Bankruptcy Analysis for Credit Risk using Manifold Learning

B Ribeiro[1], A Vieira[2], J Duarte[2], C Silva[1,3], J Carvalho das Neves[4], Q Liu[5], and A H Sung[5]

[1]CISUC, Department of Informatics Engineering,University of Coimbra, Portugal
[2]Physics Department, Polytechnic Institute of Porto, Portugal
[3]ESTG- Polytechnic Institute of Leiria, Portugal
[4]ISEG - School of Economics, Technical University of Lisbon, Portugal
[5]Computer Science Department, University of New Mexico Tech, USA
bribeiro@dei.uc.pt,asv@isep.pt,jcneves@iseg.utl.pt,sung@cs.nmt.edu

1 Introduction

In this work we apply manifold learning to a real data set of distressed and healthy companies for proper geometric tunning of similarity data points and visualization. While Isomap [1] algorithm is often used in unsupervised learning, our approach combines this algorithm with information of class labels for bankruptcy prediction. We compare prediction results with classifiers such as Support Vector Machines (SVM), Relevance Vector Machines (RVM) and the simple k-Nearest Neighbor (KNN) in the same data set, showing comparable accuracy. Furthermore, the proposed approach is shown to have excellent visualization capabilities, as a result of the incorporation of prior knowledge of a variable (indicating bankruptcy risk) into a dissimilarity matrix.

2 Methods

Supervised nonlinear dimensionality reduction can be used as a preprocessing step before classification. The rationale here is to map the high-dimensional data space into a lower dimensional space where classification methods do not suffer from the curse of dimensionality. As the explicit mapping is not found by the algorithm some learning methodology must be used. Our approach uses the training labels in the data set to provide a better construction of features. We apply the dissimilarity measure (1) [2]

$$D(x_i, x_j) = \begin{cases} ((a-1)/a)^{1/2} & \text{if } c_i = c_j \\ a^{1/2} - d_0 & \text{if } c_i \neq c_j \end{cases} \tag{1}$$

where $a = 1/e^{-d_{ij}^2/\sigma}$ with d_{ij} set to one of the distance measures described above, σ is a smoothing parameter, d_o is a constant ($0 \leq d_0 \leq 1$) and c_i, c_j are the class labels.

3 Results

Testing was carried out with a sample of Diane database of financial statements of French companies. The dataset includes information about 30 financial ratios, of about 60 000 industrial French companies, for the years of 2002 to 2006, with at least 10 employees. Figure 1 shows the visible separation of patterns (from healthy to bankrupt firms) with prior knowledge of the bankruptcy risk variable incorporated into the dissimilarity matrix. In the picture on the right, trustiworthiness measures are also shown.

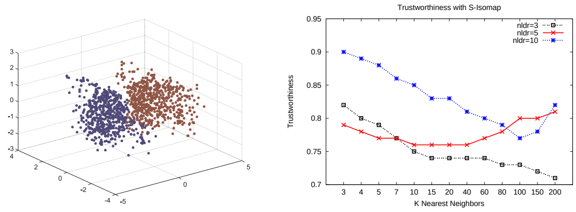

Fig. 1. Data visualization with historical Data Set by S-Isomap.

Table 1 illustrates the performance results for all considered data sets with S-Isomap and with the learning algorithms KNN, SVM and RVM. S-Isomap presents better results in testing accuracy than single KNN and RVM by 2% and 10%, respectively; and comparable one with SVM (sligthly better by 2.5% with the latter), however, with much reduced embedded space (nldr=3) whereas SVM is used with all financial ratios.

Table 1. Performance Measures on Diana Financial Data Sets

S-Isomap	Train	Test	recall	precision	errorTypeI	errorTypeII
2006	91.85 ± 0.54	**87.73** ± **1.54**	86.79 ± 2.62	87.94 ± 1.96	**11.30** ± **1.91**	**13.21** ± **2.62**
2005	78.70 ± 0.91	77.08 ± 2.02	77.13 ± 2.66	76.64 ± 3.62	22.87 ± 3.37	22.87 ± 2.66
2006-2005	94.26 ± 0.41	**89.55** ± **1.01**	89.38 ± 1.98	89.72 ± 1.94	**10.31** ± **2.30**	**10.62** ± **1.98**
2005-2004	96.74 ± 0.27	79.65 ± 1.42	77.61 ± 2.71	80.61 ± 2.12	18.38 ± 2.79	22.39 ± 2.71
KNN	Train	Test	recall	precision	errorTypeI	errorTypeII
2006	90.92 ± 0.76	85.77 ± 1.68	77.95 ± 3.29	92.51 ± 2.00	6.32 ± 1.69	22.05 ± 3.29
2005	84.78 ± 0.76	76.86 ± 1.71	73.22 ± 3.33	79.02 ± 1.98	19.46 ± 1.66	26.78 ± 3.33
2006-2005	91.18 ± 1.00	86.09 ± 1.88	76.99 ± 3.87	94.22 ± 3.03	4.74 ± 2.81	23.01 ± 3.87
2005-2004	84.39 ± 0.81	75.60 ± 1.79	64.80 ± 3.50	82.72 ± 1.65	13.58 ± 1.38	35.20 ± 3.50
SVM	Train	Test	recall	precision	errorTypeI	errorTypeII
2006	95.09 ± 0.42	**90.54** ± **1.28**	89.33 ± 2.24	91.73 ± 1.76	**8.19** ± **1.90**	**10.67** ± **2.24**
2005	86.06 ± 0.76	81.63 ± 1.76	81.01 ± 3.81	82.42 ± 2.84	17.64 ± 2.92	18.99 ± 3.81
2006-2005	95.85 ± 0.55	**91.18** ± **1.28**	92.10 ± 1.93	90.56 ± 1.69	**9.74** ± **1.72**	**7.90** ± **1.93**
2005-2004	89.93 ± 0.66	80.29 ± 1.54	81.04 ± 2.34	79.81 ± 2.58	20.42 ± 2.53	18.96 ± 2.34
RVM	Train	Test	recall	precision	errorTypeI	errorTypeII
2006	97.88 ± 0.63	81.25 ± 1.78	67.35 ± 2.98	92.31 ± 1.98	5.39 ± 2.01	32.65 ± 1.45
2005	93.25 ± 0.54	76.75 ± 1.25	72.64 ± 2.19	79.35 ± 2.34	19.09 ± 1.78	27.36 ± 2.03
2006-2005	99.68 ± 0.35	80.71 ± 2.11	72.47 ± 6.08	89.47 ± 2.55	8.71 ± 2.56	27.53 ± 6.08
2005-2004	100.00 ± 0.0	70.75 ± 1.74	65.36 ± 2.29	73.68 ± 1.53	23.46 ± 1.03	34.64 ± 2.29

4 Conclusion

In this paper we proposed an approach for bankruptcy analysis and prediction based on a supervised Isomap algorithm where class label information is incorporated. Assuming that corporate financial statuses lie in a manifold we attempt to uncover this embedded structure using manifold learning. Isomap acts as a preprocessing stage allowing financial data visualization. Results have shown that comparable testing accuracy can be obtained even using a 3-dimensional reduced space. Further work is necessary to design a method to avoid the interpolation error resulting from the mapping learning stage.

References

1. J. Tenenbaum, V. Silva, J. Langford, “A global geometric framework for nonlinear dimensionality reduction”, Science, vol.290, no.5500, pp.2319-2323, 2000.
2. X. Geng, D. Zhan, Z. Zhou, “Supervised Nonlinear Dimensionality Reduction for Visualization and Classification”, IEEE Transactions SMC, Vol. 35, No. 6, pp. 1098–1107, 2005.