Knowledge-Based Systems 47 (2013) 14–22

Contents lists available at SciVerse ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

Linear versus nonlinear dimensionality reduction for banks' credit rating prediction

C. Orsenigo *, C. Vercellis

Dept. of Management, Economics and Industrial Engineering, Politecnico di Milano, Via Lambruschini 4b, 20156 Milano, Italy

ARTICLE INFO

Article history:
Received 4 September 2012
Received in revised form 23 January 2013
Accepted 1 March 2013
Available online 14 March 2013

Keywords:
Dimensionality reduction
Manifold learning
Isometric feature mapping
Banks' credit rating prediction
Multi-category classification

ABSTRACT

Dimensionality reduction methods have shown their usefulness for both supervised and unsupervised tasks in a wide range of application domains. Several linear and nonlinear approaches have been proposed in order to derive meaningful low-dimensional representations of high-dimensional data. Among nonlinear algorithms manifold learning methods, such as isometric feature mapping (Isomap), have recently attracted great attention by providing noteworthy results on artificial and real world data sets.

The paper presents an empirical evaluation of two linear and nonlinear techniques, namely principal component analysis (PCA) and double-bounded tree-connected Isomap (dbt-Isomap), in order to assess their effectiveness for dimensionality reduction in banks' credit rating prediction, and to determine the key financial variables endowed with the most explanatory power. Extensive computational tests concerning the classification of six banks' rating data sets showed that the use of dimensionality reduction accomplished by nonlinear projections often induced an improvement in the classification accuracy, and that dbt-Isomap outperformed PCA by consistently providing more accurate predictions.

1. Introduction

Dimensionality reduction techniques aim at finding meaningful with low-dimensional representations of high-dimensional data. They may prove quite useful both for unsupervised tasks, such as clustering or data visualization, and supervised learning, to reduce the training time and increase the classification accuracy.

Several approaches have been developed for dimensionality reduction. Traditional methods dealing with linear data projections include principal component analysis (PCA) [20] and factor analysis. These techniques are easy to implement and have shown their effectiveness for data visualization and classification in a wide range of application domains.

More recently, a large number of nonlinear dimensionality reduction methods have been proposed in order to properly handle data with complex nonlinear structures. Within the family of nonlinear algorithms manifold learning methods have attracted great attention. They include, among others, isometric feature mapping (Isomap) [35], locally linear embedding [31] and Laplacian eigenmaps [2]. Manifold learning methods attempt to uncover the low-dimensional manifold along which data are supposed to lie. Given a set of m data points $\mathcal{S}_m = \{\mathbf{x}_i,\ i \in \mathcal{M} = \{1,2,\ldots,m\}\} \subset \Re^n$ arranged along a nonlinear manifold M of intrinsic dimension d, with $d \ll n$, they aim at finding a function $f : M \to \Re^d$ mapping $\mathcal{S}_m$ *into* $\mathcal{D}_m = \{\mathbf{z}_i,\ i \in \mathcal{M} = \{1,2,\ldots,m\}\} \subset \Re^d$ such that some geometrical properties of the data in the input space are preserved in the projection space. Manifold learning techniques exhibited noteworthy performance in the analysis of artificial and real world data sets in several contexts. An empirical comparison of these methods for microarray data classification is presented in [24].

Credit ratings are opinions expressed in terms of ordinal measures which reflect the current financial creditworthiness of issuers, like governments, firms or financial institutions. These ratings are conferred by rating agencies, such as Fitch Ratings, Moody's and Standard and Poor's (S&P's), and may be regarded as comprehensive evaluations of issuers' ability to fully meet their financial obligations on time. Hence, they play a crucial role by providing participants in financial markets with useful information for financial decision planning.

For banks' rating assessment, agencies resort to a broad set of financial and non-financial information, including domain experts expectations. General guidelines on the rating decision process are usually delivered, but the detailed description of the rating criteria and of the determinants of banks' rating is not explicitly provided. Thus, several research efforts have been recently devoted to the development of reliable quantitative methods for automatic banks' classification according to their financial strength.

The motivation for the present study is to evaluate whether dimensionality reduction, accomplished by linear or nonlinear data projections, may enhance the performance of classical and well-established supervised learning techniques for banks' credit rating prediction, so that the resulting methods can be used as forecasting tools for generating credit ratings on the base of a set of measurable financial variables. In particular, the paper has three main objectives.

* Corresponding author. Tel.: +39 02 23993970; fax: +39 02 23993978.
E-mail addresses: carlotta.orsenigo@polimi.it (C. Orsenigo), carlo.vercellis@polimi.it (C. Vercellis).

0950-7051/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.knosys.2013.03.001

First, we intend to determine whether linear and nonlinear dimensionality reduction methods, namely principal component analysis and double-bounded tree-connected Isomap (dbt-Isomap), may be effectively used for credit rating prediction when they are combined with support vector machines, naïve Bayes classifier and k-nearest neighbor. Then, we are interested in investigating whether nonlinear dimensionality reduction dominates its linear counterpart. Finally, we aim at analyzing the key explanatory factors exploited by both methods in order to highlight the different role of the financial variables on the prediction task. Hybrid approaches based on PCA and manifold learning were proposed in [21,29,30] for predicting business failure. We are not aware of any previous study resorting to manifold learning for dimensionality reduction to model and predict banks' credit ratings.

The remainder of the paper is organized as follows. Section 2 offers a brief review of the literature on prediction models for banks' credit rating. Section 3 provides a general description of the dimensionality reduction techniques considered in this study. Sections 4 and 5 illustrate the credit rating data sets and the experimental settings, respectively. Section 6 presents the most relevant empirical findings. Conclusions and future extensions are discussed in Section 7.

2. Related works

Extensive empirical research devoted to analyze the stability and soundness of financial institutions date back to the 1960s. We refer the reader to [28] for a comprehensive survey on the application of statistical and intelligent techniques for predicting the default of banks and firms.

Despite its relevance, however, only recently the development of reliable quantitative methods for banks' credit rating prediction has drawn great interest. These studies are mainly comprised within two broad research strands focusing on statistical and machine learning techniques, and may address both feature selection and classification.

Poon et al. [27] developed logistic regression models for predicting financial strength ratings assigned by Moody's, using bank-specific accounting variables and financial data. Factor analysis was applied to reduce the number of independent variables and retain the most relevant explanatory factors. Authors showed that loan provisions information, risk and profitability indicators added the greatest predictive value in explaining Moody's ratings.

Huang et al. [15] compared support vector machines and back-propagation neural networks to forecast the rating of financial institutions operating in the United States and the Taiwanese markets, respectively. In both cases five rating categories were considered, based on the information released by S&P's and by the Taiwan Ratings Corporation (TRC). The analysis of variance was used for discarding non informative features. In this study, support vector machines and neural networks achieved comparable classification results; however, authors noticed that the relative importance of the financial variables used as inputs by the optimal models were quite different across the two markets.

Gaganis et al. [9] presented a multicriteria decision aid model for the classification of banks into three groups, according to their financial soundness. The assignment of the banks into each group was based on Fitch individual ratings. The sample was composed by banks operating in 79 countries described in terms of six financial and four non-financial explanatory variables. Their results indicated that loan loss provisions, capitalization and the market, where the bank operates were the most important criteria in banks classification. A similar task was accomplished in [16], where models were generated incorporating country-specific variables and indicators related to the regulatory environments, and where alternative machine learning techniques, such as artificial neural networks, classification trees and k-nearest neighbor were compared. This study indicated an overall improvement in accuracy when country-level indicators were employed.

Pasiouras et al. [25] resorted to multi-group hierarchical discrimination in order to replicate Fitch individual ratings of commercial banks operating in the main South and Southeastern Asian countries. Their approach achieved considerable results in terms of classification accuracy dominating discriminant analysis and ordered logistic regression. Among the financial variables selected by means of factor analysis, equity over customer and short-term funding, net interest margin and return on average equity appeared to be as the most influential; the number of shareholders and subsidiaries and the environment of the bank's operating country resulted as the most relevant non-financial factors.

Bellotti et al. [3] employed financial variables and country-level indicators as determinants of Fitch individual ratings, and applied support vector regression and ordered choice models for predicting the rating of commercial banks from 90 countries. They came to the conclusion that support vector regression was capable of better classification results with respect to two versions of ordered logit and probit models developed for comparison. They also highlighted the crucial role of country effects indicators for banks' rating prediction.

Hammer et al. [14] addressed the problem of reverse-engineering Fitch individual ratings by using multiple linear regression, ordered logistic regression, support vector machines and logical analysis of data. The creditworthiness of banks operating in 70 countries was evaluated in terms of a set of 24 representative predictors, including financial variables, financial ratios and a risk indicator which modeled country effects. They showed that ordered logistic regression and logical analysis of data provided superior conformity of banks' ratings, the latter exhibiting outperforming results in the classification task.

Chen [5] proposed a hybrid procedure for predicting Fitch long-term ratings of a sample of Asian banks that covered five rating categories. It relied on an integrated feature selection method used to identify the relevant variables represented by 16 financial ratios, which was further combined with a cumulative probability distribution approach for discretizing the data, and with rough set theory for generating the decision rules. Two hybrid models based on rough set theory were also implemented in [6] for classifying large banks from various countries according to Fitch long-term ratings. Feature selection was applied on the original set of 37 financial variables after discarding the uncorrelated attributes: in particular, the first model resorted to factor analysis, whereas the second to the reduct method within rough set theory.

A summary of these studies in terms of rating agency, time horizon, sample size, number of classes and explicit use of feature selection is reported in Table 1.

3. Linear and nonlinear dimensionality reduction

This section presents a general description of the linear and nonlinear dimensionality reduction techniques investigated in the present study: these are principal component analysis and a variant of isometric feature mapping. In what follows, $\mathbf{X}$ denotes the $m \times n$ matrix whose rows represent the input vectors $\mathbf{x}_i$, $i \in \mathcal{M}$, and $\mathbf{Z}$ the corresponding $m \times d$ matrix whose rows are the projected vectors $\mathbf{z}_i$ in the low d-dimensional space.

3.1. Principal component analysis

Principal component analysis (PCA) is a widely used statistical technique for linear dimensionality reduction. It is most effective