An automated literature analysis on data mining applications to credit risk assessment

Sérgio Moro[1,2], Paulo Cortez[2], and Paulo Rita[1]

Abstract:

This paper presents an automated literature analysis of data mining applications to credit risk assessment, encompassing the recent timeframe of the last five years, from 2010 to 2014. The widespread academic search engine Google Scholar was used for collecting the 100 most relevant articles published in management and information systems conferences and journals and containing the keywords "data mining" and "credit risk". Such set of articles served as a basis for assessing the main trends of research in data mining applications to credit risk, first by using text mining, then through the latent Dirichlet Allocation Algorithm for grouping the articles in logical topics.

Five types of problems in credit risk were assessed: credit scoring, bankruptcy, credit fraud, credit cards and regulatory issues. From these, credit scoring gets the most of the attention, while bankruptcy and credit fraud also encompass a significant number of articles. The most interesting finding is that the most advanced data mining techniques such as support vector machines and ensembles are being applied to credit risk problems more for tuning these techniques than to benefit credit risk assessment. This represents an interesting research gap to be addressed. The trends discovered prove the value of the automated procedure undertaken, which is a novel in credit risk applications. Credit scoring was confirmed as the dominant subject regarding data mining applications. On the other hand, several studies are mostly focused on tuning data mining techniques rather than on showing the benefits achieved by applying such techniques. More focus should be given to the value of data mining to risk assessment. Also, findings suggest that regulatory issues are demanding research in data quality, in alignment with banking regulation leveraged by the global crisis.

JEL classification: C51; C52; C55; G20; G17

[1] Business Research Unit, ISCTE – University Institute of Lisbon
[2] ALGORITMI Research Centre, University of Minho

1. Introduction

Financial crises are catalysts to an outstanding increase in investigation for finding innovative techniques to anticipate such financial distresses, providing a major advantage to act accordingly (Claessens et al., 2014). Therefore, research in financial credit risk has become one of the most prominent and prolific subjects in recent years. In fact, the 2008 global financial crisis arose due to a poor assessment of the risk associated with the various ways through which banks have transferred credit risk in the financial system (Nijskens and Wagner, 2011). Thereafter, contagion took place by spreading through countries worldwide, originating a global systemic risk.

Several studies have shown a large increase in publications related to credit risk after 2009 (Galati and Moessner, 2013; Moro et al., 2015). There are numerous reasons for such event. First, the financial crisis sounded massive alarms which triggered research in two domains: bankruptcy detection for anticipating the impact of defaulting and regulatory reporting for providing a tighter control of financial institutions. Large scale regulatory projects potentially benefiting from advanced machine learning techniques gained relevance in the post-crisis financial market, such as the IRB (Internal Ratings-Based Approach) in Europe (Tobback et al., 2014). The more rigid control on banks also allowed detecting some severe fraud situations, with the institutions being affected by trust loss (Macey, 2012). Other credit risk related domains are credit scoring and research on the specific product of credit cards. While these were widely studied subjects prior to the crisis, such event has also influenced research on these matters, given that the global crisis affected every individual, leading credit institutions to require a tighter control on individual loans.

Data mining (DM) is a concept that encompasses techniques and methodologies for unveiling patterns of knowledge from raw data (Turban et al., 2011). Typical DM projects include data understanding and preparation followed by the application of machine learning algorithms for finding interrelations between data that can be translated into valuable knowledge. The previous steps may involve data sampling and feature selection, depending on the data, and also data quality operations to improve the value of information before it can be used for feeding next steps. The latter steps may include one among several widely studied algorithms, such as decision trees, artificial neural networks and support vector machines, or even an ensemble of a few different algorithms.

Real world problems are often based on collected data from which a new insight is needed for leveraging business. Such problems may be addressed through a data-driven approach including DM. The credit risk domain typically involves an analysis of past history to understand which variables influence behavior of credit holders, making it an excellent subject for applying DM techniques. Being such an interesting applied field of research, a few literature analysis and reviews were published recently on DM applications to credit risk. Marques et al. (2013) conducted a literature review by collecting 56 papers published from 2002 to 2012 on the application of evolutionary computing to credit scoring. Their methodology consisted in a manual analysis by dividing into sub-problems to which evolutionary computing techniques have been applied: classification, variable selection, parameter optimization, and other miscellaneous problems. The conclusions show that variable selection gets most of the attention regarding evolutionary techniques. Guerrero-Baena et al. (2014) evaluated literature in terms of the application of multi-criteria decision making techniques to corporate finance issues during the period 1980-2012, in a total of 347 publications from the Scopus database. The method used was a manual classification of every article and a descriptive statistical analysis over that classification. The results presented by their study show bankruptcy prediction and credit risk assessment as receiving 4.6% and 3.7% of the attention among the total publications. Nevertheless, that paper focused specifically on corporate finance, with capital budgeting receiving the most of the attention (64%), while there exists a high number of publications on DM applications to individual credit risk and default (e.g., Oreski and Oreski, 2014).

Literature analyses can be conducted through automated methods that are able to parse the relevant terms from each publication and then build logical clusters of articles, providing a meaningful structure from which new insights can be obtained. One of the most recent used methods includes Text Mining (TM), such as the work of Delen and Crossland (2008) demonstrates. Furthermore, the latent Dirichlet allocation (LDA) algorithm may be used for organizing the articles in logical topics (Moro et al., 2015). It should be noted that such procedure has not yet been applied to credit risk assessment applications; therefore it would be interesting to understand if the insights achieved by this procedure allow identifying possible research gaps.

This article presents an automated literature analysis over a significant set of articles about DM applications to credit risk assessment. The main highlights are the following:

- Collecting the hundred most relevant articles on credit risk using DM, according to Google Scholar relevance criterion;
- Using articles' keywords for building a lexical dictionary of relevant terms, followed by the application of TM for understanding the subjects which are deserving the most of the attention;
- Applying LDA algorithm for building logical topics of articles, identifiable by the terms that characterize each of these topics.

The next section elucidates on the materials and methods used for the experiments, whereas Section 3 analyzes the results achieved. Finally, Section 4 completes the article with conclusions and final remarks.

2. Materials and methods

2.1. Search criteria

Several academic search engines currently available are widely known to the scientific community, such as Google Scholar (GS), Web of Science (WoS) and Scopus (Harzing, 2013). The former (GS) is a free service that uses web crawlers for retrieving scholarly publications from several sources available on the Internet, while the two latter (WoS and Scopus) are indexing systems for specifically selected sources. Recent articles on this thematic have shown that GS has progressively improved and is already at a maturity stage which enables it to compete with the more traditional source enclosed indexing services (Harzing, 2013; De Winter et al., 2014). Furthermore, GS provides by default search results ordered by relevance, measuring it not only by the number of citations but also considering the full text of each source as well as the source's author and the publication in which the source appeared, according to the Google Scholar website. The usage of GS rank to obtain a certain number of most relevant articles has been used in several relevant reviews, such as Hall (2006) and Tabuenca et al. (2014). Therefore, for the present analysis, GS was the chosen search engine for selecting the most relevant articles on DM applications to credit risk.

The search query included every publication that contained both “credit risk” and “data mining” within the 2010–2014 timeframe, with both the “include patents” and “include citations” unchecked, and leaving the default option of “sort by relevance” for allowing the most relevant hits to be presented at the top pages. Also, it should be stressed that only English written journals were considered. While leaving a large number of publications out of the present scope, as Englishis the major research dissemination language (Crystal, 2012). Besides, the proposed automated approach would not be viable considering the need to fully translate each article to a common language and the fact that most human languages have an intrinsic subjectivity, hence usually a direct word to word does not encompass this subjectivity (Hatim and Mason, 2014). The search was executed on May 16, 2015, resulting in 2970 hits at that date. Then, articles began to be collected starting at the top of the first page, benefiting from the sort by relevance of GS. It should be noted that only journal and conference articles were included, leaving behind books, book chapters, presentations and other sorts of material. From May 16 to 18, 2015 articles were collected one by one, and each one of them was evaluated to validate if it matched the subjects in analysis, for including only those that matched. This process stopped when the number of valid articles for the present study reached one hundred. For that, 116 articles needed to be evaluated, implying that sixteen of those, most of them in the latter pages from the search results list, were not relevant for the present analysis, thus were excluded. Hence, above a certain threshold of articles it would become less likely finding relevant articles. Therefore, the next steps of this literature analysis were performed on the one hundred most relevant articles in DM applications to credit risk, according to GS.

2.2. Text mining

Text mining (TM) allows extracting knowledge from unstructured data such as a collection of texts (Fan et al., 2006). Therefore, it can be useful in the process of analyzing a large number of literature publications, providing an automated mean of summarizing the corresponding contents. Previous works followed two distinct approaches: by extracting all the lexical words, excluding only the more common words such as pronouns (Delen and Crossland, 2008); or by using specific dictionaries of terms composed of one or more words (Soper and Turel, 2012) defined by experts in the studied subjects (Moro et al., 2015). The current work followed the latter approach, with a significant enhancement: instead of asking the assistance of experts to define an unguided dictionary, all the keywords for the one hundred articles were collected, resulting in a list with 466 words. Then all duplicates were removed, and similar words in different formats (i.e., singular versus plural) were also reduced to a single word. Also, common or too generic terms such as “banking” or “data mining” were removed, considering this literature analysis focus on the specific DM methods and techniques applied to a subset of problems within credit risk. Finally, the dictionary of equivalent terms was built with the remaining 111 terms, as displayed in Table 1 (similar terms in different formats such as singular versus plural are not shown, for simplification purposes).

The table is logically divided in two areas: the first for the specific DM methods, and the last for the sub-problems within credit risk. Also, abbreviations were included when these do not have a meaning in the English language (to avoid a mismatch), considering most articles use them. It should be emphasized that the procedure of using articles’ keywords is less prone to the subjectivity associated with human experts’ definition of relevant terms.

For the experiments, the R statistical environment was adopted, considering it has the advantage of being open source with the support of a large community, providing a vast number of packages in a wide scope of applications. Moreover, the “tm” package was chosen for the TM functions, and the “wordcloud” package for generating visually appealing word clouds, with a few other packages also included for supporting auxiliary functions.

Using the dictionary and the R packages, the procedure adopted can be summarized as follows:

1. Create a corpus of documents (i.e., articles);
2. Remove extra spaces and convert all words in lower case for simplifying term matching;
3. Apply a transformation to convert every equivalent term in the dictionary in a unique term;
4. Build the document term matrix.

Table 1 – Dictionary of terms (in lower case)

Term	**Equivalents**
Data mining methods and techniques	
anfis	adaptive neuro-fuzzy inference system, adaptive network-based fuzzy inference system
bagging	bootstrap aggregating
bayesian	
case-based reasoning	cba
clustering	self-organizing map, k-nearest neighbor
data quality	information quality
decision support system	dss, expert systems
decision tree	dt, random forest, rotation forest, chaid
discriminant analysis	cpda,lpda
ensemble	
feature selection	filtering, variable selection
genetic algorithm	ga
hybrid	
logistic regression	lr
multiple criteria	mcdm
neural network	nn, ann, multilayer perceptrons, mlp
particle swarm	
rough set	set theory, fuzzy sets
sampling	sample selection, random subspace
support vector machine	svm
Credit risk sub-problems	
bankruptcy	insolvency, default detection, early warning, financial distress
credit card	
fraud	money laundering
regulatory	loss given default, probability of default, irb, internal ratings-based, basel
scoring	credit risk classification, rating

The document term matrix is a bi-dimensional matrix that counts the frequency that each term (in columns) occurs in each of the documents (in rows). Such structure is of paramount relevance for TM, considering it is the basic input for constructing easy to interpret structures such as a table of frequencies and a word cloud.

2.3. Topics of articles

A simple TM consisting in word counting and displaying summarized information about the contents of documents may be interesting by its own for providing some insights. However, it is more compelling to build a body of knowledge using clustering DM techniques for unveiling previously unknown trends that may enrich understanding on a given subject. The main goal is to group articles in logical clusters which may be characterized by some common denominators. For this task, the latent Dirichlet allocation (LDA) algorithm was adopted for both its simplicity and effectiveness given the large amount of publications with interesting results using this technique (Campbell et al., 2014).

LDA is the most popular topic-analysis method, with applications in a wide range of domains. It produces a weighted list of topics for every document in a collection dependent on the properties of the whole (Campbell et al., 2014). The number of topics is a needed input for LDA. Following similar approaches (Delen and Crossland, 2008; Moro et al., 2015), this value was set to half of the terms considered. Each word w_j has a probability of matching a given topic z_i given by Equation 1.

$$\beta_{ij} = p(w_j=1|z_i=1);$$

Equation 1 – β distribution

Thus the LDA computes a bi-dimensional probability matrix β of k topics versus V different words. Therefore β provides a simple metric for measuring the relation of each term to a given topic. A value closer to zero indicates a stronger relation to that topic (Blei et al., 2003). For the experiments presented next section, the R package “topicmodels” was adopted, since it can be fed directly with the document term matrix produced from the “tm” package, facilitating the procedure.

The LDA output is a tridimensional matrix encompassing terms, documents and topics built by the algorithm. Thus, for every topic it is possible to obtain a measure of its relationship to one of the dictionary terms through the β distribution. Also, for every document it is possible to check to which topic it suits better. Considering the 25 reduced terms defined in Table 1, the one hundred articles, and the thirteen topics, that imply a structure containing 32.500 values. Since the goal is to analyze the groups of articles represented by the topics and its characterization and more specifically different DM approaches to credit risk problems, for each topic only the most relevant credit risk problem and the most relevant DM method (as measured by the β distribution) are scrutinized.

2.4. Proposed approach

This section underlines the whole approach undertaken, drawn on the procedures described in both Sections 2.2 and 2.3. Such approach is illustrated in Figure 1.

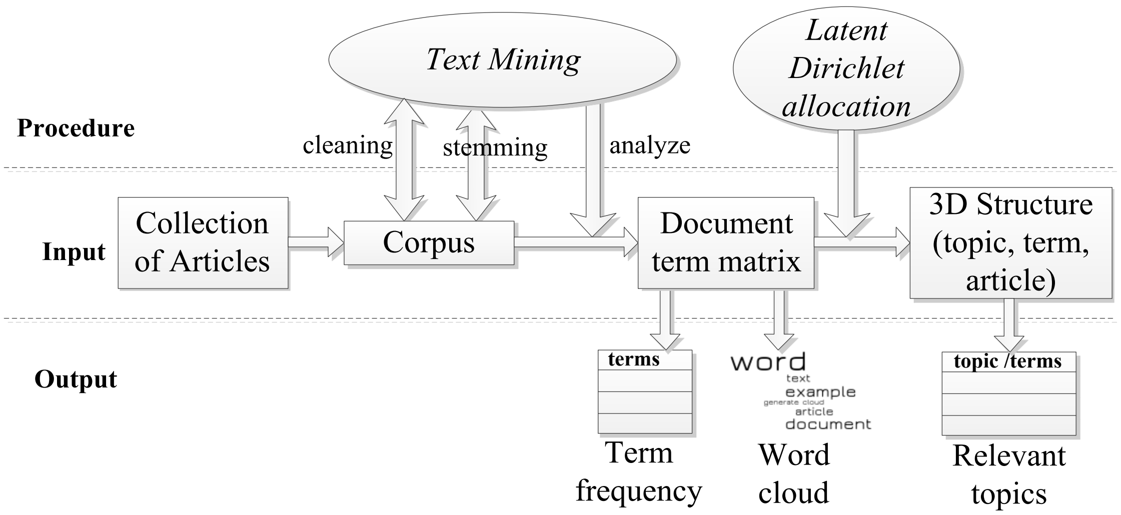

Figure 1 - Proposed approach

The input is the set of articles collected through the methods explained in Section 2.1. From each article, the title, abstract, keywords and body of text are retained, discarding images and also references section, the latter for avoiding including terms that appear only in the titles of each reference. These articles constitute the corpus of documents used for the text mining procedures. The text mining procedure then takes place over the corpus by preparing and analyzing the contents of each document. Code excerpt presented next illustrates such steps.

```
articles <- Corpus(DirSource(path), readerControl = list(language = "en"))
articles <- tm_map(articles, content_transformer(stripWhitespace)) # remove
extra space
articles <- tm_map(articles, content_transformer(tolower)) # lower case
equivTerms <- stemFromFileLoad("equivalent.txt")
reducedDictionary <- as.vector(intersect(unique(equivTerms[[1]]),
dictionary))
articles <- tm_map(articles, content_transformer(function(x)
stemFromFile(doc=x, equivTerms=equivTerms)))
phraseTokenizer <- function(x) RWeka::NGramTokenizer(x, Weka_control(min = 1,
max = 6))
dtm <- DocumentTermMatrix(articles, control = list(
  tokenize = phraseTokenizer,
  dictionary = reducedDictionary))
dtmMatrix <- as.matrix(dtm)
v <- sort(colSums(dtmMatrix),decreasing = TRUE)
d <- data.frame(word = names(v), freq = v) # term frequency list
wordcloud(d$word,d$freq) # generate the word cloud
```

First, the corpus of documents is read from the file system. Then, extra spaces are removed and all contents are converted to lower case, for allowing a direct comparison. Next, stemming occurs for finding all relevant keywords (the right column from Table 1) and transforming them to the corresponding reduced terms (the left column from Table 1), in replacement of their equivalents. The document term matrix is built upon the analysis of the documents and retaining the ones existing in the reduced dictionary. Furthermore, it should be noted that a tokenizer is used for finding the relevant terms considering each of them may be constituted from one to six words (e.g., "adaptive network-based fuzzy inference system"). Finally, besides the document term matrix, to be used as an input for the LDA, the other two direct outputs are the term frequency list and the word cloud.

The execution of the LDA model is very simple using the "topicmodels" package, by simply invoking the LDA function with both the document term matrix (dtm) and the number of topics to be modeled in the two parameters. Then the relation to each term is collected, as well as the most relevant topic in which each article is fitted in the first place (code below).

```
lda <- LDA(dtm, 13)
terms <- terms(lda, length(reducedDictionary))
topics <- topics(lda,1)
```

The resulting approach is relatively straightforward, allowing it to be applied to other numerous contexts involving text analysis (e.g., analysis of online news).

3. Results and analysis

3.1. Articles

In this section, the articles selected are summarized in three categories: publication names (Table 2), publication types (Table 3), and the publishers (Table 4). For both Tables 2 and 4 the publication names/publishers that contribute with just one article are not presented, for page space optimization purposes only. Notably, Elsevier's Expert Systems with Applications journal contributed with 33 articles, helping to consolidate Elsevier's dominant position, with 62 articles. In fact, from the publications contributing with more than one article (Table 2), only the International Journal of Neural Systems is not published by Elsevier. It should also be noted that from Table 2, most of the journals are strongly technology related, with the exception of the International Journal of Forecasting and the European Journal of Operational Research, that are more management related, although also both encourage contributions benefiting from technology approaches. Finally, Table 2 includes only journals, emphasizing the result from Table 3, in which is displayed that 92 articles from the one hundred are published in journals. Hence, GS relevance order appears to favor journals.

Table 2 – Publication names for the articles

Publication name	Number of articles
Expert Systems with Applications	33
Applied Soft Computing	6
Decision Support Systems	4
European Journal of Operational Research	3
Knowledge-Based Systems	2
Procedia Computer Science	2
International Journal of Neural Systems	2
International Journal of Forecasting	2
Information Sciences	2

Table 3 – Publication types for the articles

Publication type	Number of articles
Journal	92
Conference	8

Table 4 – Publisher names for the articles

Publisher name	Number of articles
Elsevier	62
IEEE	7
Springer	5
Wiley	3
AIRCC	2
World Scientific	2

3.2. Text mining

Following the experiments on the articles' contents with TM, the frequency of the relevant terms defined in the dictionary is show on Tables 5 and 6. On the left, the results for credit risk problems are presented. It is possible to observe that credit scoring accounts for more than half of the credit risk problems being addressed by DM methods for the selected set of one hundred highly relevant articles. Next appears bankruptcy, a subject that has been largely debated due to the impact of the financial crisis. Credit cards still receive a lot of attention, even though it is a subject that has also been widely studied prior to the crisis (Huang et al., 2007). Regulatory projects including DM get almost ten percent of the attention, being one of the subjects highly boosted by the crisis (Gerding, 2009). Finally, fraud is the least studied from the subjects in analysis, with

3.5%. Nevertheless, it can be argued that the improvement in credit scoring evaluation tends to reduce fraud (Siddiqi, 2012).

Table 5 - Frequency of terms for credit risk

Term	#	%
scoring	2,082	57.0
bankruptcy	710	19.4
credit card	394	10.8
regulatory	339	9.3
fraud	129	3.5
Total	**3,654**	**100.0**

Table 6 - Frequency of terms for DM

Term	#	%
neural network	1,834	18.7
support vector machine	1,397	14.2
decision tree	1,182	12.1
ensemble	920	9.4
logistic regression	675	6.9
hybrid	616	6.3
clustering	604	6.2
sampling	522	5.3
genetic algorithm	385	3.9
feature selection	371	3.8
bagging	302	3.1
discriminant analysis	252	2.6
multiple criteria	206	2.1
rough set	130	1.3
bayesian	108	1.1
anfis	80	0.8
decision support system	76	0.8
data quality	72	0.7
case-based reasoning	55	0.6
particle swarm	17	0.2
Total	**9,804**	**100.0**

On Table 6, results are shown for the frequency of DM terms. Neural networks are advanced machine learning techniques that try to mimic human brain using artificial neurons for apprehending non-linear relations between input variables (Moro et al., 2014). This technique has been largely studied in the literature (Khashman, 2010), with good modeling results, thus is a naturally good candidate for enhancing solutions for credit risk problems, appearing at the top of our list, with almost nineteen percent of the total occurrences. Support vector machines is the second most mentioned DM method. Such modeling technique emerged in the nineties (Tian et al., 2012) becoming one of the most complex and successful among those in the machine learning domain. The traditional decision trees stand in the test of time, in the third place, considering the recent 2010-2014 timeframe. In fourth comes ensemble modeling, in which a few different techniques are combined for obtaining a better result than any of the isolated methods. Several other modeling techniques are included in the list. Also methods for selecting the appropriate records (e.g., sampling and feature selection) have been studied for improving credit risk assessment. Surprisingly, data quality, a key issue for large DM projects, particularly in the regulatory domain (Moges et al., 2013), is still weakly associated with credit risk. This is an interesting gap for researchers to fill in the near future. Figure 2 complements Tables 5 and 6 by visually helping to understand the differences between terms in the dictionary.

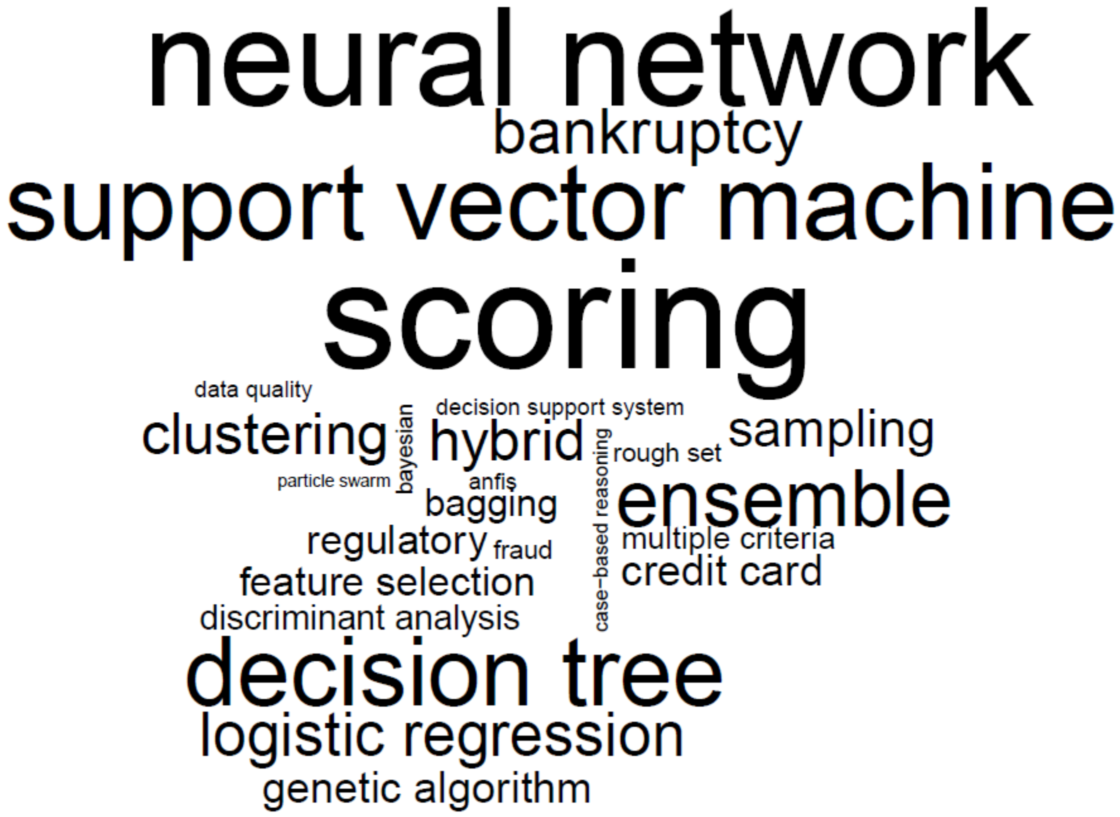

Figure 2 – Word cloud

3.3. Topics of articles

As stated in section 2.3, the result of LDA is a set of topics grouping articles logically according to the frequency of each term in the dictionary. Table 7 summarizes the findings in thirteen topics, showing the number of articles as well as the most relevant credit risk domain and DM method in each topic. Also for exemplification purposes, one article was selected from each topic.

The first impression that arises by looking at Table 7 is that scoring is receiving the most of the attention, with more than half of the articles (53 out of 100) and six topics. Nevertheless, according to the β values, for three of the topics encompassing 21 articles with β greater than four, it is a weak relation, even though scoring is the credit risk problem more closely related with those topics. Such finding reveals that those three topics are more closely related with DM techniques exploration than with benefiting credit risk scoring. In fact, both the examples selected for two of those topics confirm this hypothesis: Khashman's (2010) work focused on different neural models and learning schemes, while Zhou et al. (2010) used the nearest subspace method for improving classification, which is a technique based on training samples' selection. The two papers are highly technology related. The exception is the example chosen for Mandala et al. (2012) which tries to improve credit risk scoring in a local context through decision trees. Still analyzing scoring topics, it is also interesting to note that the topic that includes most of the articles (nineteen) is related to logistic regression, which is one of the most basic techniques. Perhaps this is another indicator that there is still room for research in advanced DM techniques that can translate a direct improvement in credit scoring.

Table 7 – Topics of articles

#	**Credit risk**		**DM method**		**Example**
	Term	**β**	**Term**	**β**	
19	scoring	0.23	logistic regression	2.25	(Yap et al., 2011)
12	scoring	4.64	neural network	0.23	(Khashman, 2010)
7	scoring	5.44	decision tree	0.08	(Mandala et al., 2012)
6	scoring	3.75	hybrid	0.59	(Tsai and Chen, 2010)
6	scoring	2.77	feature selection	0.59	(Marinaki et al., 2010)
3	scoring	4.55	sampling	0.42	(Zhou et al., 2010)
53	articles				
10	bankruptcy	5.81	ensemble	0.18	(Verikas et al., 2010)
5	bankruptcy	0.28	clustering	2.19	(De Andrés et al., 2011)
4	bankruptcy	4.88	genetic algorithm	0.30	(Oreski and Oreski, 2014)
19	articles				
9	fraud	1.94	clustering	1.00	(Wu et al., 2014)
8	fraud	5.79	support vector machine	0.03	(Hens and Tiwari, 2012)
17	articles				
6	credit card	0.26	neural network	3.37	(Chen and Huang, 2011)
5	regulatory	0.26	data quality	2.22	(Moges et al., 2013)

Bankruptcy is the second most mentioned credit risk problem, including nineteen articles and three topics. Again in an almost repetition of what was observed for scoring, one may check that the two topics mostly related with the advanced DM methods of ensemble modeling (combination of a few techniques for improving the isolated techniques' results) and genetic algorithms are barely related to bankruptcy, with β values of 5.81 and 4.88, respectively. Such finding reinforces the suspicion that DM research is still missing to explicitly focus on the benefits for bankruptcy, as it happened for scoring. On the other hand, bankruptcy is strongly associated with clustering in the topic with five articles, which provides a mean to group enterprises in terms of bankruptcy risk.

Fraud also gets some attention, with seventeen articles, being more related to clustering for the same reasons as bankruptcy, than with support vector machine, which is very weakly associated with fraud. Such result contrasts with Table 5 and may be justified by a higher concentration of credit fraud issues in a smaller number of articles. The remaining two topics show two median relations: between credit cards and neural networks; and between regulatory issues and data quality. The former topic shows through its example (Chen and Huang, 2011) a more matured trend of research where advanced neural networks are applied toward a solution for a real credit risk problem. As stated previously, such trend has been widely studied prior to the crisis (Huang et al., 2007). The latter topic emphasizes a real problem that emerged particularly after the crisis, given the governmental pressures to audit financial assets and liabilities: financial institutions are posed with a huge problem of reorganizing its management information systems for improving data quality to respond to an increasingly large number of regulatory reports. In fact, the demand for highly detailed reports such as IRB has emphasized a growing pressure over financial institutions and justifies the significant relation found between data quality and regulatory issues (β values of 2.22 and 0.26, respectively).

4. Conclusions

Credit risk poses several interesting problems for which solutions can benefit directly from data mining (DM) approaches. Some of the most widely studied problems include credit scoring, bankruptcy, credit fraud, credit cards and regulatory issues. The 2008 global financial crisis proved that previous solutions were not adequate to predict credit risk on a global scale, although specialized DM approaches to problems such as credit cards provided an already effective method. Particularly, bankruptcy and regulatory issues have received a significant attention in the 2010-2014 post-crisis period analyzed.

This paper presents an automated literature analysis approach to credit risk problems being addressed by DM methods. The automation included the usage of text mining for analyzing contents and the latent Dirichlet allocation (LDA) for organizing the articles collected in topics. One hundred of relevant articles including both “credit risk” and “data mining” were selected for analysis, according to Google Scholar relevance criterion.

Credit scoring is by far the most mentioned credit risk problem, followed by bankruptcy and fraud, while the most cited DM techniques include neural networks and support vector machines, which are two advanced methods, showing that these can be directly applied to credit risk problems. Ensembles that try to bring the best of a few known techniques by combining their results are also largely mentioned in credit risk problems.

By analyzing the topics built on the LDA algorithm, one of the major conclusions is that research on the most advanced and recent DM methods and techniques such as support vector machines and ensembles is more focused on a fine tuning of those techniques than in assessing real benefits for credit risk. More work should be done to take advantage of those techniques toward real-world credit risk applications, making it an interesting research gap to fill. Another finding is that regulatory issues are demanding research in data quality. Such trend is directly related with a huge increase in the post-crisis period of highly detailed regulatory reports that sustain more frequent auditing processes to the financial institutions.

The full approach undertaken can potentially be applied to any kind of literature analysis. In fact, it can also be used for analyzing other collections of texts such as comments within a website. Furthermore, the approach is both flexible and extensible: a full English words analysis may be used instead of a specific dictionary, and other clustering or topic analysis may also be applied.

References

D. M. Blei, A. Y. Ng, and M. I. Jordan, (2003), Latent dirichlet allocation, The Journal of Machine Learning Research, 3, 993–1022.

J. C. Campbell, A. Hindle, and E. Stroulia, (2014) 'Latent Dirichlet Allocation: Extracting Topics from Software Engineering Data', in Bird, Menzies and Zimmermann (eds), The Art and Science of Analyzing Software Data, 1st Edition, Morgan Kaufmann.

S. C. Chen, and M. Y. Huang, (2011) 'Constructing credit auditing and control & management model with data mining technique', *Expert Systems with Applications*, 38(5), 5359–5365.

S. Claessens, M. M. A. Kose, M. A. Kose, M. L. Laeven, and F. Valencia, (2014), 'Financial Crises: Causes, Consequences, and Policy Responses' *International Monetary Fund*.

D. Crystal, (2012), English as a global language, Cambridge University Press.

Y. Tian, Y. Shi, and X. Liu, (2012), 'Recent advances on support vector machines research', *Technological and Economic Development of Economy*, 18(1), 5–33.

J. De Andrés, P. Lorca, F. J. de Cos Juez, and F. Sánchez-Lasheras, (2011), 'Bankruptcy forecasting: A hybrid approach using Fuzzy c-means clustering and Multivariate Adaptive Regression Splines (MARS)', *Expert Systems with Applications*, 38(3), 1866–1875.

J. C. De Winter, A. A. Zadpoor, A. A., and D. Dodou, (2014), 'The expansion of Google Scholar versus Web of Science: a longitudinal study', *Scientometrics*, 98(2), 1547–1565.

D. Delen, and M. D. Crossland, (2008), ‘Seeding the survey and analysis of research literature with text mining’, *Expert Systems with Applications*, 34(3), 1707–1720.

W. Fan, L. Wallace, S. Rich, and Z. Zhang, (2006), ‘Tapping the power of text mining’, *Communications of the ACM*, 49(9), 76–82.

G. Galati, and R. Moessner, (2013), ‘Macroprudential policy–a literature review’, *Journal of Economic Surveys*, 27(5), 846–878.

E. F. Gerding, (2009), ‘Code, Crash, and Open Source: The Outsourcing of Financial Regulation to Risk Models and the Global Financial Crisis’, *Washington Law Review*, 84, 127–198.

M. D. Guerrero-Baena, J. A. Gómez-Limón, and J. V. Fruet Cardozo, (2014), ‘Are Multi-criteria Decision Making Techniques Useful for Solving Corporate Finance Problems? A Bibliometric Analysis’, *Revista de Metodos Cuantitativos para la Economia y la Empresa*, 17, 60–79.

C. M. Hall, (2006), ‘The impact of tourism knowledge: Google scholar, citations and the opening up of academic space’, *E-Review of Tourism Research*, 4(5), 119–136.

A. W. Harzing, (2013), ‘A preliminary test of Google Scholar as a source for citation data: a longitudinal study of Nobel prize winners’, *Scientometrics*, 94(3), 1057–1075.

B. Hatim, and I. Mason, (2014), Discourse and the Translator, Routledge.

A. B. Hens, and M. K. Tiwari, (2012), ‘Computational time reduction for credit scoring: An integrated approach based on support vector machine and stratified sampling method’, *Expert Systems with Applications*, 39(8), 6774–6781.

C. L. Huang, M. C. Chen, and C. J. Wang, (2007), ‘Credit scoring with a data mining approach based on support vector machines’, *Expert Systems with Applications*, 33(4), 847–856.

A. Khashman, (2010), ‘Neural networks for credit risk evaluation: Investigation of different neural models and learning schemes’, *Expert Systems with Applications*, 37(9), 6233–6239.

J. R. Macey, (2012), ‘Regulator Effect in Financial Regulation’, *The Cornell Law Review*, 98, 591–636.

I. G. N. N. Mandala, C. B. Nawangpalupi, and F. R. Praktikto, (2012), ‘Assessing Credit Risk: An Application of Data Mining in a Rural Bank’, *Procedia Economics and Finance*, 4, 406–412.

M. Marinaki, Y. Marinakis, and C. Zopounidis, (2010), ‘Honey bees mating optimization algorithm for financial classification problems’, *Applied Soft Computing*, 10(3), 806–812.

A. I. Marques, V. García, and J. S. Sanchez, (2013), ‘A literature review on the application of evolutionary computing to credit scoring’, *Journal of the Operational Research Society*, 64(9), 1384–1399.

H. T. Moges, K. Dejaeger, W. Lemahieu, and B. Baesens, (2013), ‘A multidimensional analysis of data quality for credit risk management: New insights and challenges’, *Information & Management*, 50(1), 43–58.

S. Moro, P. Cortez, and P. Rita, (2014), ‘A data-driven approach to predict the success of bank telemarketing’, *Decision Support Systems*, 62, 22–31.

S. Moro, P. Cortez, and P. Rita, (2015), ‘Business intelligence in banking: A literature analysis from 2002 to 2013 using text mining and latent Dirichlet allocation’, *Expert Systems with Applications*, 42(3), 1314–1324.

R. Nijskens, and W. Wagner, (2011), ‘Credit risk transfer activities and systemic risk: How banks became less risky individually but posed greater risks to the financial system at the same time’, *Journal of Banking & Finance*, 35(6), 1391–1398.

S. Oreski, and G. Oreski, (2014), ‘Genetic algorithm-based heuristic for feature selection in credit risk assessment’, *Expert Systems with Applications*, 41(4), 2052–2064.

N. Siddiqi, (2012), Credit risk scorecards: developing and implementing intelligent credit scoring (Vol. 3), John Wiley & Sons.

D. S. Soper, and O. Turel, (2012), ‘An n-gram analysis of communications 2000–2010’, *Communications of the ACM*, 55, 81–87.

B. Tabuenca, M. Kalz, S. Ternier, and M. Specht, (2014), ‘Mobile authoring of open educational resources for authentic learning scenarios’, *Universal Access in the Information Society*, pp. 1-15.

C. F. Tsai, and M. L. Chen, (2010), ‘Credit rating by hybrid machine learning techniques’, *Applied Soft Computing*, 10(2), 374–380.

E. Tobback, D. Martens, T. Van Gestel, and B. Baesens, (2014), ‘Forecasting Loss Given Default models: impact of account characteristics and the macroeconomic state’, *Journal of the Operational Research Society*, 65(3), 376–392.

E. Turban, R. Sharda, D. Delen, (2011), Decision Support and Business Intelligence Systems, 9th edition, Pearson.A. Verikas, Z. Kalsyte, M. Bacauskiene, and A. Gelzinis, (2010), ‘Hybrid and ensemble-based soft computing techniques in bankruptcy prediction: a survey’, *Soft Computing*, 14(9), 995–1010.

D. D. Wu, S. H. Chen, and D. L. Olson, (2014), ‘Business intelligence in risk management: Some recent progresses’, *Information Sciences*, 256, 1–7.

B. W. Yap, S. H. Ong, and N. H. M. Husain, (2011), ‘Using data mining to improve assessment of credit worthiness via credit scoring models’, *Expert Systems with Applications*, 38(10), 13274–13283.

X. Zhou, W. Jiang, and Y. Shi, (2010), ‘Credit risk evaluation by using nearest subspace method’, *Procedia Computer Science*, 1(1), 2449–2455.

Indexing

Credit risk, data mining, machine learning, text mining, literature analysis.