Using a Data Metric for Preprocessing Advice for Data Mining Applications

Robert Engels and Christiane Theusinger[1]

Abstract. This paper describes research that is performed in the course of a project where a methodology for providing user support for KDD processes plays a central role. Although methodologically we aim at supporting the whole process of applying inductive learning techniques, the current paper focussus on a part of this process. The main issue in this paper is the support of data preprocessing for KDD. We give some insights in the metadata we calculate from a dataset as part of the method for user support. DCT (Data Characteristion Tool) is implemented in a software environment (Clementine). Some examples are given that resulted from running the UGM/DCT (User Guidance Module combined with DCT) on the data.

1 Introduction and Motivation

As a result of its increasing popularity, an increasing number of machine learning applications were implemented. The more such applications were discussed the more it became clear that applying inductive algorithms is not as trivial as it sometimes might look. The setting in which algorithms are immediately tested on (clean) datasets clearly is too academic and not realistic in real world applications (a finding that is also repeated at workshops on this topic and in literature, see e.g. [4], [11]). Several experiences show that up to three quarters of the time might be used for transforming the data at hand in a format appropriate for learning and that this process has significant influence on the final generated models. It was only natural that from such experiences and the fact that many algorithms are in principle applicable in various situations, the idea arose to partially automate the support of algorithm selection. Such an approach can also be found in statistics, as described in [6] and projects like StatLOG project [9], and the MLT-approach [1] where the CONSULTANT [2] played an advisory role. A very important contribution of that research is that it notifies the importance for a support in defining application processes where statistical and inductive techniques are involved. An issue related to that of algorithm selection is that of data preprocessing. But at the same time we noticed that selecting a final appropriate algorithm for the problem at hand depends more on the effort one wants to put in preprocessing the data than anything else. In the next sections we discuss how knowledge about data can influence user support for machine learning, shortly describe which characteristics are used, and give some examples of preprocessing support.

[1] Institute AIFB, University of Karlsruhe, D-76128, Karlsruhe, Germany, Email: {engels, chth}@aifb.uni-karlsruhe.de

ECAI 98. *13th European Conference on Artificial Intelligence*
Edited by Henri Prade
Published in 1998 by John Wiley & Sons, Ltd.

2 Data Preprocessing for Data Mining

Cases where data is used for Data Mining directly without any kind of preprocessing are so rare (we are not aware of such cases) that data preprocessing seems to constitute an obligatory step. This step is at the same time one of the more time consuming steps, and changes made to a dataset during preprocessing can bring a solution to a KDD problem nearer or just further away. Support for data preprocessing is provided as part of a more general architecture for user support that is described in several papers (UGM approach: [3], [5]). As part of the project, the UGM approach is implemented on top of Clementine (ISL) and has the advantage that it supports generation of KDD processes besides documention and reuse aspects of KDD projects. The approach helps a user defining his problem, after which a top down selection of possibly applicable learning algorithm groups is made. According to the knowledge about the preconditions of these algorithms, an analysis of the available data is made. Based on this knowledge (problem as well as data characteristics are known), preprocessing advice is defined. The current implementation aims at analysis and implementation of techniques for division of multidimensional data spaces according to the distribution of a concept in such a space as done for classification and prediction tasks. Data preprocessing mainly influences data in one of three directions:

- Data cleansing (treatment of noise, extreme values, redundancy, etc.)
- Altering the dimensionality of the data (by attribute generation, filtering, transformation, etc.).
- Altering the data quantity (by selecting, sampling, balancing the available data records).

Where the former is not supported (we presume a cleansed data warehouse to be available), several techniques that support preprocessing in the last two directions are implemented. Thereby we aim at providing as much support as possible, although one should realize that being complete is hardly possible.

3 Measures for Characterisation of Data Sets

This section describes the techniques that are used for characterisation of datasets, where some measures are independent of our focus on statistical classification. Besides statistical measures the DCT approach also includes information theoretical measures for dealing with discrete attributes such as:

- Attribute entropy, describing the information contained in a certain (discrete) variable,
- Class entropy, describing information contained in the dependent variable,
- Joint entropy, relates the depending variable to one of the indepent variables. This measure is usually meant to get an insight in the relative importance of discrete attributes in classification problems.
- Several heuristics that we calculated in order to estimate the degree of agreement among the measures. We took a relevance measure, information gain, the Gini index and the g-function for our experiments.

However, due to space restrictions in this section we will concentrate on statistical measures.

3.1 Standard Characteristics

Some characteristics of data can be seen as standard, and are easily calculated from a dataset and are relevant for either

q :	Number of classes
n :	Number of examples
n_i :	Number of examples in class i
p :	Number of variables
num :	Number of variables with numeric data type
sym :	Number of variables with symbolic data type

Using these characteristics one can get a first indication of the complexity of the problem by analysing these indicators. When a large number of variables is symbolic one would rather concentrate on information theoretical characteristics (as in [9]). We concentrate on the opposite case, in which many independent variables are numeric. The next measurements that are calculated are *location* and *dispersion* parameters that are only computable as single dimensional measurements. Location parameters are measurements like *minimum, maximum, arithmetic middle, median,* α*-trimmed mean* and *empirical quantiles*, where dispersion parameters provide measurements that indicate the dispersion of values the variable can have. Note that location parameters can be divided in two classes, those who can deal with extreme values and those that are sensitive to them. In Data Mining it really depends on the problem definition whether one wants to take one group or the other. This is easily seen, there is for example no need for calculation of robust measures where the goal is to find interesting and unexpected values in the data. Such measures that are resistant against extreme values might just cover up interesting knowledge.

Dispersion parameters that are calculated include *standard deviation (sensitive to extreme values), quartiles deviation (less sensitive to extreme values)* and *median deviation*.

3.2 Assumption Testing

An appropriate technique for prediction of complexity of a domain and relevance of independent variables is discriminant analysis (see [8] for a good introduction). Discriminant Analysis has as one of its purposes to provide one or more mathematical equations in order to separate a data space. This set of equations can be used in order to classify as well. Since we also deal with classification tasks in quantitative domains, our idea is that when performing a classification task on numerical data in Data Mining measurements from discriminant analysis might be used in order to provide indicators that point to certain preprocessing steps for the data. As basis for DCT, the assumptions of discriminant analysis are taken:

1. Linear independence of the discriminating variables
2. Multivariate normal distribution of the discriminating variables
3. Equal covariance matrices for all classes

Now these assumptions can be tested, whereas some tests deliver more useful measurements as strictly necessary. A useful technique that we used here is the Multiple Correlation coefficient.

Definition 3.1 (Multiple Correlation coefficient) *The Multiple Correlation coefficient* $\bar{R}_{1,2,\ldots,p}$ *between* X_1 *and variables* $X_2,\ldots,X_p$ *is the maximal correlation between* X_1 *and some linear function* $\boldsymbol{\alpha}'\mathbf{X_2}$ *of* $X_2,\ldots,X_p$ *where* $\boldsymbol{\alpha} \in I\!R^{p-1}$.

$$\bar{R}_{1,2,\ldots,p} \quad := \quad \sqrt{\frac{\boldsymbol{\sigma}_{12}'\Sigma_{22}^{-1}\boldsymbol{\sigma}_{12}}{\sigma_{11}}}$$

Let $\boldsymbol{X_1},\ldots,\boldsymbol{X_n} \in I\!R^p$ *be* n *independent observation of the randomvector* $\boldsymbol{X}$. *The Sample Multiple Correlation Coefficient between variable 1 and the other* $p-1$ *variables is defined as*

$$R \quad := \quad \sqrt{\frac{\mathbf{s}_{12}'S_{22}^{-1}\mathbf{s}_{12}}{s_{11}}} \;,$$

where $S = \frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{X}_i-\bar{\boldsymbol{X}})(\boldsymbol{X}_i-\bar{\boldsymbol{X}})'$, $S = \begin{pmatrix} s_{11} & \mathbf{s}_{12}' \\ \mathbf{s}_{12} & S_{22} \end{pmatrix}$ *and* $s_{11} \in I\!R$ *and* $\mathbf{s}_{12} \in I\!R^p$, *is the sample covariance matrix.*

This test enables us to provide the user with a dataset that is non redundant. Depending on which variables are defined as important to the user (a user might manually include variables in the UGM), the set of variables is then shrinked. The lost variables are reconstructible in case of total correlation, while the complexity of the inductive task decreases.

3.2.1 Test on Multivariate Normal distribution

Linear discrimination has as assumption that the independent variables show a multivariate normal distribution.

For testing on normal distribution, usually simple measures like kurtosis and skewness are taken. However, both these measures are not robust and distributions exists that, although they do not show a normal distribution, are incorrectly recognised as having one. Another problem is that in the above mentioned case kurtosis and skewness tests do not deliver information on why the assumption of normal distribution is rejected. A robust test on normal distribution is provided by the BHEP-test,[2] and is for this purpose implemented

[2] See [7] for a thorough description of the BHEP-test.

in our system. This BHEP test has a few test properties that make it more suitable on real world datasets as the other discussed alternative: affine invariance, consistency against every non-normal distribution, applicable on datasets of every size and dimension.

In order to test for homogeneity among the independent variables we use the Boxian M-statistic, while this statistic seems to deliver a good estimation of homogeneity. M-statistics basically tests the hypothesis that all classes posses the same covariance structure and has an asymptotical chi-square distribution. From M-statistics the SD-ratio is calculated, which basically means that a transformation is performed which delivers a real that is equal to unity when all individual covariance matrices are equal.

4 Estimating Space Complexity for Classificational Problems

In Discriminant Analysis several covariance matrices are calculated in order to be able to determine 'eigenvalues' and eigenvectors. The covariance matrices that we calculate are not different from the ones known from ordinary statistics, we will therefore only mention which matrices we calculate. First of all we calculate the covariance matrix of each class i and the pooled covariance matrix (over all classes) Furthermore we calculate the within-groups sums of squares and cross-products matrix (W). Matrix (T) recollects the total sums of squares and cross-products matrix. From these both we retrieve **B=T-W**. Here matrix (B) represents the between groups sums of squares and cross-products matrix. This gives us an estimation of how different the several groups really are.

Covariance matrices are used to calculate eigenvalues and eigenvectors of matrix $W^{-1}B$. The motivation to use $W^{-1}B$ stems from the wish to maximize the ratio between the elements of matrix B to the elements of matrix W. Since our restriction is that we aim at applying the chosen techniques on real world datasets, we have chosen the so called QR-algorithm [10], since this algorithm is resistant against non symmetric covariance matrices, whereas other algorithms (like the JACOBI algorithm) are only able to calculate valid eigenvalue iff the covariance matrices are symmetric. Since $W^{-1}B$ is positive definite with order $r = \min(q-1, p)$, $\lambda_{max} = \lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_r$ positive eigenvalues with the belonging eigenvectors $b_{max} = b_1, \ldots, b_r$ can be defined.
The larger such an eigenvalue, the larger is the role the belonging discriminant function plays in the final classification. The number of significantly positive eigenvalues gives us a first insight in how many dimensions might be needed in order to be able to distinct the classes from one another. We also provide the relative importance of each eigenvalue according to the Total Discrimination Power (TDP). The relative importance of the most important eigenvalue (an indication for the importance of the 1st discriminant function) is also provided.

Next we provide the canonical correlation, which is a value that (if close to unity) tells us that there is a strong connection between the classes and the 1st discriminant function. This function is defined as:

$$FirstCanonicalCorrelation = \sqrt{\frac{\lambda_{max}}{1+\lambda_{max}}}.$$

Note that although the 1st discriminant function might be the strongest (as seen by the most important eigenvalue), the canonical correlation might show that the relation to the classes is only a weak one. In this case we are not likely to be able to make a good prediction with the current dataset when using linear discriminating methods. The next measure we calculate is Wilks Lambda (also known as U-statistic). Wilks Lambda is a multivariate measure of group differences (over the independent variables). It is defined as follows:

$$\Lambda = \Pi_{j=1}^{r}(1+\lambda_j)^{-1}$$

where λ_j are the eigenvalues that are greater then zero mentioned before and r is the number of positive eigenvalues. Wilks Λ is used as follows: if the value of Wilks Λ is near 1, the groups centres are identical, and while there are obviously no group differences, the discrimination is bad. On the other hand, given a value for Wilks Λ near zero, the group centres are really distinctive, and a good discrimination can be found.

Now we are interested in a measure that can support us in deciding upon the number of significant discriminant functions that should be considered. Bartlett's statistic is calculated, which is applicable in the case that the assumption of multivariate normal distribution holds. Bartlett's statistic has an asymptotic $\chi^2_{p(q-1)}$-distribution and is defined as:

$$V = \{(n-1) - \frac{1}{2}(p+q)\} \sum_{j=1}^{r} \ln(1+\lambda_j)$$

Iff V is larger than the critical value of this χ^2-distribution with $p(q-1)$ degrees of freedom, then all r discriminant functions can be treated as relevant. This statistic delivers the discriminant functions that are relevant by repeatedly retracting the most important discriminant function from V, after which this function is deleted from the ordered list with significances of the discriminant functions. The relevance of a single discriminant function j is calculated through:

$$V_j = \{(n-1) - \frac{1}{2}(p+q)\} \ln(1+\lambda_j)$$

with an asymptotic χ^2_{p+q-2j}-distribution.

Based on the statistics mentioned above, several kinds of support could be defined. For example, there is the possibility to make statements about the kind of algorithms that one wants to use. In the case that the Canonical Correlation and the relative proportion of an eigenvalue to the Total Discriminating Power is close to 1, algorithms that are able to describe linear relations should be favoured, whereas in case the relative proportion of an eigenvalue is large where the canonical correlation is near zero (no real relation to the classes), then one would prefer to look at the information gain measures, since algorithms that are able to describe linear relations are less applicable. At the same time, the calculated measures provides insight in the complexity of a problem, which might either give rise to certain data preprocessing steps in order to allow a better depiction of the complex space on a lower dimensionality.

5 Examples of Advice for Preprocessing based on DCT

We took UCI-datasets in order to demonstrate our hypothesis that the data characteristics as we took them are suf-

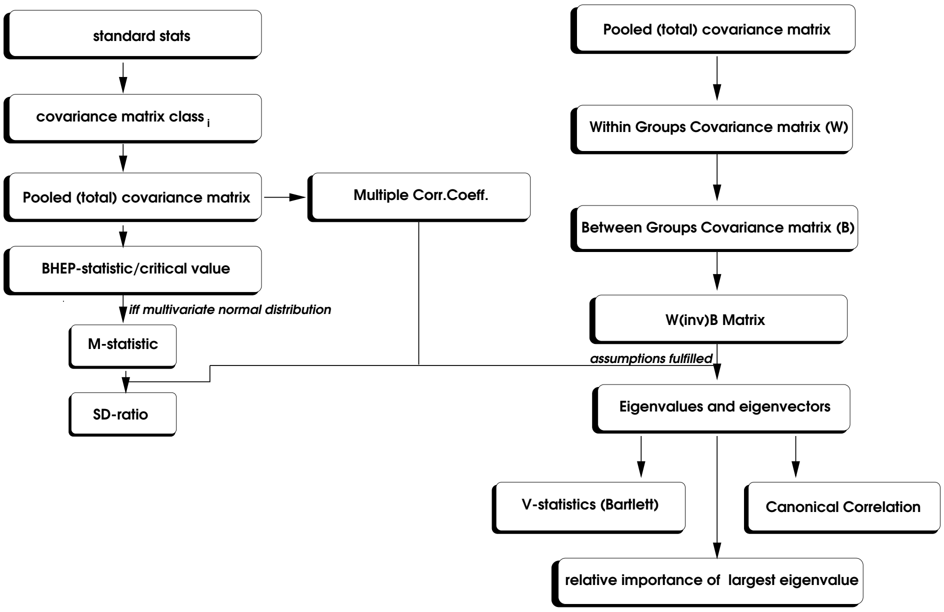

Figure 1. Used measures and their relations

	Range:	Median	Mean	α-trimmed Mean	Std. Dev.	Quartil
vegde-sd:	991,719	0,833	5,709	1,25	44,837	1,451
hegde-sd:	1386,33	0,963	8,244	1,683	58,799	1,763
⋮	⋮	⋮	⋮	⋮	⋮	⋮

Table 1. Values for several of the "standard statistics" of two independent variables (Segment)

ficient for definition of certain preprocessing steps. We will deal with some of the calculated measurements w.r.t. these datasets and show that there is a possibility to make statements about which preprocessing is appropriate, or which kind of algorithms can be used. For every dataset we also made a testrun using several data mining techniques, so we could verify whether our assumptions about the interpretation of the characteristics was right or not.

5.1 Example: Segment

The used dataset originates from 7 outdoor images that where divided in 3x3 regions and classified according to their membership of one of 7 classes. Each region forms an instance that is described by 19 numerical variables. Three of these 19 variables are eliminated while containing constant variables which do not contain any information that might contribute. Table 1 shows the generated output for the variables vegde-sd and hegde-sd. The difference between the arithmetical mean and the α-trimmed mean[3] we see that these variables might suffer from extreme values quite a lot. This view is supported by the range, standard deviation and quartile distance as dispersion measures. The information that is gathered in the last three measures is based on a different quantity of extreme values that is neglected which gives us feedback on the number of such extremes that is present in a certain variable.

The output for the Multiple Correlation Coefficients (MCC) in table 2 shows that at least 7 variables are perfectly correlated. A correlation of 100 percent is also known as a functional dependency and is not contributing any information to the dataset, except for redundancy. Based on the results of DCT we therefore eliminated 6 of these 7 variables from our dataset. We also tested our DCT tool on the resulting database after preprocessing and indeed found that the number of discriminant functions needed for classification remained constant whereas the dimensionality of the input data was reduced with 6 dimensions. Post evaluation of training a neural net delivers a less complex net with fewer input neurons as well as fewer neurons in the hidden layer, while the classification accuracy stays the same and training time decreases dramatically. When evaluating the other measures (see table 3) we find that the Canonical Correlation remains high (0.98) after eliminating these 6 variables, indicating that the correlation between the groups and the discriminant functions is high. Also interesting is that a relative large proportion of the total discrimination power is explained by the first discriminant function. Since Wilks Lambda as close to zero (0.000616) we can conclude that there is a high discrimination power available in the dataset, which tells us that the group centroids are very different. This is also found in the

[3] The α-trimmed mean basically cuts the $2^*\alpha$ percent extreme values of a variable away. α is set to 5 percent in our example.

Var:	reg.col	reg.row	vdg-mn	vgd-sd	hdg-mn	hdg-sd	intsy-mn	rwred-mn
MCC:	0,185	0,717	0,726	0,794	0,760	0,809	1,000	1,000
Var:	**rwblue-mn**	**rwgrn-mn**	**xred-mn**	**xblue-mn**	**xgrn-mn**	**value-mn**	**sat.-mn**	**hue-mn**
MCC:	1,000	1,000	1,000	1,000	1,000	1,000	0,774	0,94

Table 2. Multiple Correlation Coefficients for the sixteen variables of the segment dataset

Stats:	BHEP	M-Stat	SDratio	EV/TDP	1stCanCor	Wilks	Nr.SDiscF.
After Preproc:	77,789	50141,72	11,236	0,564	0,976	0,000616	6

Table 3. Results for V-statistic, number of significant eigenvalues (Nr.SDiscF.) and canonical correlation of the segment dataset

experimental results, where classifiers were build with high accuracies (default class: 0.14 percent, classification results between 90 - 95 percent, depending on the algorithms used).

5.2 Example: Post-operative

When considering the dataset Post-operative (classification of patients in one of three groups: transfer to Intensive Care, Normal Care or send them home), one sees that:

- the class entropy H(C) is 0.98, whereas
- the mean information gain ($\bar{I}_{gain}$) is 0.018.
- Gini-index shows a valuerange of [0 , 0.021] and
- g-function has a range of [6.97^{-32}, 1.78^{-30}]

These characteristics tell us that there is significantly less information available in the attributeset ($\bar{I}_{gain}$) as is required by the class entropy (H(C)). The class entropy provides a measure for the number of binary decisions that is necessary to be able to differentiate between the classes. The low values of the Gini-index show that there are no attributes that really contribute to the classification. As a last indicator one can interpret the near zero values of the g-function as an indication for the fact that there is no high probability to find the joint distribution of the classes and attributes.

6 Conclusions

Beginning with ideas from the knowledge acquisition field and statistical principles and guided by projects as MLT [1] and StatLOG [9] we finally came up with a framework for User Guidance Modelling (UGM: [3], [5]) that recollects a top down (problem definition etc.) and bottom up (data characteristics) process. The results of DCT are used for initiation or support of preprocessing of the data. Several techniques from the field of statistics as well as information theory are implemented in our UGM prototype. We are aware of the fact that due to our focus on statistical techniques we neglected situations in which combinations of continuous and discrete data are found (as in most real world settings). Therefore, in the future we have to focus on the relationship between information theoretical approaches and the approach described here. The examples show us that indeed interesting and helpful preprocessing can be defined that is based on the several test statistics that are calculated. Other research at our institute aims at using DCT for algorithm selection and will be integrated in the UGM framework.

Acknowledgements We thank Norbert Henze for helping us understand the BHEP-test statistic. DCT's first version is implemented based on Gamma's and Brazdil's analyst tool (StatLOG project). Statistical and Information Theoretical techniques mentioned in this paper where added. Thanx to our colleagues Guido Lindner, Ulrike Zintz and Rudi Studer for support and discussion.

REFERENCES

[1] MLT Consortium, 'Final public report', Technical report, (1993). Esprit II Project 2154.

[2] S. Craw, D. Sleeman, N. Granger, M. Rissakis, and S. Sharma, 'Consultant: Providing advice for the machine learning toolbox', in *Research and Development in Expert Systems*, eds., M.A. Bramer and R.W. Milne, pp. 5–23, (1992).

[3] R. Engels, 'Planning tasks for knowledge discovery in databases; performing task-oriented user-guidance.', in *Proceedings of the 2nd Int. Conference on Knowledge Discovery and Data Mining*, eds., E. Simounis, J. Han, and U. Fayyad, pp. 170–175, Portland, Oregon, August 2-4, (1996). AAAI-Press.

[4] R. Engels, B. Evans, J. Herrmann, and F. Verdenius, eds. *Workshop on Machine Learning Application in the real world; Methodological Aspects and Implications (at the ICML-97)*, Nashville, TN, July 12th, 1997. at: 14th International Conference on Machine Learning.

[5] R. Engels, G. Lindner, and R. Studer, 'A guided tour through the data mining jungle', in *Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining*, ed., D.Pregibon D. Heckerman, H.Manilla, Newport Beach, CA, August 14 -17, (1997). AAAI Press, Menlo Park, CA.

[6] D.J. Hand, 'Deconstructing statistical questions', *Journal of the Royal Statistical Society*, 317–356, (1994).

[7] N. Henze and Th. Wagner, 'A new approach to the bhep tests for multivariate normality', *Journal of Multivariate Analysis*, **62**(1), (1997).

[8] W.R. Klecka, 'Discriminant analysis', *Sage University Paper Series on Quantitive Applications in the Social Sciences*, **07-019**, (1980). Beverly Hills and London: Sage Pubns.

[9] D. Michie, D.J. Spiegelhalter, and C.C. Taylor, *Machine Learning, Neural and Statistical Classification*, Ellis Horwood, 1994.

[10] J. Stoer and R. Bulirsch, *Numerische Mathematik 2*, Springer Verlag, 3 edn., 1990.

[11] F. Verdenius, 'Applications of inductive learning techniques: A survey in the netherlands', *AI communications*, **10**(1), (1997).