Uniwersytet Ekonomiczny w Katowicach

Studia Ekonomiczne. Zeszyty Naukowe
Uniwersytetu Ekonomicznego w Katowicach
ISSN 2083-8611 Nr 288 · 2016

Informatyka i Ekonometria 5

Katarzyna Stąpor
Silesian University of Technology
Institute of Computer Science Institute of Computer Science
Faculty of Automatic Control, Electronics and Computer Science
katarzyna.stapor@polsl.pl

A CRITICAL COMPARISON OF DISCRIMINANT ANALYSIS AND SVM-BASED APPROACHES TO CREDIT SCORING

Summary: Credit scoring models are the basis for financial institutions like retail and consumer credit banks. The purpose of these models is to evaluate the likelihood of credit applicants defaulting in order to decide whether to grant them credit. The paper compares two methodologies for building credit scoring models: heteroscedastic discriminant analysis-based with the support vector machines. The real-world credit dataset is used for comparison.

Keywords: discriminant analysis, support vector machines, credit scoring model.

Introduction

The phenomenon of borrowing and lending has a long history associated with human behaviour. Credit is perhaps a phenomenon as old as trade and commerce. Despite the very long history of credit, the history of credit scoring is very short, beginning only about six decades ago. Information collected by financial institutions of a credit applicant is used to develop a numerical score for each applicant [Thomas, 2000]. Credit scoring is applied at the point of application for a loan to predict the risk of default (nonpayment) and to make the decision whether to approve, that application for credit.

The set of decision models and their underlying methods, that serve lenders in granting consumer credits by assessing the risk of lending to different consumers are called “credit scoring models”. Credit scoring models are the basis for financial institutions like retail and consumer credit banks. The purpose of these

models is to evaluate the likelihood of credit applicants defaulting in order to decide whether to grant them credit [Matuszczyk, 2012]. Credit scoring is although an important area of research, that enables financial institutions to develop lending strategies to optimize profit. The use of credit scoring models is now a key component in retail banking.

A range of different statistical as well as data mining techniques [e.g. Stąpor, 2011] have been used in building credit scoring models. Discriminant analysis (AD), linear regression, logistic regression, neural networks, k-nearest neighbors, support vector machines (SVM) and classification trees cover the range of different surveys on credit scoring models (for an overview see: [Thomas, 2000; Crook et al., 2007]). Most of these techniques are applicable to build an efficient and effective credit scoring system, that can be effectively used for predictive purposes.

Advanced statistical techniques, such as support vector machines and neural networks provide an alternative to conventional statistical techniques, such as discriminant analysis, probit analysis and linear or logistic regression. The point of using sophisticated techniques, is their capability of modelling extremely complex functions, and, of course, this stands in contrast to traditional linear techniques, such as, linear regression and linear discriminant analysis.

Several papers have recently been published assessing the performance of SVM for credit scoring. They report, that SVM perform slightly better in comparison with other algorithms, buy not significantly so.

The purpose of this paper is to compare the performance of discriminant analysis based methods for building credit scoring models (i.e. classical) against those based on the support vector machines. As we have only the one real world credit dataset, the German dataset – this file will be used in our comparisons. For the comparison, we have selected the algorithms with the best prediction accuracies (according to the literature). Discriminant analysis-based approach is represented here by the paper K. Stąpor et al. [2016], while support vector machines with the paper F. Chen and F. Li [2010]. For the reasons discussed later, the last algorithm was re-implemented and tested on the transformed German credit dataset.

This paper is organized as follows. Section 2 and 3 give the short description of the approaches being compared: discriminant analysis and support vector machines, while in the section 4 the credit dataset used in the comparison is described (it's detailed structure is given in the Appendix). Section 5 deals with the comparison and is followed by conclusions in section 6.

1. Discriminant analysis

Fisher Discriminant Analysis (FDA) [Fisher, 1936; Krzyśko, 1990] is a multivariate technique to classify study instances into groups and/or describe group differences. Discriminant analysis is widely used in many areas such as biomedical studies, banking environment (for credit evaluation), financial management, bankruptcy prediction, marketing and many others.

There are many formulations of FDA, a typical one for pattern recognition community is given below.

FDA is concerned with the search for a linear transformation, that reduces the dimension of a given *n*-dimensional statistical model to *d* ($d < n$) dimensions, while maximally preserving the discriminatory information for the several classes within the model. It determines a linear mapping A, a $d \times n$ matrix A, that maximizes the so-called Fisher criterion J_F:

$$J_F(A) = tr\left(\left(AS_W A^T\right)^{-1}\left(AS_B A^T\right)\right)$$

Here, $S_B = \sum_{i=1}^{c} \frac{n_i}{n}\left(m_i - \overline{m}\right)\left(m_i - \overline{m}\right)^T$ and $S_W = \sum_{i=1}^{c} \frac{n_i}{n} S_i$ are the between-class and the average within-class scatter matrices, respectively; c is the number of classes, m_i is the mean vector of class i, n_i is a number of samples in class i, $n = \sum_{i=1}^{c} n_i$, and the estimated overall mean equals $\overline{m} = \sum_{i=1}^{c} \frac{n_i}{n} m_i$, $S_i = \frac{1}{n_i}\sum_{j=1}^{n_i}\left(X_{ij} - m_i\right)\left(X_{ij} - m_i\right)^T$ is the within-class covariance matrix of class i.

Optimizing (1) comes down to determining an eigenvalue decomposition of $S_W^{-1} S_B$, and taking the rows of A equal to d eigenvectors corresponding to d largest eigenvalues.

The most important assumption of discriminant analysis is the *homogeneity* of variance/covariance matrices (*homoscedasticity*). Moreover it can be applied only to data measured at least on the ordinal measurement scale or higher.

For the two-class case we have:

$$S_B = (m_1 - m_2)(m_1 - m_2)^T \text{ and } S_W = p_1 S_1 + p_2 S_2,\ p_2 = 1 - p_1$$

where $p_i = n_i / n$, $i = 1, 2$. A limitation of FDA is that it merely tries to separate class means as good as possible and it does not take the discriminatory informa-

tion, that is present in the difference of the covariance matrices into account. It is incapable of dealing explicitly with heteroscedastic data, i.e. data in which classes do not have equal covariance matrices.

The two most important extensions of FDA have been given by M. Krzyśko, W. Wołyński [1996] and M. Loog, R. Duin [2002].

The heteroscedastic extension in [Loog, Duin, 2002] and the one which we adopted, is based on the notion of *Distance Directed Matrices* (DDM), which capture not only the difference in means between two classes, but describe, in a certain way, their difference in covariance as well. They proposed DDM based on the *Chernoff distance* between two probability density functions d_1, d_2:

$$\partial_C = -\log \int d_1^{\alpha}(x) d_2^{1-\alpha}(x) dx$$

where $\alpha \in (0,1)$.

For two normally distributed densities, the DDM is a positive semi-definite matrix S_C:

$$S_C = S^{-\frac{1}{2}}(m_1 - m_2)(m_1 - m_2)^T S^{-\frac{1}{2}} + \frac{1}{p_1 p_2}(\log S - p_1 \log S_1 - p_2 \log S_2)$$

where $\alpha = p_1$, $S = p_1 S_1 + p_2 S_2$. The trace of S_C is the *Chernoff distance* ∂_C between those two densities. Determining transformation A by an eigenvalue decomposition of S_C, means that we determine a transform, which preserves as much of the *Chernoff distance* in the lower dimensional space as possible. The heteroscedastic two-class Chernoff criterion J_C is defined as:

$$J_C(A) = tr\left(\left(AS_W A^T\right)^{-1} A(m_1 - m_2)(m_1 - m_2)^T A^T \right.$$
$$\left. - AS_W^{\frac{1}{2}} \frac{p_1 \log\left(S_W^{-\frac{1}{2}} S_1 S_W^{-\frac{1}{2}}\right) + p_2 \log\left(S_W^{-\frac{1}{2}} S_2 S_W^{-\frac{1}{2}}\right)}{p_1 p_2} S_W^{\frac{1}{2}} A^T \right)$$

This is maximized by determining an eigenvalue decomposition of:

$$S_W^{-1}\left(S_B - S_W^{\frac{1}{2}} \frac{p_1 \log\left(S_W^{-\frac{1}{2}} S_1 S_W^{-\frac{1}{2}} \right) + p_2 \log\left(S_W^{-\frac{1}{2}} S_2 S_W^{-\frac{1}{2}} \right)}{p_1 p_2} S_W^{\frac{1}{2}} \right)$$

and taking the rows of the transform A equal to d eigenvectors (called here "discriminant directions") corresponding to the d largest eigenvalues.

Another interesting approach to heteroscedastic linear discriminant analysis can be found in [Krzyśko, Wołyński, 1996], where authors proposed the optimal classification rules based on linear functions, which maximize probabilistic distances: the Chernoff or the Morisita or the Kullback-Leibler ones.

2. SVM classifier

SVM classifier [Vapnik, 1995] separates training examples from two classes by a hyperplane, such that the margin width between the hyperplane and the examples is maximized. In the case of nonlinear separability, training examples are allowed to be on the wrong side of a margin, but they are assigned a penalty proportional to how far they are on the wrong side. The sum of penalties is minimized, while maximizing the margin width. A parameter C controls the relative cost of each goal in the overall optimization process.

The SVM optimization problem can be expressed algebraically as a dual form quadratic programming problem. Let $S = \{(x_i, y_i), 1 \le i \le n\}$ where $y_i \in \{1,-1\}$ be a training set. The optimization problem is:

$$\max_{\alpha}\left(\sum_{i=1}^{n} \alpha_i + \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j k(x_i x_j) \right)$$

subject to the constraints:

$$0 \le \alpha_i \le C,\ i = 1,...,n \quad \sum_{i=1}^{n} y_i \alpha_i = 0$$

where α_i is a Lagrange multiplier for each training example i. The kernel function k can be used to implement non-linear models of the data. We consider here Gaussian kernel:

$$k(x_i, x_j) = \exp(-\frac{\left\| x_i - x_j \right\|^2}{2\sigma^2})$$

where σ is a kernel parameter specified by a user.

Finally, the decision function of classifying a new data point *x* can be written as follows:

$$f(x) = \text{sgn}(\sum_{i=1}^{n} \alpha_i y_i k(x_i, x) + b)$$

where b is a threshold term computed as:

$$b = \sum_{i=1}^{n} \alpha_i y_i k(x_i, x_j) \text{ for any } j \in \{1, ..., n\} \text{ such that } 0 < \alpha_i < C$$

Training examples are called support vectors, if they are on the margin or are on the wrong side of the margin.

3. Real world credit data set

The presented comparison is based on the prediction accuracies coming from the experiments conducted with the real world credit data set, the German credit dataset, available from the UCI Repository of Machine Learning Databases [Murphy, 1994]. The German dataset consist of 700 instances of creditworthy borrowers and 300 of bad borrowers. It is composed of 20 numeric as well as nominal attributes containing information about credit duration, history, purpose, amount, savings, age, job and other personal information (see Appendix 1).

4. Comparing algorithms

To compare discriminant analysis and support vector machines based approaches to building credit scoring models we have selected two methodologies with the best prediction results.

The first, described in [Stąpor et al., 2016], is based on the heteroscedastic discriminant analysis combined with feature selection. We have proved in that publication, that using heteroscedastic extension of the classical linear Fisher

Discriminant Analysis results in a better prediction accuracy than in the previous studies. The other reason for such a choice is that all other discriminant analysis-based approaches [Crook et al., 2007] are based on the classical Fisher linear discriminant analysis in which the major assumption is the homogeneity of variance/covariance matrices (homoscedasticity). None of the described in the literature discriminant analysis-based approaches [Crook et al., 2007] has checked the fulfillment of this important assumption. In the case of German credit dataset this assumption is violated. We performed permutation statistical test implemented based on the method described in [Zhu et al., 2002].

The second approach we have chosen, based on the support vector machines, is represented by the article [Chen, Li, 2010] with the best prediction accuracy obtained so far on the German credit dataset. Unfortunatelly, there is methodological error in the publication – the authors made an unallowed transformation from the nominal to the ratio measurement scales on the German credit dataset – some attributes are qualitative, not numeric [see: Appendix 1]. This strengthening of the nominal scale adds a new information, which is methodologically not allowed.

Thus, to be able to compare the results, we re-implemented the algorithm from F. Chen and F. Li [2010] and tested it on the German credit dataset converted as described below.

According to the recommendations from M. Walesiak [2003], the conversion from the weaker to the stronger measurement scale, for example from the nominal to the ratio one is methodologically unallowed strengthening of the nominal scale, because one cannot have more information from the less amount.

Thus, we propose the following transformation, namely, all the nominal features were transformed to the "binary" features, each one representing one of its possible states/labels (1/0value if the object is/is not assigned a given label). After such conversion, each attribute is transformed to as many new attributes as there are different labels/states. Preprocessed German dataset contained 59 attributes.

Heteroscedastic discriminant analysis based approach

The whole methodology for building the credit scoring model based on the heteroscedastic extension of Fisher linear discriminant analysis was described by K. Stąpor et al. [2016] and will not be described here. The best result using this proposed methodology was achieved by using the combination of filter-based feature selection based on *F-score* with feature extraction by heteroscedastic discriminant analysis. The *F-score* [Duda et al., 2001] is defined as:

$$F\text{-}score = \frac{|S_B|}{|S_W|}$$

where |.| is a determinant. The larger the *F-score* is, the more likely this feature is more discriminative. *F-score* was calculated for each feature and they were then sorted in increasing order.

The proposed model was able to achieve 75.10% ± 3.38% accuracy rate with 18 features selected and 3 discriminant directions (as described in section 2) [Stąpor et al., 2016].

SVM based approach

For the reasons described above, we have re-implemented the algorithm for building the credit scoring model based on SVM classifier with nonlinear kernel (Gaussian) proposed by F. Chen anf F. Li [2010].

This methodology is a based on setting two SVM parameters using grid-search and selecting input features using F-score. In the grid-search approach, pairs of (C,σ) are tried and the one with the best cross-validation accuracy is chosen. After identifying a "better" region on the grid, a finer grid search on that region can be conducted. To get good generalization ability, grid search approach uses a validation process to decide parameters. That is, for each of the *l* subsets D_i (*i=1,...,l*) of the data set *D*, create a training set $T_l = D \setminus D_l$, then run a cross-validation process. Overall accuracy is averaged across all *k* partitions. These *l* accuracy values also give an estimate of the accuracy variance of the algorithm. Using 5-fold cross-validation in the iterative procedure described by F. Chen and F. Li [2010], only the best first *f* features were passed to the finite model.

Using this methodology on the transformed German dataset, the classification accuracy achieved 76.10% ± 6.10% and the average number of selected features was 20.

According to the above described results – SVM performs slightly better, although the method based on the heteroscedastic discriminant analysis gives the model with less number of features. Bearing in mind the standard error of prediction rule – the classification accuracies of the two approaches are almost identical.

Moreover, using nonlinear SVM classifier (with Gaussian kernel), makes the learning process more complex – one should estimate the parameters of the SVM classifier in the separate validation procedure, the grid search, which requires the additional dataset and is computationally intensive process. Additionally, such complex learning procedure is prone to overtraining. We found

that, unlike many other learning tasks, a large number of support vectors are required to achieve the best performance. This is due to the nature of the credit data for which the available application data can only be broadly indicative of default.

Conclusion

SVMs are a relatively new technique for application in credit scoring. We test them on the German credit dataset used in the previous studies. We find, that SVMs are successful in building credit scoring models, but a the cost of intensive learning procedure, which require separate validation set to avoid overtaining. SVMs with non-linear kernel does not give the significant improvement over the simpler models like (properly conducted) discriminant analysis.

This indicates, that the data is broadly linearly separable [Gayler, 2006].

In credit scoring, more important than the goodness of fit to the developmental sample is the anticipation of possible changes in the operational systems and data. Practitioners in credit scoring achieve this aim by biasing their models towards simple models. Such models yield most of the predictive power of more complex models and, which is more important, are more likely to generalize across potential data sets.

However, techniques enabling better generalization to the distribution of possible data sets would be welcome.

References

Chen F., Li F. (2010), *Combination of Feature Selection Approaches with SVM in Credit Scoring*, "Expert Systems with Applications", Vol. 37, s. 4902-4909.

Crook J.N., Edelman D.B., Thomas L.C. (2007), *Recent Developments in Consumer Credit Risk Assessment*, "European Journal of Operational Research", Vol. 183(3), s. 1447-1465.

Duda R., Hart P., Stork D. (2001), *Pattern Classification*, 2 ed., John Wiley & Sons, New York.

Fisher R. (1936), *The Use of Multiple Measurements in Taxonomic Problems*, "Annals of Eugenics", No. 7, s. 179-188.

Gayler R. (2006), *Comment: Classifier Technology and the Illusion of Progress – Credit Soring*, "Statistical Science", Vol. 21(1), s. 19-23.

Krzyśko M. (1990), *Discriminant Analysis*, WNT, Warszawa.

Krzyśko M., Wołyński W. (1996), *Discriminant Rules Based on Distances*, "Tatra Mountains Math. Publ.", No. 7, s. 289-196.

Loog M., Duin R. (2002), *Non-iterative Heteroscedastic Linear Dimension Reduction for Two-class Data: From Fisher to Chernoff*, Proc. 4th Int. Workshop S+SSPR, s. 508-517.

Matuszczyk A. (2012), *Credit Scoring*, CeDeWu.pl, Warszawa.

Murphy P.M., Aha D.W. (1994), *UCI Repository of Machine Learning*, Department of Information and Computer Science, University of California, http://www.ics.uci.edu/-~mlearn/MLRepository.html.

Stąpor K. (2011), *Classification Methods in Computer Vision*, Wydawnictwo Naukowe PWN, Warszawa.

Stąpor K., Smolarczyk T., Fabian P. (2016), *Heteroscedastic Discriminant Analysis Combined with Feature Selection for Credit Scoring*, "Statistics in Transition new series", June.

Thomas L.C. (2000), *A Survey of Credit and Behavioural Scoring: Forecasting Financial Risk of Lending to Consumers*, "International Journal of Forecasting", Vol. 16(2), s. 149-172.

Vapnik V. (1995), *The Nature of Statistical Learning Theory*, Springer Verlag, New York.

Walesiak M. (2003), *Strategies in Statistical Methods in the Case of Variables Measured on Different Scales,* "Operational Research and Decisions", No. 1, s. 71-77.

Zhu L., Ng K., Jing P. (2002), *Resampling Methods for Homogeneity Tests of Covariance Matrices*, "Statistica Sinica", No. 12, s. 769-783.

Appendix 1

The structure of the German credit data set

Attribute	Description	Values
1	Status of existing checking account (qualitative)	A11 :... < 0 DM A12 : 0 <=... < 200 DM A13 :... >= 200 DM /salary assignments for at least 1 year A14 : no checking account
2	Duration in month (numerical)	
3	Credit history (qualitative)	A30 : no credits granted/all credits paid back duly A31 : all credits at this bank paid back duly A32 : existing credits paid back duly until now A33 : delay in paying off in the past A34 : critical account/other credits existing (not at this bank)
4	Purpose (qualitative)	A40 : car (new) A41 : car (used) A42 : furniture/equipment A43 : radio/television A44 : domestic appliances A45 : repairs A46 : education A47 : (vacation – does not exist?) A48 : retraining A49 : business A410 : others
5	Credit amount (numerical)	
6	Savings account/bonds (qualitative)	A61 :... < 100 DM A62 : 100 <=... < 500 DM A63 : 500 <=... < 1000 DM A64 :.. >= 1000 DM A65 : unknown/ no savings account
7	Present employment since (qualitative)	A71 : unemployed A72 :... < 1 year A73 : 1 <=... < 4 years A74 : 4 <=... < 7 years A75 :.. >= 7 years
8	Instalment rate in percentage of disposable income (numerical)	
9	Personal status and sex (qualitative)	A91 : male : divorced/separated A92 : female : divorced/separated/married A93 : male : single A94 : male : married/widowed A95 : female : single
10	Other debtors/guarantors (qualitative)	A101 : none A102 : co-applicant A103 : guarantor
11	Present residence since (numerical)	

12	Property (qualitative)	A121 : real estate A122 : if not A121 : building society savings agreement/life insurance A123 : if not A121/A122 : car or other, not in attribute 6 A124 : unknown/no property
13	Age in years (numerical)	
14	Other instalment plans (qualitative)	A141 : bank A142 : stores A143 : none
15	Housing (qualitative)	A151 : rent A152 : own A153 : for free
16	Number of existing credits at this bank (numerical)	
17	Job (qualitative)	A171 : unemployed/unskilled - non-resident A172 : unskilled – resident A173 : skilled employee/official A174 : management/self-employed/highly qualified employee/officer
18	Number of people being liable to provide maintenance (numerical)	
19	Telephone (qualitative)	A191 : none A192 : yes, registered under the customer's name
20	Foreign worker (qualitative)	A201 : yes A202 : no

PORÓWNANIE ANALIZY DYSKRYMINACYJNEJ I MASZYN WEKTORÓW PODPIERAJĄCYCH W ANALIZIE RYZYKA KREDYTOWEGO

Streszczenie: Modele oceny ryzyka kredytowego stanowią podstawę działalności większości instytucji finansowych, zajmujących się udzielaniem kredytów. Celem takich modeli jest ewaluacja prawdopodobieństwa zaprzestania przez kredytobiorcę spłaty udzielonego mu kredytu. W artykule dokonano porównania dwóch modeli oceny ryzyka kredytowego, które wykorzystują nowe metody statystyczne, a także metody uczenia maszynowego do ich konstrukcji: heteroscedastyczną analizę dyskryminacyjną oraz maszyny wektorów podpierających. Dla dokonania porównania tych metod wykorzystany został ogólnie dostępny, niemiecki zbiór kredytowy.

Słowa kluczowe: analiza dyskryminacyjna, maszyny wektorów podpierających, model oceny ryzyka kredytowego.