Bankruptcy forecasting using case-based reasoning: The CRePERIE approach

Fabio Sartori, Alice Mazzucchelli, Angelo Di Gregorio

A R T I C L E   I N F O
Article history:
Received 17 May 2016
Revised 12 July 2016
Accepted 22 July 2016
Available online 26 July 2016

Keywords:
Case-based reasoning
Bankruptcy prediction
Overall similarity

A B S T R A C T
Bankruptcy prediction is a very important research trend: although statistical methods are mainly used in literature, techniques based on Artificial Intelligence are interesting from many points of view. Among them, Case-Based Reasoning (CBR) could be useful to cluster enterprises according to opportune similarity metrics as well as suggest proper actions to take for avoiding bankruptcy in border-line situations. In this paper, we present a new and still under development CBR approach to this problem, that seems to return better results than previous attempts. The approach is based on different kinds of similarity metrics and is focused on the implementation of innovative revise algorithms. In particular, the paper shows how the revise step is crucial to improve the accuracy of the bankruptcy prediction model.

© 2016 Elsevier Ltd. All rights reserved.

1. Introduction

Over the last decade, the globalization phenomenon has determined the growing integration of markets, as well as a concrete structural modification of enterprise capability to compete. The increasing level of uncertainty and the complexity of the competitive scenario imply the need for continuous investments to improve both process and product innovation: many firms are not able to revise them, with the consequence to close from one year to another. The recent global market crisis, as demonstrated by recent cases of enterprise bankruptcy, has amplified this problem. In this scenario, how to forecast enterprise bankruptcy has become an important and multidisciplinary research topic.

According to Aziz and Dar (2006), prediction models can be divided into three main categories: Statistical models, which include the well known Multiple Discriminant Analysis - MDA (Altman, 1968) and Logit approaches; Artificial Intelligence Expert System models (AIES), like Artificial Neural Networks-ANN (Alici, 1996), Decision Trees and Ensemble Classifier (Nanni & Luminii, 2009; Zieba, Tomczak, & Tomczak, 2016); and Theoretical models, like BSDM and Cash management theories. Although statistical models have been widely adopted in the past, their main limitation is the subjective nature of the prediction, which makes difficult to perform consistent estimates, as explained in Atyia (2001). Also Liang, Lu, Tsai, and Shih (2016) state that among many techniques employed to develop bankruptcy prediction models statistical techniques show to fail with Artificial Intelligence Expert System models, especially with machine learning techniques.

In this paper, we present the results of a research conducted on Italian Small and Medium Enterprises (SMEs) that aims at identifying potentially bankrupting firms. In Italy, one of the countries most affected by the financial crisis with 306000 business closures registered in 2013-2015¹, the bankruptcy condition can be one of three following legal status: Inactive, Failed, or Liquidation; on the other hand, a non-bankrupt enterprise is characterized by an Active legal status. Given a collection of certainly bankrupt companies and a collection of Active enterprises, the goal of the system is comparing the non-bankrupt firms’ dataset with the bankrupt firms’ one. Doing so, it should be possible to discover enterprises potentially moving from the Active status to the Inactive, Failed or Liquidation ones. Our research exploited Case-Based Reasoning (Kolodner, 1993), from the case structure construction to the revise phase of the system development.

The development of CBR systems is typically divided into four steps, as stated by the well-known 4Rs’ cycle (Aamodt & Plaza, 1994):

- Retrieve: the current case $C_C$ is compared with the case base (possibly indexed to increase efficiency) in order to find the most similar past case, namely the retrieved case $C_R$;
- Reuse: the $C_F$ solution is associated to $C_C$;

• Revise: since \( C_t \) is the most similar past problem to \( C_r \), but it is not equal to \( C_r \), the reused solution is modified in order to transform it in the real \( C_r \) solution;
• Retain: finally, the current case is solved and added to the case base together with its solution. Doing so, the CBR system learns new problem solving experiences that can be exploited in the future.

A general-purpose CBR platform, namely CREPERIE (Manenti \& Sartori, 2011), has been adopted in the research. This platform implements two kinds of similarity metrics, Overall similarity, to index current case with respect to the whole case base, and Local similarity, to retrieve the most similar cases from the archive using the Overall similarity as a threshold. For this reason, CREPERIE is very suitable to analyze potentially bankrupting enterprises, since Active enterprises can be indexed to understand their behavior according to different bankruptcy definitions and then compared to the most similar case (on the basis of indexing calculus) to understand how their legal status could be maintained in the future.

Moreover, CREPERIE is able to revise reused solutions to make them more suitable to solve new problems. Starting from the Overall and Local similarity definitions, we have developed an algorithm, preserving the peculiarities of CBR with respect to hybrid or different approaches. In this sense, CREPERIE implements a pure and complete CBR strategy, exploiting substitutional adaptation in the revise step. Substitutional adaptation is based on the principle that similar differences in the case description of the current case and the retrieved case induce similar differences in their solutions. In CREPERIE, we have implemented the \( \Delta \) analysis technique complying with this definition.

We have tested the proposed method on a dataset of bankrupt and non-bankrupt Italian enterprises, related to the manufacturing district of Monza and Brianza province. As pointed out in Gordini (2014), most of Italian enterprises (especially the ones belonging to our dataset) are SMEs, with the consequence that traditional models of bankruptcy prediction have difficulties in clustering them. This opinion has been confirmed by the early step of our test, where only the retrieval and reuse steps of the CBR cycle have been considered. Most significant discussion emerges from the analysis of revise results: the CREPERIE approach has been compared to the \( Z^2 \)-score evaluation of our dataset, being currently the most used MDA index to predict Italian enterprises bankruptcy, as highlighted in Altman, Danovi, and Falini (2013).

The rest of the paper is organized as follows: Section 2 briefly reviews the literature on bankruptcy forecasting and introduces the CBR paradigm. Section 3 describes the main elements of the CREPERIE platform, focusing on the CBR cycle designed and implemented in it. Section 4 presents the research case study, showing the main characteristics of CREPERIE method from the practical point of view. A detailed analysis of results obtained is pointed out in Section 5, where the CREPERIE performance is compared to \( Z^2 \)-Score by means of a Student’s \( t \)-test. Finally, Section 6 ends the paper with a brief explanation of managerial implications and possible future works.

2. Related work

To develop bankruptcy predictions, many techniques have been used in previous studies (Liang et al., 2016). One of the most complete review of methods for bankruptcy prediction was proposed in Bellovary, Giacomino, and Akers (2007). These authors analyzed 165 bankruptcy prediction studies published from 1965 to 2007, pointing out four interesting conclusions:

1. MDA and ANN are the most promising methods to forecast bankruptcy. The main reason for this conclusion is their great accuracy, due to the capability to minimize the Type I errors percentage;
2. The authors suggested to use existing bankruptcy prediction models as opposed to the development of new methods given that more than 150 models were available, and many of them show high predictive ability;
3. Another interesting conclusion pointed out that a large number of factors does not necessarily increase the model’s predictive ability;
4. Finally, they highlighted how a stronger connection between research and practice is necessary.

Also Mousavi, Ouenniche, and Xu (2015) proposed a literature review of the main bankruptcy prediction models in order to assess their different performance and select the best one(s), through the DEA-based framework. The authors suggested that some models perform better than others, primarily the survival analysis, and that the choice and the nature of explanatory variables seem to affect models performance.

Regarding AI-based techniques, they have been widely applied to bankruptcy prediction, in particular, Artificial Neural Networks (Ciampi \& Gordini, 2013; Kim & Kang, 2010; Wilson \& Shar da, 1994; Zhang, Hu, Patuwo, \& Indro, 1999) and support vector machines (Gestel et al., 2003; Kim \& Sohn, 2010; Min, Lee, \& Han, 2006; Shin, Lee, \& Jung Kim, 2005). Many studies have demonstrated the ANN higher level of accuracy than other methods (Ciampi \& Gordini, 2013; Zhang et al., 1999), as well as some problems like the dependence on researcher’s experience in control parameters selection (Kim \& Sohn, 2010) and difficulties in generalizing the results due to over-fitting phenomena (Santos, Sabourin, \& Maupin, 2009). Boritz, Kennedy, and Albuquerque (1995) demonstrated how the ANNs performance is sensitive to the choice of variables. Bryant (1997) highlighted that ANNs often break down with a great number of predictor variables and they are not able to provide explanations or justifications for their classifications.

To avoid these problems other approaches have been profitably exploited in the recent past: for example, Min and Lee (2005) used genetic algorithms in association with other methods, while Gordini (2014) approached them in a stand-alone way to build up innovative bankruptcy prediction models for Italian small and medium enterprises.

Some attempts to apply Case-Based Reasoning methodology have been proposed too. The basic principle of CBR is that similar problems have similar solutions. In other words, if we are trying to solve a problem at time \( t_i \), and we are not able to produce an analytical problem solving strategy for it, we can proceed by analogy with problems solved in the past at times \( t_0, t_1, ..., t_{i-1} \). The problem without solution is called current case \( C_r \), while the solved problems are defined past cases \( C_{p_0}, ..., C_{p_{i-1}} \). The past cases are typically collected in a case base.

Park and Han (2002) and Peng and Zhuang (2007) focused on the retrieval step based on K-NN metrics, with the need for a careful and deep definition of weights to apply; Ahn and Kim (2009) defined a hybrid approach, where CBR was combined with genetic algorithms to improve the classificatory performance of the reasoning by analogy paradigm; similar attempts (Chuang, 2013; Sungbin, Hyojung, \& Byoung-Chun, 2010) were made using decision trees and rough set theory to support CBR retrieval in classifying bankrupt firms. What emerges from these references is the statement of CBR as a promising, relatively simple and good from the explanatory perspective approach to bankruptcy forecasting, if used in association with other methods to reduce the extremely high error percentage.

Bryant (1997) applied it in a stand-alone way, analyzing the performance of indexing and retrieval algorithms provided by ReMind platform (Watson \& Marit, 1994) on a case base of 85
enterprise bankrupt between 1978 and 1994. Jo, Han, and Lee (1997) compared CBR with MDA and ANN on three sets of Korean firms bankrupt during 1991–1993, for a total of 261 failed enterprises. CBR was judged inappropriate for bankruptcy prediction in both studies, although the second one recognized its utility in case of insufficient training data.

This divergence of opinions shows how the adoption of CBR in this domain needs further research; in fact, although it is averagely worse than others in classifying bankruptcy enterprises, it is very useful to understand which actions to take for avoiding bad situations become critical. In our opinion, the main limitation of past comparative studies is that they didn't consider fully developed CBR systems: the Revise step is defined in Bryant (1997, p. 200) not applicable for classification tasks, while Jo et al. (1997) simply ignore it. Doing so it is not possible, in our opinion, to understand the real potentialities of CBR paradigm in predicting bankruptcy. Moreover, it is important to highlight that classification is only one aspect of CBR systems, whose final goal is finding solutions to complex problems characterized by episodic and heterogeneous knowledge.

In particular, to compare problem descriptions and reuse past solution (possibly revising them) to solve new problems can be profitably exploited to build up complete decision support systems for: 1) identifying firms closed to bankruptcy; 2) identifying the most probable causes of their criticalities; 3) suggesting how to eliminate those causes before the critical situation become irreversible.

In this sense, CBR should be evaluated in its complete development life cycle. Although the revise step has been recognized in the past too complex, time-consuming and just the opposite of foundational CBR issues (Wilke & Bergmann, 1998), it should be always considered when case-based systems are going to be developed. This is particularly true in domains like bankruptcy prediction, where errors in characterizing a firm could have serious consequences: the definition of proper algorithms for the revise step would allow the prediction system to correct errors made in the retrieval phase.

3. CBR for bankruptcy forecasting: the CRePERIE elements

To implement the CBR life cycle, the following elements must be defined:

- Case Structure: being the CBR based on problem comparisons, it is necessary to define how these problems can be described, i.e. which are the main attributes and relationships among them;
- Similarity function(s): the retrieve step aims at ranking the past cases according to their similarity with the current case. To this aim, functions working on the case structures must be involved, returning values in the range [0...1];
- Revise algorithms: the Revise step of the CBR methodology is the most difficult one, as reported in literature (Greene, Freyne, Smyth, & Cunningham, 2008; Wilke & Bergmann, 1998) since it is necessary to know deeply the knowledge domain to understand how modifying a reused solution. It is generally left to the user the possibility to update solutions at hand (i.e. the null adaptation approach) or hybrid approaches are implemented based on genetic algorithms, Artificial Neural Networks and so on (Portinale & Torasso, 2001). Anyway, some successful attempts to use pure CBR in this step have been registered, like in Manzoni, Sartori, and Vizzari (2007).

Our approach exploits the CRePERIE platform: in the following we describe how the points above have been considered to fit the paper case study.

3.1. Case structure

In our approach, a case is a collection of case elements which correspond to nodes of a tree-structure. Formally, a case element ce is a member of the CaseElement set: $\forall ce \in \text{CaseElement}$, $ce = (id, t, n)$ where: $id \in \mathbb{Z}^+ = \{0\}$ is the identifier of the case element; $t \in T$ identifies the range of value associated to ce (i.e. String, Integer, Double); $n \in \text{Name}$ is the name used to refer to the case element.

A case base $C = \{c_1, ..., c_n\}$, with $n \in [1, \infty)$, is a finite and not empty collection of cases and every case is defined as a set of couples $(ce_i, v_i) \ldots (ce_n, v_n)$ where $V(ce_i, v_i), ce_i \in \text{CaseElement}$ and $t(ce_i) = t$ and $v_i \in \text{v}(\{\}, \{\})$, where $\bot$ means the bottom type (a.k.a. the zero type or empty type); bottom is the subtype of every type in the type system. Thus, $v_i$ is the value associated to $ce_i$.

Each case has to be organized following a particular tree-structure, where inner nodes and outer nodes can be identified: outer nodes, also named attributes overlap with leaves whereas inner nodes represent categories which the attributes belong to. In our approach, only one structure can be defined for each case base.

The structure defines the three parts of a generic case according to literature: $\forall x \in \text{StructBase}, x = (d, s, o)$, where StructBase is a finite and not empty collection of case structures; $d(x) = d$ denotes the description part; $s(x) = s$ denotes the solution part; $o(x) = o$ denotes the outcome part. An XML schema has been adopted to represent the case structure in CRePERIE (Manenti & Sartori, 2011).

In the structure, each node overlaps with a particular $ce \in \text{CaseElement}$. Moreover, in order to support the fundamental methodology of CBR approach based on comparison among cases, every node is coupled with a similarity function $sf$, defined as $sf: C \times C \rightarrow [0, 1]$, where $C_1$ and $C_2$ denote the case descriptions of the current and retrieved cases respectively.

3.2. Similarity functions

Similarity metrics in CBR have been recently classified (Cunningham, 2009) as direct mechanisms, transformation-based mechanisms, information-theoretic measures and emergent measures. The first category is the most applied in the CBR research, since they operate on feature vector representations. Another important category is the second one, where metrics suitable for taking care of not only feature values, but also their relationships within more complex structures are considered.

CRePERIE allows to manage both those kinds of similarity measures, grouping features into inner or outer nodes. Outer nodes contain the concrete values of the case description (both numeric and alpha-numeric values) to be compared on the basis of direct mechanisms, while inner nodes represent a category, i.e. a collection of correlated outer nodes. Depending on the application domain, it is possible to choose the most suitable similarity function to adopt in the retrieval step.

A first type of similarity functions works on number intervals (both integer and double): it is necessary to define minimum and maximum values of the range allowed for every node. The similarity value on a node $n$ corresponding to case element $ce$ between two cases $x$ and $y$ is defined as follows:

$$ sf(n, x, y) = 1 \frac{|v_{ce}(x) - v_{ce}(y)|}{\max - \min} \tag{1} $$

where $\max = v_{ce}(n) \in x \cup y : v_{ce}(m) \geq v_{ce}(m), \forall v_{ce}(m) \in x \cup y$ and $\min = v_{ce}(k) \in x \cup y : v_{ce}(k) \leq v_{ce}(f), \forall v_{ce}(f) \in x \cup y$. In practice, $\max$ and $\min$ values can be substituted by the extremes of the normalization interval in case of need.

A second type of similarity functions works on strings: they compare strings in case-sensitive or case-no-sensitive ways.
according to different methods [Mazzucchelli & Sartori, 2014], returning a value in the range [0…1].

New similarity functions can be defined according to the application domain: they will be added to the case structure and used in the case instantiation.

3.2.1. Local similarity and overall similarity

The similarity functions $s$ defined above allow to compare the values assumed by an attribute in two distinct case descriptions; the similarity of the current case $x$ with respect to a case $y$ in the case base is calculated through the following equation:

$$sim(x, y) = \frac{\sum_{n \in d} w_n s_f(n, x, y)}{\sum_{n \in d} w_n}$$  \hspace{1cm} (2)

where $w_n \in [0, 1]$ is the weight of the attribute $n$; $sim(x, y)$ is called Local similarity, since it focuses on a specific case in the unsolved cases' set. Local similarity is the representation in CRePERIE of the ranking method adopted in the retrieve step of the CBR methodology, with the aim to find the most similar $y$ case to the current $x$ case and reuse its solution. Local similarity is a typical direct mechanism, based on the well known K-NN algorithm (Finnie & Sun, 2002).

Another type of similarity metric defined in the CRePERIE platform is the Overall similarity:

$$SIM(x, C) = \frac{\sum_{y \in C} sim(x, y)}{\#C}$$  \hspace{1cm} (3)

where $\#C$ is the cardinality of the case base $C$.

In Psychology (Wills, Milton, Longmore, Hester, & Robinson, 2013), **overall similarity** is a technique to classify subjects, typically humans, according to various and heterogeneous information. For this reason, it is also called **multi-dimension classification**. Many authors highlight how multi-dimension classification is less performing of single-dimension approach, where subjects are classified on the basis of only one specific indicator. Overall similarity is becoming an important research topic in many disciplines, like Philosophy (Guigon, 2014), Medicine (Milton, Wills, & Hodgson, 2009) and Economy (Brack & Benkenstein, 2012).

Also Computer Science, and in particular CBR research, can profitably exploit the overall similarity concept. At a first glance, a CBR cycle could be seen as an example of multi-dimension analysis, where the different features adopted to describe the case are aggregated through the adoption of a similarity mechanism in the retrieval step. Anyway, if we look at it more deeply, we can observe that the typical CBR cycle is an instance of single-dimension classification, where an unsolved case is compared with solved ones, and the dimension used for classification is the whole case description. For this reason, the Local similarity introduced above is a single-dimension classification approach, that is present in CRePERIE since its first development (Manenti & Sartori, 2011). More recently, the similarity mechanisms of CRePERIE have been extended with the possibility to compare cases and/or set of cases with the entire content of a case base: in this sense, the analysis offered is multi-dimensional, given that the classification of a subject (i.e. a case) considers various and heterogeneous cases. For this reason, we have called the similarity metric defined by Eq. (3) **Overall similarity**.

Overall similarity goal is to understand the average behavior of an unsolved case $x$ with respect to the whole case base $C$ used in the similarity ranking. By calculating it on every case $x$, it is possible to cluster the unsolved cases, distributing them according to their similarity value in a given number of sectors, depending on application scenario (see Section 4 for an example). It is useful in the maintenance and analysis of the case base, because it provides information about the algorithm performance on the whole set of unsolved cases rather than a single one and it is not related to the reuse step of CBR, different from Local similarity. In this sense Overall similarity is a **transformation based mechanism**, that exploits a direct mechanism like Local similarity together with structural information coming from the case base nature (depending on the application domain) to return different kinds of outputs.

3.2.2. Reuse and revise steps in CRePERIE: the role of $\Delta$ analysis

The obtained Overall similarity value is used in the next retrieval-reuse phase: the goal of the reuse phase in CBR systems is to find the solution of the most similar solved case and use it as the starting point for deriving the new solution. In the CRePERIE approach, given the current case $x$, the reused solution $s_x$ is returned by the following formula:

$$s_x = s_y, y \in C : |sim(x, y) - SIM(x, C)| < \epsilon, \epsilon \in \mathbb{N}^+$$  \hspace{1cm} (4)

In this way, the current status of the most similar enterprise in the case base, on the basis of Local similarity notion, is associated to the unsolved case. Moreover, the solution is completed with the indication of differences between indexes' values of the $x$ and $y$ descriptions:

$$\Delta(x, y) = [\delta_1(x, y), \delta_2(x, y), \delta_3(x, y)]$$  \hspace{1cm} (5)

3.3. The CBR cycle

Fig 1 depicts the CBR cycle adopted in the research. The problem description contains enterprises balance indexes (economic-financial ratios), selected according to the rules described in Section 4. An Active enterprise is characterized by it, while the case base is a collection of firms presenting a problem solution too. The problem solution is described by the current legal status of the enterprise together with the so called $\Delta$ Indexes. The current legal status of the enterprise is one of the following (a sort of Italian Chapter 7 and Chapter 112):

- Inactive, about an enterprise that no longer practices;
- Failed, about a firm having a non-revoked bankruptcy proceeding such as bankruptcy, bankruptcy agreement, arrangement with creditors;
- Liquidation, about an enterprise having a non-revoked winding-up proceeding (compulsory or voluntary winding-up).

$\Delta$ Indexes are the differences between the balance indexes of the enterprise and the retrieved case whose solution has been reused in the past: for the firms in the original case base (i.e. the case base without any new case retained), these differences are equal to the balance indexes themselves.

The whole forecasting strategy in CRePERIE is shown in Algorithm 1. When an Active enterprise $x$ is going to be analyzed,

**Algorithm 1** Bankruptcy forecasting in CRePERIE.

**Require**: $x, C$

**Ensure**: $s_x$

1: $(s_x, y) = \text{Retrieve-Reuse}(x, C)$
2: $s_x = \text{Revise-FirstStep}(s_x, s_y, y)$
3: $s_x = \text{Revise-SecondStep}(x, s_x, C)$


it is compared with the whole case base $C$ according to the Overall similarity notion, calculated on the problem description normalized balance indexes. This indexing step allows to cluster Active enterprises into one of three areas (see Section 4 for further details), namely RED, YELLOW or GREEN, according to their level of risk (RED area is closer to bankruptcy condition).

The indexing, retrieve and reuse steps are described in Algorithm 2. Three initially empty sets are defined to contain

---

the clustering results (GreenArea, YellowArea and RedArea respectively). The current case x is compared with the whole case base according to the similarity functions (1), (2) and (3). The obtained OverallSimilarity, is then compared with opportune thresholds. Thresholds are determined according to the following formula:

$$Threshold_d = \frac{\sum_{i=1}^{n} (d - i)}{n}$$

(6)

where n is the number of completely different indexes and d is the number of indexes included in the case description d. Given that two completely equal cases should have exactly #d equivalent indexes and two completely different cases should have 0 equivalent indexes, Eq. (6) calculates the areas’ boundaries according to the average number of equivalent indexes between the current case and the whole case base. This number will vary according to the number of areas we want to define. Since our algorithm divides the enterprises into three groups, we’ll have to define two thresholds, to decide which area x belongs to (Threshold, and Thresholds), depending on the case study (see Section 4 for further details).

The Overall similarity value is then used as the entry point to retrieve the past case to compare the current case with, on the basis of the Local similarity metric. This is the most significant difference between CRePERIE and traditional CBR retrieval algorithms: the scope of retrieval phase in CRePERIE is not to find the most similar past case, but at least one case to confirm the classification of x as bankrupt or not-bankrupt firm. Given that x will probably have at least one inactive enterprise with a very high level of similarity (due to the normalization of balance indexes), the application of a traditional retrieval algorithm would address towards misleading results. Instead, the CRePERIE approach aims at comparing x with Inactive firms close to it in the related area. To do this, LocalSimilarity, y value obtained at line 18 of Algorithm 2 is evaluated with respect to OverallSimilarity, x y: if at least one past case y is found, its solution s_y is reused in x (see line 20).

In order to decide if reusing the retrieved solution or not, a Δ Indexes calculus is performed between the two case descriptions in Algorithm 3, that explains how the first-step revise is accomplished; each (not normalized) balance index defined in d_y, namely \(v_i(x)\) is decreased by the corresponding index in d_y, namely \(v_i(y)\) to get the related \(\delta_i(x, y)\) value. This value gives useful information according to the balance indexes set chosen in the case structure definition (further details can be found in Section 4). The number of negative and/or positive (depending on the case study) \(\delta\) values is determined, increasing a BadIndex variable keeping trace of bad indexes in d_y with respect to d_x. If the final value of BadIndex is higher than an opportune threshold, s_y is stored in the case structure of x, that can be possibly retained for future comparisons. Otherwise, the s_y solution is discarded and the Active status of enterprise x is restored.

Finally a Revise-SecondStep phase can be executed (see Algorithm 4), during which the Δ Indexes value is calculated again between the problem description of x and the problem
Algorithm 2 Retrieve-reuse.

Require: $x, C$
Ensure: $s_x$

GreenArea $\leftarrow \emptyset$
2: YellowArea $\leftarrow \emptyset$
3: RedArea $\leftarrow \emptyset$
4: for $y \in C$ do
5: \hspace{1em} $s_f \leftarrow s_f(n.x.y)$
6: \hspace{1em} $\text{sim} \leftarrow \text{sim}(x, y)$
7: \hspace{1em} $\text{SIM}(x, C) \leftarrow \text{SIM}(x, C) + \text{sim}$
8: end for
9: $\text{OverallSimilarity}_{x,C} \leftarrow \text{SIM}(x, C)/\#C$
10: if $\text{OverallSimilarity}_{x,C} \in [0, \text{Threshold}_1)$ then
\hspace{2em} GreenArea $\leftarrow \text{GreenArea} \cup \{x\}$
11: else if $\text{OverallSimilarity}_{x,C} \in [\text{Threshold}_1, \text{Threshold}_2)$ then
\hspace{2em} YellowArea $\leftarrow \text{YellowArea} \cup \{x\}$
12: end if
13: else if $\text{OverallSimilarity}_{x,C} \in [\text{Threshold}_2, 100]$ then
\hspace{2em} RedArea $\leftarrow \text{RedArea} \cup \{x\}$
14: end if
15: end if
16: for $y \in C$ do
17: \hspace{1em} $\text{LocalSimilarity}_{x,y} \leftarrow \text{sim}(x, y)$
18: \hspace{1em} if $|\text{LocalSimilarity}_{x,y} - \text{OverallSimilarity}_{x,C}| < \epsilon$ then
\hspace{2em} $s_x \leftarrow s_y$
19: \hspace{1em} end if
20: end for
21: return $s_x, y$

Algorithm 3 Revise-FirstStep.

Require: $x, s_x, y$
Ensure: $s_x$

BadIndex $\leftarrow 0$
2: for $v_i(x) \in d_x \land v_i(y) \in d_y$ do
3: \hspace{1em} if $v_i(x) - v_i(y) < (or >) 0$ then
\hspace{2em} BadIndex $\leftarrow \text{BadIndex} + 1$
4: \hspace{1em} end if
5: end for
6: if BadIndex $< \text{Threshold}_3$ then
\hspace{2em} $s_x \leftarrow \text{Active}$
7: end if
8: return $s_x$

Algorithm 4 Revise-SecondStep.

Require: $x, s_x, C$
Ensure: $s_x$

MaximumSimilarity $\leftarrow -1$
2: MostSimilarCase $\leftarrow x$
3: for $y \in C$ do
4: \hspace{1em} $\text{LocalSimilarity}_{x,y} \leftarrow \text{sim}(x, y)$
5: \hspace{1em} if MaximumSimilarity $< \text{LocalSimilarity}_{x,y}$ then
\hspace{2em} MaximumSimilarity $\leftarrow \text{LocalSimilarity}_{x,y}$
\hspace{2em} MostSimilarCase $\leftarrow y$
6: end if
7: end for
8: BadIndex $\leftarrow 0$
9: for $v_i(x) \in d_x \land v_i(\text{MostSimilarCase}) \in d_{\text{MostSimilarCase}}$ do
10: \hspace{1em} if $v_i(x) - v_i(\text{MostSimilarCase}) < (or >) 0$ then
\hspace{2em} BadIndex $\leftarrow \text{BadIndex} + 1$
11: \hspace{1em} end if
12: end for
13: if BadIndex $< \text{Threshold}_4$ then
\hspace{2em} $s_x \leftarrow \text{Active}$
14: end if
15: return $s_x$

The description of the most similar past case according to Local similarity. If $\Delta$ Indexes analysis is coherent with the previous one, the reused solution will be confirmed, otherwise it would be discarded. The following section shows how this model works in practice.

4. Case study

Our sample was constructed from 2013 income statement and balance sheet information stored in the AIDA database, a Bureau Van Dijk (BvD) database containing demographic information, geographical location, industry, financial statement, balance sheet and assessment (at least five years) of more than one million of Italian enterprises, mainly limited companies.

4.1. Case base definition

As for the 14218 Monza and Brianza enterprises, we have extracted demographic information, legal status, balance sheet, income statement and any information about bankruptcy procedures activated in the considered year. This original dataset was divided based on the legal status of the enterprises into two distinct subsets: non-bankrupt enterprises and bankrupt enterprises. In order to select the subset of bankrupt enterprises, we have considered the following definition of bankruptcy: bankruptcy occurs when a company files a formal legal document in a federal district court for the purpose of either liquidation or reorganization (Altman, 1968). Thus, the bankrupt firms in this research have triggered a bankruptcy procedure prescribed by Italian law in 2012. On the other hand, non-bankrupt enterprises are firms that have not started any procedure and therefore they are considered healthy within industry.

The two subsets have been cleaned of all enterprises with at least one balance sheet item not credible, to avoid those possible errors in the data sources affect the case base creation as well as the reasoning process. The rule adopted was discarding enterprises in case one of the following indexes (that should be positive or equal to zero) negativity: sales, subscribed capital unpaid, total capital assets, total tangible assets, total intangible assets, total financial assets, total current assets, inventories, credits, credits within 12 months, credits over 12 months, cash and bank balance, accrual, net equity, corporation stock, risk and charge fund, indemnities, total debts, total debts within 12 months, total debts over 12 months, accrued liability.

As a result, the number of non-bankrupt firms included in the new clean subgroup is 11637 enterprises. On the other hand, firms that have triggered a bankruptcy procedure prescribed by Italian law in 2012 are 807.

4.2. Case structure definition

For the correct implementation of Case-Based Reasoning model, it is necessary to determine the economic-financial ratios (i.e. balance indexes in Fig. 1) able to correctly specify the case base.

Through extensive literature review on bankruptcy prediction and thanks to Merwin (1942), who showed that failing firms exhibit significantly different ratios with respect to successful firms, 29 economic-financial ratios were identified as the most significant for predicting corporate crisis, according to Dimitras, Zanakis, and Zopounidis (1996).

Afterwards, to strengthen the analysis we have tried to both consolidate and turn into suitable form the enterprises value of each economic-financial ratios, because the value of economic-financial ratios were distributed on a too wide and inhomogeneous value range. The indexes have been normalized in the range [0, 100] by means of the min-max technique, as presented in Hanit and Kamber (2000).
Then, data were processed through the correlation analysis, the multicollinearity analysis and the stepwise logistic regression techniques. The correlation analysis let us highlight those variables that could best predict enterprises bankruptcy: we selected only those economic-financial ratios with both the highest correlation levels with respect to the legal status, and also the lowest correlation levels with respect to the economic-financial ratios themselves (Gordini, 2014). Furthermore, as suggested by Pompe and Bilderbeek (2005) and Gordini (2014) we carried out the multicollinearity analysis using the Variance Inflation Factor (VIF) method, in order to exclude the economic-financial ratios with a VIF value of above 3 (Leow & Mues, 2012). Finally, in order to select the final economic-financial ratios, the stepwise logistic regression analysis (see Fig. 2) was carried out using as dependent variable a dummy variable equal to 1 for bankrupt enterprises and to 0 for non-bankrupt enterprises and as independent variables the economic-financial ratios with a VIF value less than 3 (Etemadi, Rostamy, & Dehkordi, 2009; Gordini, 2014).

These processes enabled us to reduce the economic-financial ratios to the following six features from the initial set of 29 indexes: *Current liability coverage ratio* (CLCR in the following), *Debt/Equity ratio* (DER), *EBITDA to Sales* (EBITDA), *Working capital to total assets* (WCTA), *Value added to Total assets* (VATA) and *Net income to total assets* (NITA). In order to make the analysis more robust, six independent t-tests have been implemented to verify the conclusions. For each index, two vectors of 580 elements have been prepared, the first one containing the value of the current economic-financial ratio for each enterprise, the latter a variable equal to 1 if the current value indicated bankruptcy, 0 otherwise. Table 1 shows the results, confirming the indexes significance with a 99% probability for all the indexes except DER and EBITDA, that are acceptable at 95%.

These six indexes have been exploited to build up the case structure (see Appendix A for its description in XML code) in CREPERIE and allow the comparison of unsolved cases with the case base.

<table>
<thead>
<tr>
<th>Table 1</th>
<th>A t-test analysis of feature selection process in the case study.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Econ-fin. Ratio</td>
<td>t</td>
</tr>
<tr>
<td>CLCR</td>
<td>15.4283</td>
</tr>
<tr>
<td>DER</td>
<td>2.5315</td>
</tr>
<tr>
<td>EBITDA</td>
<td>2.4832</td>
</tr>
<tr>
<td>WCTA</td>
<td>8.2493</td>
</tr>
<tr>
<td>VATA</td>
<td>5.4087</td>
</tr>
<tr>
<td>NITA</td>
<td>6.7552</td>
</tr>
</tbody>
</table>

The case description is a hierarchy delimited by the (graph)/(graph) tags. Each node is characterized by an id, that allows to identify that node within the hierarchy, a mode tag to specify its belongingness to description, solution or outcome part of the case structure, a type specifying if the node is a leaf (i.e. outer node) or not (i.e. inner), a value, which will contain the attribute value when a case would be instantiated and a sim tag to define which Local similarity function will be adopted to compare the node with similar ones. Finally, the dou.min.sim. dou.max.sim and weight define the range (i.e. [0.0, 100.0]) and weights applied in the calculus of similarity between two nodes. At the current state of implementation, all the weights are equal to 1 because we preferred to analyze the CBR performance considering the contribution of all the indexes equivalent.

During the case base construction process, many instances of the structure above are created: the solved cases are completed with feature values (that are inserted into the related tag value field) together with the solution part, that is a node indicating the current legal status of the enterprise, i.e. Inactive, Failed or Liquidation as introduced above.

The XML schema takes care of case solutions too (see Appendix A), where the legal status of the enterprise is represented. The solution part is completed also by the definition of Δ Indexes values: since they depend on the retrieved case problem description, their definition and use will be introduced in the next section. Finally, the case structure of solved cases can be completed by an optional outcome node, that is useful (at current state of implementation) to insert comments about the solution adoption. The solution part of unsolved cases is inserted into the structure at the end of the retrieval-reuse steps.

### 4.3. Indexing, retrieval, reuse and revise

During the indexing step, an unsolved case x is compared with the whole case base according to the Overall similarity definition: in this way, a currently Active enterprise is clustered into one of the following three areas:

- **RED area**, if the Overall similarity value is in the range of 75% or higher;
- **YELLOW area**, if the Overall similarity value is between 65% and 75%;
- **GREEN area**, if the Overall similarity value is 65% or lower.

The areas boundaries are identified on the basis of Eq. (6). RED area is limited by the highest level of similarity, i.e. 1, and

![Threshold calculation](https://creperie.com/threshold.png)

since this region will contain Active firms characterized by at least 4 indexes equivalent to the case base, and \( # = 6 \) given that the balance indexes included in the case description are six. Solving the equation, \( Threshold_2 = \frac{\bar{X} + \sigma}{2} = 0.75 \).

---

3 Omissis in the node description contain information for the graphical user interface.
YELLOW area's limits are Threshold\textsubscript{L} and Threshold\textsubscript{H}, being characterized by firms with at least 3 indexes equivalent to the case base. The lower bound of this region is obtained solving the following equation

\[
\text{Threshold}_3 = \frac{\sum_{i=1}^{3} \left( 5.6 - i \right)}{3} = \frac{5 + 4 + 3}{3} = 0.67
\]

Then, Threshold\textsubscript{H} value has been cut off to 0.65 to better characterize YELLOW area according to the unsolved cases set used in the case study.

Finally, GREEN area is characterized by enterprises with less than three indexes equivalent to the case base, thus the Overall Similarity value of its members will be between 0 and Threshold\textsubscript{L}.

Once \( x \) has been clustered into one of the three areas above, it can be compared with a specific solved case \( y \) on the basis of the Local similarity notion. In this way, a solution for the case \( x \) can be proposed, reusing the \( y \) one. Since the current case \( x \) is about an Active enterprise, the reused solution is useful to understand which kind of future behavior it could assume, possibly changing its status to Inactive, Failed or Liquidation. The \( \Delta(x, y) \) value is taken as a suggestion of where acting before the firm situation becomes critical, in particular: \( \delta_{\text{CLCR}}(x, y) \) should be negative; \( \delta_{\text{DER}}(x, y), \delta_{\text{EBITDA}}(x, y), \delta_{\text{WCTA}}(x, y), \delta_{\text{VATA}}(x, y) \) and \( \delta_{\text{NITA}}(x, y) \) should be positive.

To make clearer this point, let’s consider the following running of Algorithm 1. Table 2 contains three enterprises extracted from GREEN, YELLOW and RED areas at the end of indexing phase. Each of them has been associated to a solution, at the end of the reuse phase. Table 3 reports the results of retrieval and reuse steps on the basis of Eq. (4). For each case \( A, F \) and \( K \), the retrieved cases attributes are indicated. The retrieved cases are called \( A_r, F_r \) and \( K_r \) respectively. Moreover, \( \Delta(x, y) \) calculus is presented for each \((x, y)\) pair, with \( x \in \{A, F, K\} \) and \( y \in \{A_r, F_r, K_r\} \).

The \( A_r \) case is classified as Liquidation: reusing this solution would be opposite to the case A classification in the GREEN area. Thus, a Revise-FirstStep algorithm can be executed: five indexes are coherent with the indexing result, with the only exception of \( \delta_{\text{WCTA}} \) value, that is negative instead of positive. This anomalous situation is indeed due to the different business sector where case \( A \) and the retrieved one operate. The analysis of \( \Delta(A, A_r) \) fully confirms the clustering of enterprise A in the GREEN area: the \( A_r \) solution will be discarded and the enterprise A will maintain its Active status.

The situation of enterprise F as described in Table 3 seems to be critical: the retrieved case \( F_r \) is classified Inactive and \( \Delta(F, F_r) \) shows that only the VATA index is good. The null value of \( \delta_{\text{CLCR}}(F, F_r) \) demonstrates that the Active firm should not ensure financial stability in the next future. The negative value of \( \delta_{\text{DER}}(F, F_r) \) indicates the higher prevalence of third-party funding sources for the Active firm than the Inactive one; the retrieved case \( F_r \) is better than \( F \) from the \( \delta_{\text{WCTA}}(F, F_r) \) perspective too. Anomalous situations are detected in \( \delta_{\text{EBITDA}}(F, F_r) \) and \( \delta_{\text{NITA}}(F, F_r) \), since source indexes are both negative, but the Active enterprise values are more negative than the Inactive ones: for this reason, the calculus of \( \delta \) is positive, but the firm \( F \) has to keep under control its EBITDA and NITA ratios. The firm \( F \) is really close to bankruptcy, and the Inactive solution coming from the retrieved case is then reused.

Finally, the \( \Delta(K, K_r) \) analysis in Table 3 for the case \( K \) and the related \( K_r \) retrieved case shows that three \( \delta \) values are bad: in particular, \( \delta_{\text{CLCR}}(K, K_r) \) and \( \delta_{\text{EBITDA}}(K, K_r) \) are really impressive, since \( K \) indexes are worse than \( K_r \) ones, while \( \delta_{\text{WCTA}}(K, K_r) \) can be considered slightly adverse. The retrieved case Inactive solution can be reused, but many doubts could emerge from this kind of decision. Then, a Revise-SecondStep algorithm can be invoked to verify the hypotheses: the most similar case to the enterprise \( K \), namely \( K' \), for which \( \text{sim}(K, K') = 83 \%), is then retrieved and \( \Delta(K, K') \) is evaluated, as presented in Table 4.

By comparison with the most similar case, only \( \delta_{\text{CLCR}}(K, K') \) index is bad. For this reason the reused Inactive solution can be reused to the current Active status; of course, it will be necessary keeping under control the financial situation in the next future, in order to avoid the critical situation forecasted by the retrieve-reuse steps would actually arise.

<table>
<thead>
<tr>
<th>Table 2</th>
<th>The three enterprises from GREEN, YELLOW and RED area respectively: the table reports the overall similarity value too.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Enterprise</td>
<td>Overall similarity</td>
</tr>
<tr>
<td>A</td>
<td>57.4%</td>
</tr>
<tr>
<td>K</td>
<td>71.22%</td>
</tr>
<tr>
<td>F</td>
<td>79.8%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 3</th>
<th>The reused solution for the enterprise A, F and K from Table 2.</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>A_r</td>
</tr>
<tr>
<td>----------------</td>
<td>----------------</td>
</tr>
<tr>
<td>CLCR</td>
<td>0.50</td>
</tr>
<tr>
<td>DER</td>
<td>1.17</td>
</tr>
<tr>
<td>EBITDA</td>
<td>24.86</td>
</tr>
<tr>
<td>WCTA</td>
<td>0.02</td>
</tr>
<tr>
<td>VATA</td>
<td>0.17</td>
</tr>
<tr>
<td>NITA</td>
<td>0.32</td>
</tr>
</tbody>
</table>

5. Results and discussion

The final training sets were composed in the following way: we built up a case base \( C \) containing data about 580 bankrupt enterprises, by removing the firms with at least one ratio tending to infinity (and therefore not comparable in our CBR cycle) from the initial set, and an unsolved case base UC containing data about 580 non-bankrupt enterprises. The UC elements were chosen applying the nonprobability sampling method (Trolle & Molteni, 2003) on the initial set, in order to have comparable \( C \) and UC sets from the numerosity point of view, as shown in Park and Han (2002). The \( C \) and UC sets have been compared according to the Overall similarity formula. The results are shown in Fig. 3.

The CREPERIE platform has divided, in the indexing step, the 580 Active enterprises into three subsets, composed of 71 (GREEN area), 440 (YELLOW area) and 69 (RED area) elements respectively.
the firms with higher similarity (highlighted by a solid rectangle in Fig. 3) should be close to bankruptcy, the ones with lower similarity (highlighted by a pointed rectangle in Fig. 3) should be completely healthy. In order to evaluate the CBR performance, the ranking rating assigned them by the Z-score Model (Altman & Hottchkiss, 2006) has been considered; this rating is very useful in forecasting Italian enterprises’ bankruptcy, as shown in Altman et al. (2013), and divides companies into three main categories: Safe Zone (ratings AAA, AA+, AA, AA-, A, A-, BBB+, BBB), Grey Zone (ratings BBB-, BB+, BB, BB-, B+), Distress Zone (ratings B-, CCC+, CCC, CCC–, D).

The first line of Table 5 shows the correspondence between CBR results and Z-score ratings, after the revision step, according to the real state of the enterprise as extracted from 2014 income statement and balance sheet information, stored in the AIDA database: the CBR is better than Z-score in predicting the future situation for GREEN and YELLOW areas firms, while seems to be significantly worse in forecasting future critical situations (64 errors with respect to only 7 mistakes made by Z-score).

Anyway, the Overall similarity calculus is only a part of the whole CREPERIE CBR cycle: in order to reduce the error percentage in RED area definition, we could try to apply a revise step by means of Δ analysis described in Section 4.

First, the balance indexes of each misleading enterprise $E_i$, with $i \in [1..64]$ have been compared with the ones of the most similar failed firm $E_j$ on the basis of Local similarity calculus, using the Overall similarity as a threshold; then, the current enterprise $E_i$ has been put in GREEN area if $\Delta(E_i, E_j)$ contained at least four well defined $\delta(i, E_i, E_j)$ (see Section 4.3), in YELLOW area if the number of well defined $\delta(i, E_i, E_j)$ was exactly equal to 3, in RED area otherwise. The revise step has allowed to redistribute 49 of the previously misleading 64 enterprises, reducing to 15 the number of incorrectly classified firms in RED area: 25 of them have been addressed to GREEN area, and 22 to YELLOW area. A further analysis on them has allowed to conclude that 27 of those 49 firms were correctly classified and 20 were not (4 in the GREEN area and 16 in the YELLOW area). Third line in Table 5 summarizes the CREPERIE performance after this first revision round (note that the number of correct enterprises in RED area in the CBR row moves from 5 to 7 to include two enterprises closed due to properties’ will).

A second revise step (see fourth line in Table 5) has been then conducted on the 22 enterprises erroneously clustered in GREEN area at the end of the RED area revise: 11 of them have been correctly classified in the RED area, moving the total to 18. Finally, the 121 errors made in YELLOW area definition have been analyzed. We were able to execute the revise step on 95 of them, since data about the remaining 26 were lost passing from 2013 to 2014 AIDA databases; 79 enterprise moved from YELLOW area to GREEN or RED ones: 57 of them were correctly clustered as GREEN firms, 8 of them in the RED area, while 14 Delta calculus sessions brought to mistakes.

To highlight the benefits of revise steps based on Δ analysis, a Student’s t-test between CBR and Z-score results has been accomplished. Table 6 shows the results of three different tests, executed on the results obtained from CBR in three different sessions. The conclusions are the same: the retrieval step is not sufficient to forecast bankruptcy, but the implementation of a complete CBR cycle improves significantly the method capabilities.

The possibility of revising retrieval result is the most crucial improvement introduced by CREPERIE with respect to past CBR approaches: doing so, the extremely high percentage error obtained in the first step can be significantly reduced to be comparable with traditional methods like Z-score, as shown in Table 7. The CBR error percentage decreases from the initial value of 32.24% to the 14.13% final one, increasing the accuracy level of our approach from the initial 68% rate to the final 86% one. Furthermore, in order to evaluate the misclassification costs of the bankruptcy prediction models, the analysis of both the Type I error (when a bankrupt firm is classified as a non-bankrupt one) and Type II error (when a non-bankrupt firm is consider as a bankrupt one) was carried out (Dopuch, Holthausen, & Leftwich, 1987; Koh, 1992; Nanda & Pendharkar, 2001). In general terms and according to some authors (Etheridge, Sriram, & Hsu, 2000; Watts & Zimmerman, 1986), it seems to be worse to commit a Type I error than a Type II error, but in a SME context, characterized by low level of cash flow and liquidity, the Type II error seems to lead to much higher

**Fig. 3.** Overall similarity calculated by CREPERIE: Active enterprises are clustered into three groups, namely RED, GREEN and YELLOW areas.

**Table 5** Comparison between CREPERIE and Z-score performance: for each zone defined by Overall Similarity, the number of correct (YES columns) and incorrect (NO columns) predictions is shown.

<table>
<thead>
<tr>
<th></th>
<th>GREEN</th>
<th>YELLOW</th>
<th>RED</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>CBR-retrieve</td>
<td>YES 53</td>
<td>NO 18</td>
<td>YES 105</td>
<td>NO 5</td>
</tr>
<tr>
<td>Z-score</td>
<td>YES 43</td>
<td>NO 28</td>
<td>YES 145</td>
<td>NO 7</td>
</tr>
<tr>
<td>Revise RED area</td>
<td>YES 74</td>
<td>NO 11</td>
<td>YES 121</td>
<td>NO 15</td>
</tr>
<tr>
<td>Revise GREEN area</td>
<td>YES 146</td>
<td>NO 15</td>
<td>YES 131</td>
<td>NO 16</td>
</tr>
<tr>
<td>Revise YELLOW area</td>
<td>YES 131</td>
<td>NO 15</td>
<td>YES 142</td>
<td>NO 26</td>
</tr>
</tbody>
</table>

**Table 6** Effects of revise step on RED, GREEN and YELLOW areas: the comparison between CBR and Z-score.

<table>
<thead>
<tr>
<th></th>
<th>$t$</th>
<th>$p(\alpha)$</th>
<th>$\alpha = 0.01$</th>
<th>$\alpha = 0.05$</th>
<th>$\alpha = 0.25$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Retrieve</td>
<td>0.4416</td>
<td>0.6589</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Z-score</td>
<td>1.1587</td>
<td>0.2468</td>
<td>–</td>
<td>–</td>
<td>X</td>
</tr>
<tr>
<td>Revise GREEN area</td>
<td>1.8868</td>
<td>0.0594</td>
<td>–</td>
<td>–</td>
<td>X</td>
</tr>
<tr>
<td>Revise YELLOW area</td>
<td>5.2464</td>
<td>&lt; 0.001</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
</tbody>
</table>

**Table 7** A comparison between CBR and Z’ accuracy.

<table>
<thead>
<tr>
<th></th>
<th>Total Er.</th>
<th>I Type Er.</th>
<th>II Type Er.</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Retrieve</td>
<td>187</td>
<td>51</td>
<td>136</td>
<td>68%</td>
</tr>
<tr>
<td>Z-score</td>
<td>180</td>
<td>35</td>
<td>145</td>
<td>69%</td>
</tr>
<tr>
<td>Revise</td>
<td>82</td>
<td>29</td>
<td>53</td>
<td>86%</td>
</tr>
</tbody>
</table>
miscategorization costs (Gordini, 2014). This because Type II error can leads especially financial institutions to make inappropriate lending decisions, reducing the SMEs possibility to have access to credit, bringing them to bankruptcy and, finally, damaging their brand image towards their stakeholders.

Another interesting feature of the CRePERIE revise step is its development according to substitutional approaches (Manzoni et al., 2007), avoiding hybrid solutions (or null adaptation) adopted in the past. None of the previous approaches exploited pure CBR to forecast enterprise default, and none of them developed a complete 4R’s cycle in their evaluation. Finally, the user (i.e. the enterprise management) can obtain a clear and immediate view of the balance indexes status, being possible to exploit the Δ analysis for anomalies correction.

6. Conclusion

Prediction of an enterprise bankruptcy is a crucial task and a very important topic not only for firms, but also for many practitioners, stakeholders, governments and researchers.

In this paper, an approach to bankruptcy prediction based on CBR has been presented. This research is still on-going: although initial results seem to show the applicability of the approach, further work is necessary to change the traditional literature point of view on this.

According to the third conclusion by Bellovary et al. (2007), we have applied the CBR approach on a very reduced set of indexes to characterize firms. Doing so, we have tested the characterizations of the Δ analysis in the revise algorithm definition. Moreover, we have implemented a complete, pure CBR cycle where the adaptation step is more significant than the retrieval one. With respect to other methods proper of Artificial Intelligence, CBR pros are the objectivity of determinations and the possibility to suggest corrections not based on any knowledge model difficult to build.

In this sense, it could be preferred to more traditional AI techniques like ANNs, that have difficulties in explaining the prediction results due to the lack of explanatory power, and suffer from the generalization point of view because of overfitting phenomena as well as time-consuming efforts to construct a suitable architecture (Ahn & Kim, 2009; Lawrence, Giles, & Tsoi, 1997; Sarle. 1995). Moreover, different from them, our approach allows not only to discriminate between bankrupt and non-bankrupt enterprise, but also to identify the most probable causes of firms critical situations, in terms of bad economic-financial ratios, and suggest how to solve them.

First results are satisfying: the Overall similarity metric demonstrated good indexing capabilities if compared with the Z*-Score. This was a crucial step for us, since the Italian characterisation of bankrupt and non-bankrupt enterprises is mainly based on Z*-Score. Moreover, it is important to notice how the indexing step is only the first point in our approach: the most interesting aspect is the possibility to exploit the indexing results to retrieve the most similar bankrupt firm from the Δ indexes point of view and reuse its current legal status, possibly revising it. These three phases are conducted according to a pure CBR approach, with no need to use other methods like in the previous works. Nevertheless, further investigation is necessary to improve the model accuracy, given that the classification of enterprises in RED area is still not completely effective.

These findings have several managerial implications for firms, financial institutions, and project financing, such as public procurement, public concessions, tender notices, and European projects. On one hand, for firms the model presented could be useful to assess their real bankruptcy likelihood and to make proper strategy improvements. In particular, SMEs that rely mainly on external liquidity and funding have to better understand their real status in order to have easily access to credit. Obviously, a firm classified in the GREEN area will have a higher probability to have access to credit. On the contrary, an enterprise classified in the RED area will have a lower probability to have access to credit than others. In this latter case, the Δ analysis and revise step of our model can support both the entrepreneurs and the managers either in bringing proper improvements or in adopting new business strategies in order to avoid the occurrence of bankruptcy status and of many difficulties in obtaining credit. As regards to large enterprises, our model could be used to better understand in which subsidiaries or affiliates and in which projects to invest in. This decision can be supported by the proposed model thanks to the realization of a matrix that correlates the bankruptcy likelihood with the profitability of the investment.

On the other hand, for financial institutions an effective bankruptcy prediction model is essential in order to make appropriate lending decisions. In fact, reliable models are crucial to assess the real status and select healthy businesses to collaborate with and invest in, and eventually support entrepreneurs to avoid bankruptcy situations and, then, have access to credit. Furthermore, given the flexibility and the accuracy of the proposed CBR model, financial institutions could build their own case base and select their set of indicators in order to build up an ad-hoc bankruptcy prediction model. To do this, financial institutions should improve their customer information systems and increase the transparency of their assessment techniques. Finally, from the project financing perspective, the proposed model might be useful in assessing the financial and market reliability of the firms to invest in, possibly using international benchmarks as thresholds.

We believe that our work could be generalized in the future. Anyway, as introduced at the beginning, we need to deepen the research from many perspectives before being sure of this. As a first step, we will implement soon an indexing phase based on Overall similarity taking into account the business industry attribute too, as presented in Mazzucchelli and Sartori (2014): we think this innovation would significantly improve the performance of the method in clusters creation. This improvement could also help us in taking under control the complexity of retrieve-reuse step of our algorithm. Given that such complexity depends on the number of cases in the case base, the exclusion of enterprises not similar from the business industry point of view will reduce the time necessary to get results. Then, we will further analyze the Overall similarity performance with respect to other known methods as presented in Tsai, Hsu, and Yen (2014).

Moreover, we are planning to implement new definitions of Overall similarity in order to increase the ranking capability of CRePERIE: results are very compressed at the current state of development, probably due to the Overall similarity definition as the average value of Local similarity results. It could be useful to modify the similarity metric through an in-depth analysis of case base nature, in order both to weight the features of the case description in the calculus, and to understand if some Local similarities should be considered more important than others.

Another improvement could drive the analysis towards the use of methods to assess the relationship, either positive or negative, between the legal status and the economic-financial ratios, to improve the current Δ set method. Doing so, the importance and the impact of economic-financial ratios on firms performance and status could be taken into account, calculating opportune ratios weights w_i depending on the case base nature and/or the business industry attribute. Finally, it could be really interesting to extend our model to also include corporate governance indicators, as a key ratios in predicting bankruptcy.
Appendix A. Case structure XML schema

The following code is an abstract of the XML schema adopted by CREPERIE to represent the case structure:

A1. Case description

```xml
<graph>
    <node id="1" name="ENTERPRISE ID" node="description" type="inner" value="" sim="String Case Sensitive" />
    <att name="weight" value="1"/>
</node>

<node id="2" name="DER" node="description" type="outer" value="" sim="Interval Double" />
    <att name="dou_min_sim" value="0.0"/>
    <att name="dou_max_sim" value="100.0"/>
    <att name="weight" value="1"/>
</node>

<node id="3" name="CLCR" node="description" type="outer" value="" sim="Interval Double" />
    <att name="dou_min_sim" value="0.0"/>
    <att name="dou_max_sim" value="100.0"/>
    <att name="weight" value="1"/>
</node>

<node id="4" name="ERIDTA" node="description" type="outer" value="" sim="Interval Double" />
    <att name="dou_min_sim" value="0.0"/>
    <att name="dou_max_sim" value="100.0"/>
    <att name="weight" value="1"/>
</node>

<node id="5" name="WCTA" node="description" type="outer" value="" sim="Interval Double" />
    <att name="dou_min_sim" value="0.0"/>
    <att name="dou_max_sim" value="100.0"/>
    <att name="weight" value="1"/>
</node>

<node id="6" name="YATA" node="description" type="outer" value="" sim="Interval Double" />
    <att name="dou_min_sim" value="0.0"/>
    <att name="dou_max_sim" value="100.0"/>
    <att name="weight" value="1"/>
</node>
</graph>
```

A2. Case solution

```xml
<node id="8" name="Current Legal Status" node="solution" type="outer" value="INACTIVE" sim="String Case Sensitive" />
    <att name="dou_min_sim" value="0.0"/>
    <att name="dou_max_sim" value="0.0"/>
    <att name="weight" value="0"/>
</node>

<node id="9" name="Comments" node="outcome" type="outer" value="INSERT COMMENTS HERE" sim="String Case Sensitive" />
    <att name="dou_min_sim" value="0.0"/>
    <att name="dou_max_sim" value="0.0"/>
    <att name="weight" value="0"/>
</node>
```

References


