Computational Statistics and Data Analysis 56 (2012) 981–994

Contents lists available at SciVerse ScienceDirect

Computational Statistics and Data Analysis

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

Credit scoring analysis using a fuzzy probabilistic rough set model

Andrea Capotorti *, Eva Barbanera

Università degli Studi di Perugia, Italy

ARTICLE INFO

Article history:
Available online 12 July 2011

Keywords:
Rough fuzzy multicriteria classification
Coherent conditional probability assessments
Credit scoring

ABSTRACT

Credit scoring analysis is an important activity, especially nowadays after a huge number of defaults has been one of the main causes of the financial crisis. Among the many different tools used to model credit risk, the recent development of rough set models has proved effective. The original development of rough set theory has been widely generalized and combined with other approaches to uncertain reasoning, especially probability and fuzzy set theories. Since coherent conditional probability assessments cope well with the problem of unifying these different approaches, a merging of fuzzy rough set theory with this subjectivist approach is proposed. Specifically, expert partial probabilistic evaluations are encompassed inside a gradual decision rule structure, with coherence of the conclusion as a guideline. In line with Bayesian rough set models, credibility degrees of multiple premises are introduced through conditional probability assessments. Nonetheless, discernibility with this method remains too fine. Therefore, the basic partition is coarsened by equivalence classes based on the arity of positively, negatively and neutrally related criteria. A membership function, which grades the likelihood of default, is introduced by a peculiar choice of t-norms and t-conorms. To build and test the model, real data related to a sample of firms are used.

1. Introduction

In this paper we propose a hybrid methodology for classification based on the methodologies of rough sets, partial conditional probability assessments and fuzzy sets. Due to common concomitance of uncertainty and imprecision in real life data, the generalization of original Pawlak rough set theory (Pawlak, 1982) and its combination with probability and fuzzy set theory has already been proposed (see, among others Dubois and Prade, 1992; Greco et al., 1999b, 2004, 2006, 2007; Yasdi, 1995; Yao, 2008; Yao and Wong, 1992; Yao et al., 1990; Ziarko, 2005, 2008). Even though a further proposal in this direction might appear unnecessary, we believe the use of the more general setting of *coherent partial conditional probability assessments*, instead of the common probabilistic approaches, could bring new insights. The framework of coherent partial conditional probability assessments finds its roots in the work of de Finetti (1974–1975), and it has been recently shown (Coletti and Scozzafava, 2002, 2004, 2006) to be a powerful tool for unifying different approaches to uncertain reasoning.

Similarly to Lyra et al. (2010), our approach has been inspired by *credit scoring*, which is the general term used to indicate methods utilized for classifying credit applicants into classes of risk on the basis of *probability of default* values. The probability of default, also named *expected default frequency* or *probability of insolvency*, is a basic parameter of the Basel II Accord used in the calculation of economic capital. The probability of default expresses the likelihood that a loan will not be repaid and will fall into default. Consequently a firm that is either unwilling or unable to pay its debt is classified in a *state of default*, or as a *defaulting firm*.

* Corresponding address: Dip. Matematica e Informatica, via Vanvitelli 1, 06126 Perugia, Italy. Tel.: +39 0755855011; fax: +39 0755855024.
E-mail address: capot@dmi.unipg.it (A. Capotorti).

0167-9473/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2011.06.036

Credit scoring analysis is an important activity, especially nowadays that a huge number of defaults has been one of the main causes of the current continuing world financial crisis. In an attempt to reduce costs of protection for firms, both academics and practitioners have studied models for evaluating business failures, according to their differing interests and the differing conditions of the firms being examined. Nevertheless, until now there is no generally accepted model for the prediction of business failures.

In the past, classical statistical methods were predominantly used to predict credit risk, (see Balcaen and Ooghe, 2006 for an overview of applications using univariate statistical models, multiple discriminant analysis, linear probability models, logit regression, and probit analysis). However, these models remain ineffective as they have to rely on various restrictive assumptions, such as a large number of samples, normally distributed independent variables, and a linear relationship between all variables. Nevertheless, the main drawback of these statistical methods is that, as stressed in Doumpos and Zopounidis (2002, Sec. 3), statistical properties of the data are rarely known, and it is difficult to specify the underlying population being considered. Hence non-parametric techniques are preferable. Such approaches can be flexible enough to adjust themselves to the peculiarities of the data under consideration.

Recently, rough set has shown its effectiveness in constructing a prediction model or in being combined with other approaches to identify key attributes relevant to risk. This procedure usually starts with a pre-process of analyzing the attributes crucial for detecting similar cases and predicting the value of the target variables (e.g. see, among others, Słowinski and Zopounidis, 1995; Słowinski et al., 1997; Tay and Shen, 2002 and Lin et al., 2009). In particular, Słowinski et al. (1997) demonstrate how the rough set approach can outperform classical discriminant analysis; and Greco et al. (1998) show how a generalization through the dominance principle of rough set theory properly deals with financial classification problems. Here we begin by following the same line. The rough set approach is used to aggregate similar instances in equivalence classes, and fuzzy set theory is used to grade the likelihood of getting in default for elements in each class. The difference in our development arises when the original attributes are transformed by experts' probabilistic evaluations into criteria; the partition generated by such criteria will be coarsened into meta-classes by profiting from a reasonable assumption of conditional exchangeability; and these meta-classes will then be graded by values built by t-norms and t-conorms. These are induced naturally by the interpretation of membership functions as coherent conditional probability assessments.

The development of our reasoning is similar to the historical evolution of the original rough set theory (RS) into the so called stochastic-dominance-based rough set approach (Stochastic-DRSA) (Dembczynski et al., 2007). For this reason, in Section 2 we reformulate the main steps of the passage from RS to a Bayesian rough set model (BRS) (Slezak and Ziarko, 2002) following the probability reinterpretation illustrated in Yao (2008). However, a BRS could involve several drawbacks that have inspired further generalizations, such as those briefly listed in Section 2.4. These have led to the development of Stochastic-DRSA. In contrast, we propose avoiding the same drawbacks in a different way, which we illustrate in Section 4 through the exemplification of default risk analysis performed over a sample of firms of the Umbria region of central Italy. The main novelties introduced are the coarsening of the original ordered categories into a meta-criterion, and the methodology adopted to obtain the membership function. Finally, Section 5 concludes the paper with a short comment.

2. Evolution of rough set classification methods

In this section we briefly review the evolution of the rough set classification methods that led to the so called "two-parameter" probabilistic rough set approximation. In doing this we mainly follow Yao (2008), taking advantage of the unifying approach that can be expressed through conditional probabilities.

2.1. Rough set theory

Classic rough set theory (RS) was proposed by Pawlak (1982) in order to approach approximate classification problems. The starting point is the availability of a finite set U of objects described through a set of attributes $\mathcal{A}$.

Let $E \subseteq U \times U$ be an equivalence relation on U. That is, E is reflexive, symmetric, and transitive. The basic building blocks of rough set theory are the equivalence classes of E. For an element $x \in U$, the equivalence class containing x is given by:

$$[x]_E = \{y \in U : xEy\}. \tag{1}$$

When no confusion arises, we also simply write $[x]$. The family of all equivalence classes is also known as the *quotient set* of U, and is denoted by $U/E = \{[x] : x \in U\}$. It defines a partition of the universe, namely, a family of pairwise disjoint subsets the union of which is the universe. Usually, the equivalence relations adopted are those induced by the subsets of attributes $\mathcal{A}$. Moreover $\mathcal{A}$ is commonly divided into disjoint sets of condition attributes $\mathcal{C}$ and decision attributes $\mathcal{D}$. For simplicity, we assume the set $\mathcal{D}$ to be a singleton $\mathcal{D} = \{D\}$. Equivalence classes are the elementary definable, measurable, or observable sets in the *approximation space* $\text{apr} = (U, E)$. By taking unions of elementary definable sets, one can derive larger definable sets.

The key idea of rough sets is to approximate one body of knowledge by another less detailed one. In classical RS the approximated knowledge is the partition of a fine set of elements U into classes induced by the decision attribute D; the knowledge used for approximation is another partition of U into elementary sets of objects that are indiscernible by the set of condition attributes $\mathcal{C}$. The elementary sets are seen as "granules of knowledge" used for approximation.