = \max_{j} D_{j} \; . \qquad \textbf{(9)}$$

The matrix $B_{\max}$ is step by step reconstructed as follows

$$b_{ik} = \max \left(b_{ik}, Q_{iks} \right). \qquad \textbf{(10)}$$

We continue in this way as long as the required number $d \leq n$ of features is selected. The time complexity of algorithm is $O\left((n + d).R^{2} \right)$ - it means it depends on the number of features, the number of selected features and with square on the number of classes. The estimation of space complexity is $O\left((R + n)^{2} \right)$ - it increase with square over the sum of number of classes and number of features.

The method tries to optimize the selection of features using sequential connecting of before selected features into the analysis of between-classes distances.

4 Conclusions

The contribution is devoted to the problem of the reduction of high dimensional feature space. The proposed method combines the advantages of both approaches and time and storage complexity is acceptable.

5 Acknowledgement

The Project Information and knowledge support of strategic control - MSM 6046070904 supports this work.

Reference

1. Devijver, P.,Kittler, J. *Pattern Recognition.* A statistical Approach. Prentice Hall, Upper Saddle River, N.J. 1982
2. Ferri, F.J., Pudil, P., Novovičová, J. Methods of selecting features from high dimensional pattern spaces. In: *Proc. of the IEEE European Workshop on Computer-Intensive Methods in Control and Signal Proccesing,* 1994 pp. 68-75.
3. Jungwoo, L. Optimized Quadtree for Karhunen-Loeve Transform in Multispectral Image Coding. *IEEE Transaction on image Processing,* vol 8, No. 4, 1999, pp. 453-461.
4. Klimešová, D., Saic, S. Feature Selection Algorithm and Correlation Cobweb. *Pattern Recognition Letters* 19 ,1998, pp.681-685.
5. Liu, H., Motoda, H. Feature Extraction, *Construction and Selection. A Data Mining perspective.* Cluwer Academic Publishers, Boston, 1998.
6. Kiranagi B., Guru D., Gudivada V. Unsupervised Feature Selection Scheme for Clustering of Symbolic Data Using the Multivalued Type Similarity Measure, *SIAM International Conference on Data Mining*, 2006.
7. Kittler, J. Feature Selection and Extraction. In: *Handbook of Pattern Recognition and Image Processing,* Young, T.Y. and Fu, K.S. Edition, Academic Press, Inc., 1986, pp. 60-81.
8. Lin, T.-S., Meador, J., “Statistical Feature Extraction and Selection for IC Test Pattern Analysis”, Circuits and systems, 1992, pp.391-394 vol 1.
9. Lisboa, P.J.G., Mehri-Dehnavi, R., “Sensitivity Methods for Variable Selection Using the MLP”, International Workshop on Neural Networks for Identification, Control, Robotics and Signal/Image, 1996, pp. 330-338.