International Journal of Business Performance Management

ISSN online: 1741-5039 - ISSN print: 1368-4892
https://www.inderscience.com/ijbpm

An application to rate banks using a new variant of agglomerative clustering algorithm

Hari Hara Krishna Kumar Viswanathan, Punniyamoorthy Murugesan, Vijaya Prabhagar Murugesan, Lavanya Vilvanathan

DOI: 10.1504/IJBPM.2024.10057347

Article History:

Received:	31 May 2021
Last revised:	22 December 2022
Accepted:	26 April 2023
Published online:	03 January 2025

An application to rate banks using a new variant of agglomerative clustering algorithm

Hari Hara Krishna Kumar Viswanathan and Punniyamoorthy Murugesan*

Department of Management Studies,
National Institute of Technology Tiruchirappalli,
Tamil Nadu, 620015, India
Email: harihara.kkv@gmail.com
Email: punniya@nitt.edu
*Corresponding author

Vijaya Prabhagar Murugesan

Indian Institute of Management Jammu,
Canal Road Nawabad, Cantonment,
Jammu and Kashmir, 180016, India
Email: vijaimano92@gmail.com

Lavanya Vilvanathan

Department of Management Studies,
National Institute of Technology Tiruchirappalli,
Tamil Nadu, 620015, India
Email: lavanya@nitt.edu

Abstract: The study aims to contribute to the field of credit ratings, by presenting models grounded on new variants of neighbourhood linkage method (NLM), an agglomerative hierarchical clustering technique. These models have been applied so as to analyse and predict the long-term bank credit ratings provided by an international rating agency. For this cause, the long-term ratings provided by an Indian arm of international rating agency have been considered. The dataset consists of 35 banks operating in India; this consists of 21 rated banks and 14 unrated banks. In order to validate the optimal number of cluster formations, the study uses a novel performance measure called 'modified structure strength'. Ultimately, based on the best performing NLM variant's cluster formations (of rated banks), unrated banks' potential ratings have been predicted. This model is agnostic to country or region and can be employed to forecast credit ratings of any bank across geography.

Keywords: neighbourhood linkage method; NLM; modified structure strength; credit ratings; rating banks.

Reference to this paper should be made as follows: Viswanathan, H.H.K.K., Murugesan, P., Murugesan, V.P. and Vilvanathan, L. (2025) 'An application to rate banks using a new variant of agglomerative clustering algorithm', *Int. J. Business Performance Management*, Vol. 26, No. 1, pp.1–27.

Biographical notes: Hari Hara Krishna Kumar Viswanathan is a research scholar at NIT Trichy, India. He holds an MBA in Finance from NIT Trichy. His research interests include machine learning applications in credit ratings, risk management and stock markets.

Punniyamoorthy Murugesan is a Professor (HAG) at Department of Management Studies, NIT Trichy, India. He has authored over 100 papers/articles that have been published/presented at national and international journals and conferences. His research interests include machine learning, deep learning, applied statistics and data analytics, operations management, project management and costing, logistics and supply chain management, and operations research and decision sciences.

Vijaya Prabhagar Murugesan is an Assistant Professor at Indian Institute of Management – Jammu, India. His academic experience has been in the area of IT systems and analytics. He has authored research papers that have been published at international journals. He holds a PhD from NIT Trichy, India.

Lavanya Vilvanathan is an Assistant Professor at Department of Management Studies, NIT Trichy, India. She has over 15 years of academic experience and has an extensive background as an academician. Her academic interests include organisational structure and design, human resource management and training and development. She holds a PhD from NIT Trichy, India.

1 Introduction

Rating agencies like the Standard and Poor's, Moody's Investors Service and Fitch Investors Service assign ratings to a large spectrum of financial instruments issued by corporate entities including banks and other financial institutions; these credit ratings reflect the rated entity's solvency and financial strength. As the global market complexity has increased over time, both regulators and investors have started to rely heavily on the rating agencies' credit ratings. Further, the number of rating agencies across the emerging and developed markets has been increasing sharply. When investors try to invest in the share of a bank or say a long-term bond, they are in need of an accurate and unbiased rating of the banking entity. Generally, a rating agency assigns its ratings based on the solicitation from the bank; however, there is also a chance of unsolicited ratings; in this scenario, the rating agencies are neither solicited nor compensated by the rated banks. Frequently, entities that have been rated on an unsolicited basis complain about this exercise as there is a general feeling of unfair treatment from the rating agencies (Behr and Güttler, 2008). Potential conflicts of interests could be encountered as rating agencies serve both investors and issuers (in our case, the bank). Investors are the chief consumers of credit rating reports; however, the fees are paid by the issuers. The fees can be considered the major income source for the rating agencies (Baker and Mansi, 2002). In general, credit rating philosophies can be classified into two – through the cycle (TTC) versus point in time (PIT) ratings. The big three credit rating agencies (CRAs) predominantly adopt the former across markets. Rating agencies tend to focus more on the long-term horizon covering more than one business cycle; in other words, they generally assign forward-looking ratings in nature. They do not necessarily reflect the present condition or the immediate future. On the other hand, the PIT approach focuses

on the obligor's present condition, and the most likely future condition over a specified horizon say one year (Rösch, 2005).

A large non-financial corporation filing for bankruptcy is typically less likely to have an effect on the economy. In contrast, the failure of one bank could trigger systemic crises and have a significant negative influence on the economy as a whole. The lack of bank-specific information makes bank failures, often known as bank runs, transmissible, and according to Park (1991), bank runs and the ensuing liquidity issue could lead to the demise of both insolvent and solvent banks. Since banks are so important to the economy, regulators and other interested parties, including depositors, governments, and investors, keep a close eye on their credit ratings. These ratings, however, were given by rating organisations under the TTC method. Ultimately, these ratings are based on the expected long-term credit quality and non-reactive to transitory fluctuations (Cantor, 2001). By designating new material as temporary in nature, rating agencies, who support the TTC approach, may escape the charge of having a delayed reaction to new information. Additionally, investors believe that rating agencies take too long to improve or decrease their ratings (Baker and Mansi, 2002). Gogas et al. (2014) claim that the bank ratings given by rating firms are not routinely updated. Investors should be aware that TTC ratings may perform poorly in anticipating short-term defaults, according to Löffler (2004). The PIT, or current-condition method, is the alternate strategy typically used in quantitative default prediction models as opposed to the TTC design of the rating agencies. In this methodology, the probability of default is estimated over a fixed horizon, which is typically one year, based on current information (Löffler, 2004).

1.1 Research problem

Although bank ratings largely reflect their financial strength, it is likely that there may be some bias present. In response, the importance of soft computing-based rating procedures is increasing (Viswanathan et al., 2020). Rating agencies often base their bank ratings on a number of quantitative and qualitative characteristics; these parameters can be viewed as model input for the clustering approach. Agglomerative hierarchical clustering (AHC) is a well-known and proven method for unsupervised learning. The AHC method has a lot to offer clustering banks and is important to classification research.

From a regulator or an investors' perspective, the application of the novel neighbourhood linkage method (NLM), an AHC technique, in the area of bank ratings might aid them in rightly assessing the creditworthiness of the banks; segregation of banks into clusters can greatly aid in the understanding the risk faced by them. The algorithm can also be used by unrated banks to determine their possible credit ratings. The research objectives (ROs) are formulated by taking into account all of the aforementioned factors. The main goal of the study is to investigate the feasibility of developing an NLM-based machine learning model to analyse bank ratings given by an Indian branch of an international CRA; based on the model's performance, an attempt may then be made to assign credit ratings for unrated banks. Since the NLM model is geographically and politically agnostic, it is anticipated that the method might be expanded to rate banks globally.

In particular, the RO seeks to:

RO 1 Identify the key quantitative factors utilised by CRAs to rate the banks in India.

RO 2 Develop new variants of the NLM algorithm and construct models in order to group the banks into clusters based on the aforementioned quantitative factors.

RO 3 Apply new performance measures to find out the optimal number of clusters for the given dataset.

RO 4 Explain the differences between cluster assignment and the original credit ratings of the banks (assigned by the CRA).

RO 5 Assign (potential) ratings to unrated banks (that were not assigned long-term credit ratings by the CRA).

For the credit rating problem, a wide variety of research studies involving soft computing techniques have been conducted (Huang et al., 2004; Chen and Shih, 2006; Lee, 2007); this study has introduced the application of NLM technique to the credit rating problem (Murugesan and Punniyamoorthy, 2020). In this study, an attempt has been made to improve our recently published NLM technique, by developing two new variants, namely NLM1 and NLM2.

Based on the original NLM technique, two new variants, namely NLM1 and NLM2, have been developed. The original method (NLM) works based on the concept of choosing the maximum variant of the coordinates involved in the cell update of the pairwise distance matrix. Subsequently, the average distance is calculated between the coordinates of the data points and the coordinates' maximum dimension values; this distance has been computed to update the pairwise distance matrix's cell value. In the second method (NLM1), the average of both maximum and the minimum dimension values of the coordinates involved in the cell update of the pairwise distance matrix is calculated. With the help of the average variant of the coordinates, the average distance is calculated between the coordinates of the data points and the average dimension values of the coordinate to update the cell value of the pairwise distance matrix. The third method (NLM2) is based on the idea of choosing the minimum variant of the coordinates and computing the distance with the coordinates of the data points. The pairwise distance matrix is updated with the average distance from the calculated distance values. Following this step, in order to determine the best variant of our algorithm, all the three variants (NLM, NLM1, and NLM2) have been compared against each other through the usage of a new performance measure called $MS(c)$.

The paper proceeds as follows. Section 2 is literature survey that describes cluster techniques and the various aspects of the NLM. Section 3 elucidates the modifications done to the original NLM. Section 4 describes the research data; in particular, the details on the ratios and ratings, data collection and data pre-processing. Subsequently, Section 5 elaborates on the experimental setup. Section 6 deals with experimental results and discussions. The future works and implications of the study are given in Section 7. Section 8 will provide the conclusion of the paper.

2 Literature survey

This section aims to offer an overview of cluster techniques and the generalised workings of the AHC algorithm; further, salient features of NLM are also discussed.

2.1 Brief overview of cluster techniques

Clustering techniques can be considered as the method of grouping similar objects into a cluster. The core aim of clustering methods is to enhance the similarity of the members in a cluster and make the data within each cluster have the greatest similarity, but the highest dissimilarity among diverse clusters (Michalski et al., 1983). Cluster techniques can be largely classified into three: hierarchical clustering, optimisation clustering, and model-based clustering. The most prevalent technique in the literature is the hierarchical clustering. A hierarchical clustering method essentially groups data points into a hierarchy of clusters; hierarchical clustering techniques can be divided into two – namely agglomerative and divisive. In the agglomerative method, clusters are formed in a bottom-up (integrating) manner; the divisive method adopts a top-down (breaking-down) fashion (Han et al., 2012; Yoshino et al., 2015). Within the hierarchical clustering, AHC technique initiates from the division of the dataset into singleton nodes. It combines step-by-step the existing pair of mutually closest nodes into a new node until one final node is left, which encompasses the full dataset. In AHC, under the linkage technique, there are single, complete and average methods. One of the key advantages of using AHC is that it needs no prior information on the '*c*' values. This feature offers a hierarchy relationship between the clusters and makes it quite easy to execute and gives the best results (Moore, 2001; Müllner, 2015; Yoshino et al., 2015). Some of the popular AHC techniques in the research world are single, complete, linkage, wards and weighted methods.

2.2 Neighbourhood linkage method

In the hierarchical clustering methods such as single, complete and average, the linkage distance is taken as a base for clustering. This study employs the neighbourhood clustering method based on the idea of the neighbourhood in order to improve the performance of the clustering. In the prior research, the neighbourhood's performance was compared against other conventional agglomerative clustering methods. A variety of label and non-label datasets have been utilised as a part of the experiments. The study has clearly demonstrated the outperformance of neighbourhood-based clustering method in all the datasets.

3 Proposed changes to NLM

In this section, the key features and changes made to $S(c)$, and the NLM algorithm are discussed; further, the algorithm's workings are elucidated.

3.1 Structure strength (S(c)) and modified structured strength (MS(c))

When it comes to hierarchical clustering, one can see that there are numerous performance measures to identify the number of clusters (c); in non-hierarchical clustering, there is a concept called the structure strength $S(c)$ which is used to determine the 'c' value. Li and Mukaidono (1995) introduced a performance measure incorporating both accuracy and effectiveness below equation

$$S(c) = (1-\alpha)\log\left(\frac{L(1)}{L(c)}\right) + \alpha\log\left(\frac{N}{c}\right) \tag{1}$$

$$L(c) = \sum_{k=1}^{N}\sum_{i=1}^{c} d_{ik}^2$$

α 0.5 (fixed)

$L(1)$ variance of the entire dataset

$L(c)$ within group sum of squares

N number of observations

c number of clusters.

But there is no trade-off between the accuracy and effectiveness of classification for all dataset. So, the existing formula does not have the trade-off between accuracy and effectiveness of classification. As a result, we are not able to determine the number of clusters by using the existing structure strength $S(c)$ expression. Since this expression was designed for non-hierarchical clustering and did not find a place for the similarity index (SI). So, the expression is modified such that the SI should find a place in equation (1). The calculation of the weightage ($\beta = 1 - \alpha$) in the existing formula does not have the trade-off between accuracy and effectiveness. So, the place of weightage value 'β' is interchanged to calculate the modified structure strength $S(c)$. Also, one can ensure that there is trade-off exist between accuracy and effectiveness. This will help to determine the number of clusters (c) for non-class label datasets (Murugesan and Punniyamoorthy, 2020).

The expression (2) for the new performance measure, $MS(c)$ is given below:

$$MS(c) = (\beta)\log\left(\frac{L(1)}{L(c)}\right) + \alpha\log\left(\frac{N}{c}\right) \tag{2}$$

The equation (1) holds good only when a beta value is less than 0.5.

α = function of β

$$\beta = \frac{SI - SI_{\min}}{SI_{\max} - SI_{\min}} \text{ where } \beta = 1-\alpha$$

where

β weightage calculated through the index (α)

SI similarity index

$SI_{\min}$ lowest value of the SI

$SI_{\max}$ highest value of the SI

c number of clusters.

Table 1 Pseudo-code for NLM

Input:

X – Data matrix = $\{x_i\}$ *where* $i = 1 \ldots N$; *N – Number of observations*;

Pd – Pair-wise distance matrix of X = $\{d_{jk}\}$ *where* $j, k = 1 \ldots N$; *P-variables*

Start

Initialise $s = 0$;

Input X;

Compute $Pd = [d_{jk}]_{N \times N}$, $\forall j, k = 1 \ldots N$ // Euclidean distance expression is used.

Choose $[\min_j, \min_k] = argmin(Pd)$, *for* $j \neq k$;

// $\min_j$ *and* $\min_k$ are row and column indices of *Pd* matrix.

Set C = $\{\min_j, \min_k\}$;

For $l = 1: N$

Exclude *if* $l \in C$

Switch 'method

Case 'max'

X_{new}= max(*X*(*all the points in set C* ∪ *l*));

// maximum variants of coordinates of given data points in *X*

Case 'avg'

X_{max} = max(*X*(*all the points in set C*∪ *l*));

X_{min} = min(*X*(*all the points in set C*∪ *l*));

Xnew = *avg* (X_{max}, X_{min});

Case 'min'

Xnew= min(*X*(*all the points in set C* ∪ *l*));

// minimum variants of coordinates of given data points in *X*

End Switch

m = |*all the points in set C*∪ *l*|; // | . | - Number of points in the set

$Nd_i = dist(Xnew, X(all\ the\ points\ in\ set\ C \cup l))$; $i = 1..m$

$d_{new,l} = average(Nd)$;

End

Count $s = s + 1$;

Reduce Pd matrix to size of $((N - s) \times (N - s))$;

Update the Pd matrix cells with the $d_{new,l}$;

Find $[\min_j, \min_k] = \min(Pd)$ *where* $j \neq k$;

Set C with new $\{\min_j, \min_k\}$;

Repeat until Pd become 2×2 *matrix size.*

End

The modification of the $S(c)$ expression deals with the justification of α value in the expression $S(c)$. In the computation of the original $S(c)$, α is always fixed at 0.5; here, alpha being constant, equal weightage is given to the classification's accuracy and effectiveness. Owing to this, there is no trade-off between the two; this inhibits the computation of an optimal number of clusters (c). In order to overcome this issue, using

our $MS(c)$ expression, weightage for the effectiveness of classification has been increased so as to get the proper trade-off between accuracy and effectiveness. This has been achieved by satisfying the below conditions:

- α values should be more than 0.5.
- α value should more than β value.
- Sum of α and β values are always equal to 1.

Newly devised $S(c)$ formula to find the α value, is based on the function of SI; this expression aids to identify a better method based on the SI. Hence, this change to the original $S(c)$ expression would enhance the α value to enable a proper trade-off between effectiveness and accuracy of classification – this aids in determining the optimal 'c' using $MS(c)$. Out of all the three variants, NLM2 provides the best $MS(c)$ and thereby the best results for our research problem. Further, as an extension to the study, unrated banks have been assigned likely rating values based on the new clusters formed using NLM2. It is opined that the technique is not restricted to Indian rating scenario, but this could very well be extended to ratings in other countries.

The pseudo-code of NLMs (NLM, NLM1, and NLM2) was presented in Table 1.

3.2 Neighbourhood linkage method 1 (NLM1)

In NLM1, the newly developed variant, the coordinates of the data points involved in the pairwise distance matrix's cell update are taken. The average of each variant's largest and smallest value would be computed from the coordinates, and it becomes a new coordinate. Subsequently, keeping the new coordinate as a reference, the computation of the distance between the coordinates of the data points involved in the cell update is performed. The average is taken to update the cell value of the pairwise distance matrix from the distance values. This process has been carried out for cell updation.

3.3 Neighbourhood linkage method 2 (NLM2)

In NLM2, the second newly developed variant, the coordinates of the data points involved in the pairwise distance matrix's cell update are taken. The smallest value of each variant is computed from the coordinates, and it becomes the minimum variant of the coordinate. Subsequently, the minimum coordinate variant is kept as a reference to calculate the distance between the coordinates of the data points involved in the cell update. The average is taken to update the cell value of the pairwise distance matrix from the distance values. This process has been carried out for cell update.

4 Research data

One of the study's core objectives is to develop a model that aids in explaining the credit ratings by reference to key ratios of the banks. This section details the key elements in the research design, including 12 key ratios, data collection, and data pre-processing.

Table 2 Ratings and the corresponding key ratios of the rated 21 SCBs – FY 2018–2019

Name of the bank	*Rating*	*Ratio 1*	*Ratio 2*	*Ratio 3*	*Ratio 4*	*Ratio 5*	*Ratio 6*	*Ratio 7*	*Ratio 8*	*Ratio 9*	*Ratio 10*	*Ratio 11*	*Ratio 12*
Rated Bank 1	AA-	0.007	0.016	0.176	0.052	0.799	0.097	0.125	-0.035	0.495	0.050	0.049	0.022
Rated Bank 2	AA+	0.014	0.017	0.162	0.057	0.738	0.084	0.110	-0.011	0.314	0.053	0.028	0.014
Rated Bank 3	AAA	0.049	0.042	0.053	0.021	0.770	0.113	0.158	0.006	0.444	0.047	0.130	0.059
Rated Bank 4	AAA	0.042	0.049	0.096	0.033	0.787	0.104	0.134	0.001	0.402	0.044	0.035	0.033
Rated Bank 5	AA+	0.031	0.040	0.158	0.056	0.770	0.110	0.142	-0.008	0.434	0.048	0.027	0.024
Rated Bank 6	A+	0.007	0.011	0.164	0.055	0.815	0.099	0.119	-0.030	0.497	0.047	0.063	0.013
Rated Bank 7	AAA	0.039	0.046	0.088	0.054	0.681	0.083	0.119	0.001	0.309	0.051	0.022	0.016
Rated Bank 8	A+	0.013	0.023	0.193	0.077	0.766	0.075	0.096	-0.017	0.462	0.052	0.048	0.017
Rated Bank 9	AAA	0.074	0.070	0.014	0.004	0.714	0.149	0.171	0.019	0.424	0.049	0.118	0.464
Rated Bank 10	AAA	0.053	0.050	0.074	0.021	0.706	0.136	0.169	0.004	0.446	0.044	0.062	0.080
Rated Bank 11	A+	0.013	0.017	0.275	0.101	0.829	0.089	0.116	-0.047	0.425	0.059	0.078	0.025
Rated Bank 12	AAA	0.016	0.018	0.071	0.038	0.657	0.110	0.132	0.001	0.355	0.047	0.015	0.029
Rated Bank 13	A+	0.012	0.017	0.220	0.108	0.714	0.078	0.102	-0.014	0.383	0.054	0.045	0.011
Rated Bank 14	AAA	0.022	0.017	0.019	0.007	0.719	0.167	0.179	0.020	0.525	0.052	0.117	0.253
Rated Bank 15	AA+	0.041	0.051	0.155	0.066	0.745	0.062	0.097	-0.013	0.435	0.048	0.048	0.015
Rated Bank 16	AA	0.006	0.008	0.118	0.072	0.595	0.078	0.109	-0.005	0.268	0.062	0.101	0.007
Rated Bank 17	AAA	0.201	0.224	0.075	0.030	0.787	0.098	0.127	0.000	0.457	0.051	0.069	0.034
Rated Bank 18	AA	0.018	0.020	0.114	0.062	0.664	0.093	0.142	-0.009	0.368	0.053	0.030	0.014
Rated Bank 19	A+	0.009	0.015	0.250	0.097	0.749	0.086	0.107	-0.018	0.445	0.049	0.015	0.018
Rated Bank 20	AA+	0.027	0.032	0.150	0.069	0.662	0.080	0.118	-0.006	0.361	0.052	0.016	0.009
Rated Bank 21	A+	0.006	0.010	0.165	0.087	0.729	0.101	0.130	-0.016	0.514	0.048	0.018	0.020

Table 3 Key ratios of the unrated 14 SCBs – FY 2018–2019

Bank label	*Ratio 1*	*Ratio 2*	*Ratio 3*	*Ratio 4*	*Ratio 5*	*Ratio 6*	*Ratio 7*	*Ratio 8*	*Ratio 9*	*Ratio 10*	*Ratio 11*	*Ratio 12*
Unrated Bank 1	0.011	0.014	0.154	0.057	0.833	0.104	0.123	-0.031	0.316	0.052	0.017	0.024
Unrated Bank 2	0.001	0.001	0.049	0.023	0.780	0.160	0.167	-0.011	0.278	0.060	0.018	0.059
Unrated Bank 3	0.001	0.001	0.075	0.024	0.847	0.106	0.138	0.001	0.320	0.058	0.005	0.050
Unrated Bank 4	0.005	0.005	0.044	0.030	0.585	0.112	0.132	0.006	0.281	0.056	0.062	0.036
Unrated Bank 5	0.005	0.004	0.014	0.007	0.653	0.121	0.135	0.013	0.250	0.054	0.151	0.203
Unrated Bank 6	0.004	0.003	0.020	0.006	0.721	0.279	0.292	0.043	0.408	0.049	0.078	0.491
Unrated Bank 7	0.003	0.003	0.030	0.018	0.630	0.132	0.156	0.016	0.250	0.055	0.013	0.082
Unrated Bank 8	0.010	0.010	0.029	0.015	0.672	0.134	0.141	0.009	0.322	0.051	0.063	0.082
Unrated Bank 9	0.008	0.005	0.024	0.013	0.559	0.153	0.155	-0.012	0.129	0.103	0.058	0.164
Unrated Bank 10	0.006	0.007	0.092	0.049	0.643	0.091	0.125	0.005	0.507	0.047	0.027	0.749
Unrated Bank 11	0.004	0.005	0.088	0.050	0.569	0.143	0.160	0.003	0.300	0.056	0.096	0.027
Unrated Bank 12	0.006	0.006	0.049	0.035	0.425	0.100	0.126	0.003	0.242	0.057	0.007	0.025
Unrated Bank 13	0.002	0.003	0.043	0.024	0.732	0.155	0.162	0.006	0.246	0.057	0.041	0.059
Unrated Bank 14	0.022	0.017	0.032	0.019	0.431	0.084	0.165	0.005	0.331	0.078	0.106	0.060

Table 4 Standardised data – ratings and the corresponding – key ratios of the rated 21 SCBs – FY 2018–2019

Name of the bank	*Rating*	*Ratio 1*	*Ratio 2*	*Ratio 3*	*Ratio 4*	*Ratio 5*	*Ratio 6*	*Ratio 7*	*Ratio 8*	*Ratio 9*	*Ratio 10*	*Ratio 11*	*Ratio 12*
Rated Bank 1	AA–	0.033	0.070	0.620	0.464	0.885	0.177	0.223	0.134	0.924	0.227	0.305	0.019
Rated Bank 2	AA+	0.069	0.071	0.569	0.512	0.743	0.122	0.153	0.402	0.467	0.271	0.155	0.010
Rated Bank 3	AAA	0.244	0.184	0.149	0.160	0.817	0.250	0.378	0.595	0.795	0.189	0.859	0.070
Rated Bank 4	AAA	0.208	0.214	0.316	0.282	0.857	0.210	0.265	0.531	0.689	0.144	0.209	0.034
Rated Bank 5	AA+	0.151	0.174	0.555	0.501	0.816	0.239	0.301	0.430	0.769	0.197	0.151	0.023
Rated Bank 6	A+	0.034	0.044	0.576	0.492	0.924	0.188	0.193	0.187	0.928	0.185	0.400	0.007
Rated Bank 7	AAA	0.190	0.201	0.286	0.478	0.607	0.117	0.195	0.531	0.453	0.233	0.116	0.011
Rated Bank 8	A+	0.063	0.099	0.687	0.704	0.808	0.080	0.088	0.334	0.841	0.255	0.295	0.013
Rated Bank 9	AAA	0.366	0.311	0.000	0.000	0.684	0.416	0.437	0.737	0.745	0.208	0.779	0.616
Rated Bank 10	AAA	0.261	0.219	0.231	0.160	0.666	0.357	0.427	0.568	0.800	0.146	0.391	0.098
Rated Bank 11	A+	0.063	0.074	1.000	0.933	0.956	0.144	0.180	0.000	0.748	0.361	0.502	0.024
Rated Bank 12	AAA	0.079	0.079	0.220	0.322	0.550	0.236	0.256	0.538	0.570	0.184	0.071	0.029
Rated Bank 13	A+	0.057	0.072	0.789	1.000	0.685	0.095	0.116	0.373	0.641	0.284	0.272	0.005
Rated Bank 14	AAA	0.107	0.073	0.021	0.030	0.697	0.496	0.474	0.748	1.000	0.257	0.772	0.331
Rated Bank 15	AA+	0.204	0.227	0.542	0.592	0.758	0.022	0.094	0.384	0.773	0.193	0.294	0.010
Rated Bank 16	AA	0.028	0.030	0.401	0.655	0.402	0.094	0.149	0.471	0.351	0.400	0.659	0.000
Rated Bank 17	AAA	1.000	1.000	0.236	0.251	0.858	0.183	0.233	0.526	0.829	0.238	0.439	0.035
Rated Bank 18	AA	0.090	0.085	0.383	0.554	0.567	0.162	0.303	0.427	0.602	0.267	0.174	0.008
Rated Bank 19	A+	0.042	0.064	0.905	0.895	0.768	0.132	0.139	0.318	0.798	0.205	0.066	0.014
Rated Bank 20	AA+	0.131	0.138	0.522	0.620	0.563	0.104	0.189	0.458	0.586	0.255	0.072	0.003
Rated Bank 21	A+	0.027	0.042	0.579	0.795	0.721	0.199	0.246	0.345	0.973	0.197	0.086	0.017

4.1 Selection of key ratios

Ratings of the 21 banks were gathered from the Indian subsidiary of the global rating agency. In this study, 12 financial and non-financial ratios were selected as model input; these 12 ratios were selected based on the research study conducted by Viswanathan et al. (2020). In the aforementioned study, the top management of Indian arm of the International CRA was asked to identify the important quantitative metrics that are considered significant in rating a bank entity. The rating agency retorted by offering a list of 12 metrics/ratios that are utilised in their internal rating process. These metrics are as follows: market share in advances (ratio 1), market share in deposits (ratio 2), gross non-performing assets ratio (ratio 3), net non-performing assets ratio (ratio 4), provisioning coverage ratio (ratio 5), cost of borrowing (ratio 6), the ratio of deposits in current and saving accounts to total deposits (ratio 7), common equity tier 1 capital ratio (ratio 8), total capital adequacy ratio (ratio 9), net worth by NPA ratio (ratio 10), return on assets (ratio 11) and share of fee-based income as a proportion of total income (ratio 12) (Viswanathan et al., 2020).

4.2 Data collection

In the study, data (for the 12 variables) has been collected predominantly from public sources such as bank's official websites and the central bank's website. The dataset consists of 35 banks as on FY 2018-19. Out of these 35 banks, 21 banks have been assigned credit ratings (long term – Basel III compliant) by the rating agency, and 15 banks are unrated (long-term ratings). The collated data for the 21 rated and 14 unrated Indian SCBs are provided in Table 2 and Table 3, respectively. The names of the unrated banks are provided in the appendix section.

4.3 Preparation of data

The min-max feature scaling technique, also known as standardisation or normalisation, is performed on the entire dataset (all independent variables) of 35 SCBs, one can observe that the data has been transformed into values ranging between 0 and 1. This transformation is mainly performed so as to alleviate the effect size. This type of scaling technique completes a linear transformation on the dataset; it is normally used so as to improve prediction accuracy as it avoids the chance of assigning higher weightage to larger value inputs and lower weightage to smaller values. The normalised values for the rated banks are provided in Table 4.

5 Experimental setup

In this section, the three NLMs' practical applicability is tested experimentally by using the data collected for the 35 scheduled commercial banks (SCBs) in India; as noted earlier, this consists of 21 rated banks and 14 unrated banks. The former is referred to as bank dataset for the sake of simplicity. As part of the experiments, the full dataset of 35 banks consisting of key ratios and corresponding ratings, despite having class labels, is considered as non-class label dataset. For conducting all the experiments in the study, the

MATLAB (2017) software system was used. In the study, there are two stages of our experiments.

In the first stage, three NLMs are applied to the 21 rated banks and clusters of banks are formed based on a performance measure called *MS*(*c*). After identifying the best performing method that exhibits the right '*c*' value with the highest *S*(*c*), the rationale behind the bank's assignment to a cluster is explained, and comparison against the original credit rating is also performed. When it comes to non-hierarchical clustering, one can see performance measures such as sum of squares error (SSE), Dunn Index, Davies-Bouldin Index and Silhouette Index are used to identify the '*c*' value; in hierarchical clustering, we have introduced a concept called the *MS*(*c*) to determine the '*c*' value. The expression for *MS*(*c*) has been provided in expression (1) of Subsection 1.2.

In the second stage, the already formed clusters (based on the best NLM) are used, and (potential) ratings are assigned to 14 unrated banks by looking at the similarity/closeness of the unrated bank against a cluster.

6 Experimental results

Using *MS*(*c*) as a performance measure; the three Neighbourhood methods have been applied to the non-class label dataset of 21 banks to identify clusters.

6.1 Importance of α value in MS(c)

The alpha (α) value is the weightage given to the effectiveness of classification in the *MS*(*c*). The equations for the *MS*(*c*) have been given in expression (1) of Subsection 1.2. For illustrative purposes, a graph of Effectiveness, accuracy and *S*(*c*) has been plotted using the same (bank) dataset for NLM2 Variant (Figure 1). This has been plotted by keeping α value as a constant at 0.5.

Figure 1 Number of clusters (c) versus *S*(*c*) – bank dataset (see online version for colours)

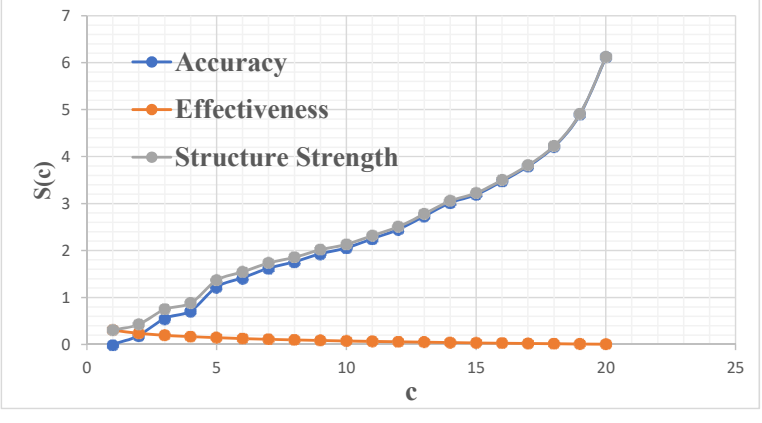
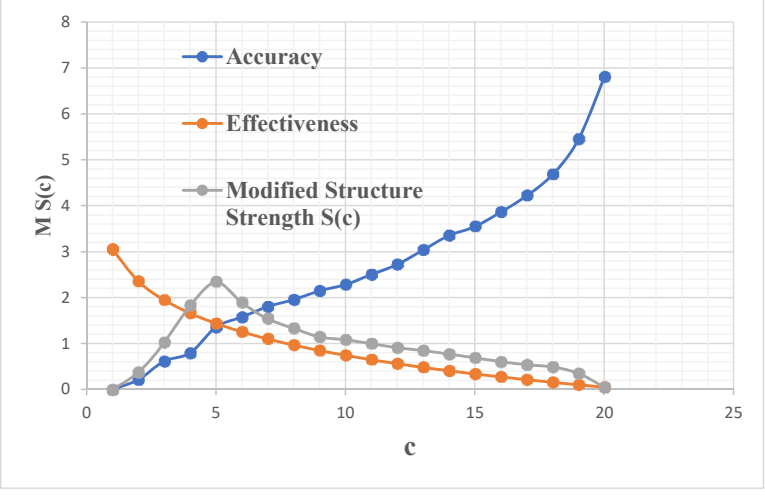


Table 5 Calculation of effectiveness, accuracy and $S(c)$ for each cluster using bank dataset

c	*Effectiveness*	*Accuracy*	*S(c)*
20	*0.004879*	*6.109989*	*6.114868*
19	*0.010008*	*4.894421*	*4.904429*
18	*0.015415*	*4.208873*	*4.224288*
17	*0.021131*	*3.79401*	*3.815141*
16	*0.027193*	*3.474147*	*3.50134*
15	*0.033647*	*3.190718*	*3.224365*
14	0.040547	3.014618	3.055165
13	0.047957	2.731258	2.779215
12	0.055962	2.448822	2.504783
11	0.064663	2.251373	2.316035
10	0.074194	2.055043	2.129236
9	0.08473	1.932841	2.017571
8	0.096508	1.759623	1.856131
7	0.109861	1.622799	1.73266
6	0.125276	1.418464	1.54374
5	0.143508	1.222349	1.365857
4	0.165823	0.715489	0.881311
3	0.194591	0.555638	0.750229
2	0.235138	0.193809	0.428947
1	0.304452	0	0.304452

Figure 2 Number of clusters (c) versus $MS(c)$ – bank dataset (see online version for colours)



From Figure 1 and Table 5, it can be inferred that the 'c' increases with both $S(c)$ and accuracy. Therefore, by keeping α value as 0.5, expression (1) of Subsection 1.2, there is no trade-off between the accuracy and effectiveness of classification. As a result, the 'c'

cannot be determined by fixing α value constant at 0.5. For all the three variants of NLM, a similar plot could be seen where there is no trade-off between the accuracy and effectiveness of classification. For this purpose, a new way to determine 'α' value was introduced based on the SI and is given in expression (2) of Subsection 1.2. Accordingly, 'α' value is calculated for each SI, and it gives high weightage to the effectiveness of classification, and low weightage to the accuracy of classification in the *MS*(*c*). Hence, it aids in getting a proper trade-off to determine the '*c*' value. The formula for the *MS*(*c*) has been given in expression (2) of Subsection 1.2. By using the same (bank) dataset, the graph of a number of clusters (c) versus *MS*(*c*) has been plotted and shown in Figure 2 and Table 6. This graph has been plotted using the bank dataset for NLM2 Variant. From the figure, one can observe the trade-off that exists between accuracy and effectiveness; this trade-off aids in determining the '*c*' value for bank dataset. Ultimately, this exhibits the importance of α value in the *MS*(*c*).

Table 6 Calculation of effectiveness, accuracy and *MS*(*c*) for each cluster using bank dataset

c	*Effectiveness*	*Accuracy*	*MS(c)*
20	0.048790	6.788876	0.04879
19	0.100083	5.438246	0.349825
18	0.154151	4.676526	0.490742
17	0.211309	4.215566	0.536679
16	0.271934	3.860163	0.603556
15	0.336472	3.545242	0.690791
14	0.405465	3.349576	0.769285
13	0.479573	3.034731	0.846237
12	0.559616	2.720913	0.909146
11	0.646627	2.501525	0.996484
10	0.741937	2.283381	1.08142
9	0.847298	2.147601	1.142373
8	0.965081	1.955137	1.328669
7	1.098612	1.80311	1.536846
6	1.252763	1.576071	1.886097
5	*1.435085*	*1.358165*	*2.33447*
4	1.658228	0.794987	1.82445
3	1.945912	0.617376	1.022644
2	2.351375	0.215344	0.377575
1	3.044522	0	0

The α value for each NLM is shown in Table 8.

Table 7 α value for each NLM – bank dataset

	NLM	*NLM1*	*NLM2*
α value	0.58	0.55	*0.63*

From Table 7, one can observe that NLM2 is getting the highest α value of 0.63 than all the NLMs. Here, the α and β values represent the weightage value for effectiveness and accuracy of classification in $MS(c)$ expression. In NLM2, getting the highest weightage ($\alpha = 0.63$) for the effectiveness of classification influences the calculation of $MS(c)$ and thereby gets the highest $MS(c)$ value for NLM2. Also, the highest α value helps provide the proper trade-off between the accuracy and effectiveness of the classification to determine the cluster number that exhibits the highest $MS(c)$ value. To further test and reaffirm the performance of all the NLM algorithms, five class label datasets have been taken and used as input in order to determine the best algorithm; the class label datasets used are iris, soybean, seed, shuttle and leaf.

Table 8 Number of clusters (c) predicted by NLMs using $MS(c)$ – Class Label Datasets

Class label datasets	*c*	*NLM*	*NLM1*	*NLM2*
Iris	3	3	3	3
Soya	4	4	*6*	4
Seed	3	3	*5*	3
Shuttle	3	3	3	3
Leaf	4	4	*5*	4

Table 8 shows that the '*c*' value is correctly predicted by NLM and NLM2 methods for all class label datasets; but, NLM1 can predict the right '*c*' value for only two datasets out of five. Therefore, it can be concluded that the NLM and NLM2 algorithms have performed better than the NLM1 method. Further, the $MS(c)$ has been computed for all the three NLM models (for all the five class label datasets) to determine the best method; the results are shown in Table 9.

Table 9 α value and the corresponding $MS(c)$ calculation for NLMs – class label datasets

	α value				
	Iris	Soya	Seed	Shuttle	Leaf
NLM	0.61	0.56	0.68	0.57	0.63
NLM1	0.58	0.53	0.62	0.53	0.59
NLM2	*0.67*	*0.62*	*0.71*	*0.59*	*0.67*
			$MS(c)$		
NLM	0.3456	0.2461	0.1587	0.1422	0.1788
NLM1	0.3319	0.2321	0.1486	0.1418	0.1752
NLM2	*0.3472*	*0.2477*	*0.1646*	*0.1449*	*0.1795*

Note: The highlighted value indicates the best α value and $MS(c)$.

In Table 9, the best α values and the corresponding $MS(c)$ values are highlighted for each class label dataset.

6.2 Calculation of MS(c) for different NLM algorithms

Using $S(c)$ as a performance measure; the three neighbourhood methods have been applied to the non-class label dataset of 21 banks to identify clusters. In Figure 3(a), a

plot of banks vs. SI for NLM has been provided; here, it is noted that with SI has a value of 0.823. Further, one can see from the plot of banks vs. $MS(c)$ for NLM [Figure 4(a)] that five clusters are formed at the $S(c)$ of 0.3396. If the SI value is more than or less than 0.823, the corresponding $S(c)$ would be less. Based on the performance measure $S(c)$, it can be observed that the dataset has been classified/clustered into five groups. In Figures 3(a)–3(c), a dendrogram for the best cut-off has been given. Similarly, NLM1 and NLM2 with SI values are calculated; the values are 0.803 and 1.02, respectively. Further, a plot of banks vs. $MS(c)$ is provided in Figure 3(b) and Figure 3(c). Further, it can be seen from the plot [Figure 4(b) and Figure 4(c)] that five clusters are formed at the $S(c)$ values are 0.3271 and 0.3999, respectively. Based on the performance measure $S(c)$, one can see that the dataset has been classified/clustered into five groups.

Figure 3 Banks versus SI, (a) NLM (b) NLM1 (c) NLM2 (see online version for colours)

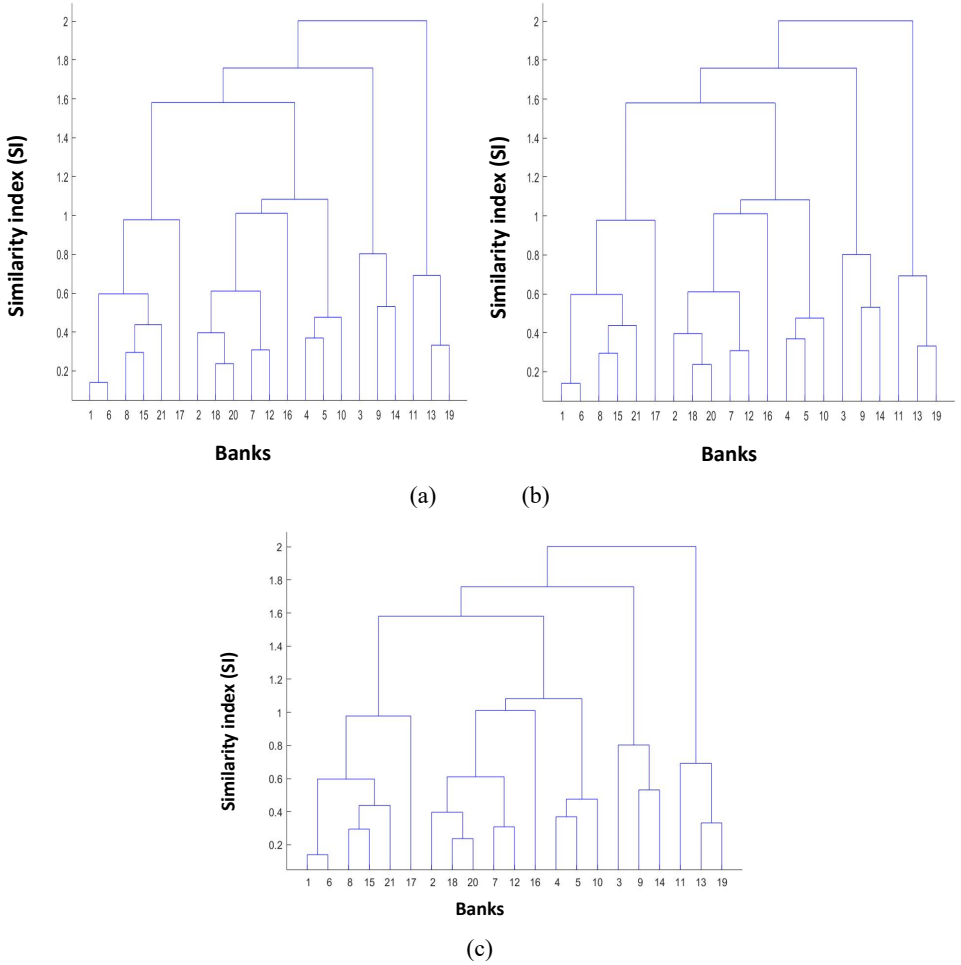
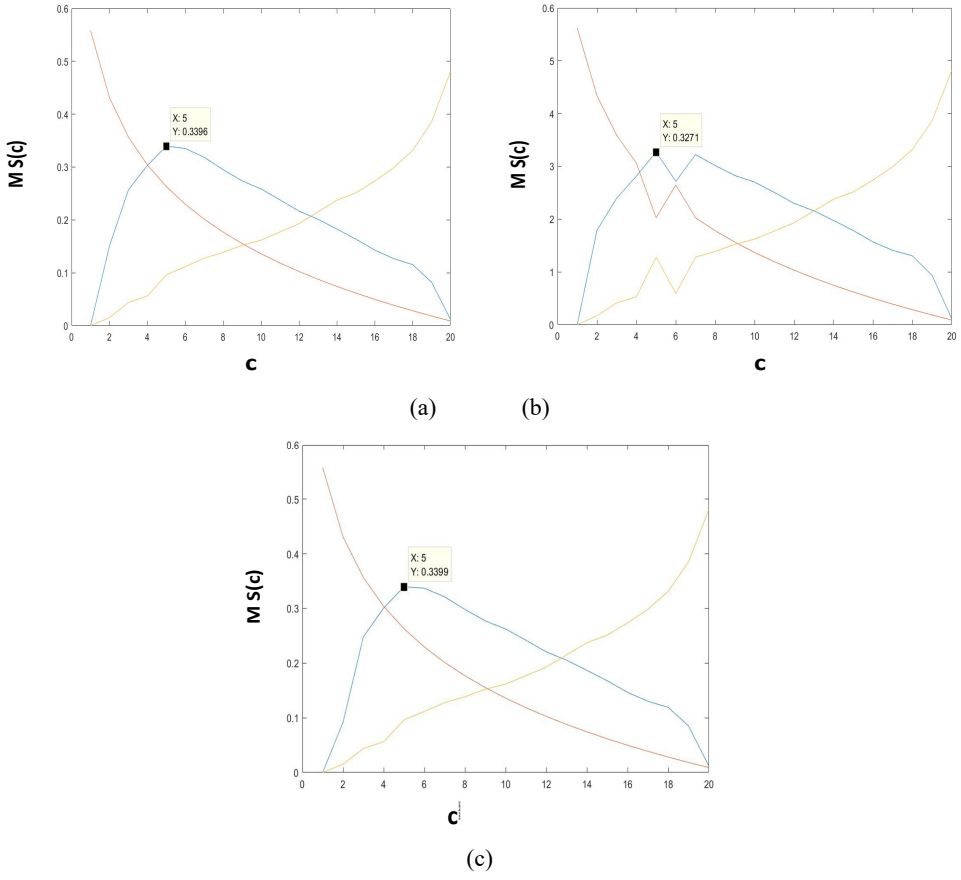


Figure 4 Number of clusters versus *MS*(*c*) – bank dataset, (a) NLM (b) NLM1 (c) NLM2 (see online version for colours)



Based on the *MS*(*c*) results, the '*c*' value for each NLM has been determined; the results are shown in Table 10.

Table 10 *MS*(*c*) values for all NLMs – Bank Dataset

Methods	*MS(c)*	*Best cut-off of similarity index value (SI)*	*Number of clusters (c) predicted*
NLM	0.3396	0.823	5
NLM1	0.3271	0.803	5
NLM2	0.3399	1.02	5

From Table 10, one can identify that each method, groups the bank dataset with five number of clusters. The banks which are there in each cluster are the same for all the three NLM algorithms. But, NLM2 exhibits the best result in terms of *MS*(*c*) value with the highest value of 0.3399 among all the methods. Further, this method grouped the banks with the highest cut-off value or SI of 1.02, as compared to all the other methods. The highest cut-off value is directly influenced by the α value of *MS*(*c*); where α value is the weightage given to compactness of the cluster.

From Table 8–10, it can be clearly seen that NLM2 has the highest α and $MS(c)$ values across all the class label datasets, including the bank datasets. The highest α and $MS(c)$ values reaffirm the superior performance of NLM2 against other NLM algorithms. Also, the dataset has been classified/clustered into five groups. The results are provided below.

Table 11 Clusters

Cluster	*Bank name*	*Original rating*
Cluster #1	Rated Bank 11	A+
Cluster #1	Rated Bank 13	A+
Cluster #1	Rated Bank 19	A+
Cluster #2	Rated Bank 3	AAA
Cluster #2	Rated Bank 9	AAA
Cluster #2	Rated Bank 14	AAA
Cluster #2	Rated Bank 17	AAA
Cluster #3	Rated Bank 4	AAA
Cluster #3	Rated Bank 10	AAA
Cluster #3	Rated Bank 5	AA+
Cluster #4	Rated Bank 7	AAA
Cluster #4	Rated Bank 12	AAA
Cluster #4	Rated Bank 2	AA+
Cluster #4	Rated Bank 20	AA+
Cluster #4	Rated Bank 18	AA
Cluster #4	Rated Bank 16	AA
Cluster #5	Rated Bank 15	AA+
Cluster #5	Rated Bank 1	AA–
Cluster #5	Rated Bank 6	A+
Cluster #5	Rated Bank 8	A+
Cluster #5	Rated Bank 21	A+

6.3 Rating the unrated banks using the NLM2 method

In the second stage of the experiment, after arriving at the five clusters (given in Table 11), each of these clusters' centroid values is calculated. The centroid values for the five clusters are provided in Table 12.

Each unrated bank is assigned to its closest cluster based on the (least possible) Euclidean distance. This Euclidean distance is calculated by considering the centroid value of the five clusters and the bank's metrics (12 standardised metrics). The formula for Euclidean distance is given below:

$$\begin{aligned} d(p,q) &= \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2} \\ &= \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2} \end{aligned} \tag{2}$$

Note: p and q represent the ratios/coordinates of each unrated bank.

Table 12 Centroid values of the five clusters

Cluster	*MSA*	*MSD*	*GNPA*	*NNPA*	*PCR*	*CET1*	*TCAR*	*RoA*	*CASA*	*COB*	*Fee*	*NW by NNPA*
Cluster 1	0.054	0.070	0.898	0.943	0.803	0.123	0.145	0.230	0.729	0.283	0.280	0.015
Cluster 2	0.429	0.392	0.102	0.110	0.764	0.336	0.380	0.651	0.842	0.223	0.712	0.263
Cluster 3	0.207	0.203	0.367	0.314	0.780	0.269	0.331	0.510	0.753	0.162	0.250	0.052
Cluster 4	0.098	0.101	0.397	0.524	0.572	0.139	0.207	0.471	0.505	0.268	0.208	0.010
Cluster 5	0.072	0.096	0.601	0.609	0.819	0.133	0.169	0.277	0.888	0.211	0.276	0.013

Table 13 Unrated banks and their corresponding Euclidean distances to the centroid of the 5 clusters

Name of the unrated bank	*Distance to cluster 5*	*Distance to cluster 4*	*Distance to cluster 3*	*Distance to cluster 2*	*Distance to cluster 1*
Unrated Bank 1	*0.518*	0.538	0.628	1.227	0.682
Unrated Bank 2	0.961	0.690	*0.653*	1.044	1.248
Unrated Bank 3	0.837	0.636	*0.581*	1.071	1.146
Unrated Bank 4	0.985	*0.535*	0.700	0.933	1.255
Unrated Bank 5	1.378	1.100	1.092	*0.874*	1.645
Unrated Bank 6	1.778	1.616	1.390	*1.223*	2.045
Unrated Bank 7	1.144	*0.692*	0.771	1.077	1.421
Unrated Bank 8	0.986	0.644	*0.590*	0.751	1.328
Unrated Bank 9	1.563	*1.168*	1.336	1.418	1.723
Unrated Bank 10	1.145	*1.113*	1.064	1.199	1.372
Unrated Bank 11	0.949	*0.610*	0.758	0.939	1.137
Unrated Bank 12	1.234	*0.747*	1.023	1.355	1.427
Unrated Bank 13	1.021	*0.650*	0.659	0.954	1.295
Unrated Bank 14	1.338	*0.992*	1.111	1.063	1.576

Table 14 Unrated banks grouping

Clusters	*Rated and unrated banks*	*Rating*	
Cluster 5	Rated Bank 1	AA–	AA
	Rated Bank 6	A+	A
	Rated Bank 8	A+	A
	Rated Bank 15	AA+	AA
	Rated Bank 20	A+	A
	Unrated Bank 1		
Cluster 4	Rated Bank 2	AA+	AA
	Rated Bank 18	AA	AA
	Rated Bank 20	AA+	AA
	Rated Bank 7	AAA	AAA
	Rated Bank 12	AAA	AAA
	Rated Bank 16	AA	AA
	Unrated Bank 11		
	Unrated Bank 12		
	Unrated Bank 13		
	Unrated Bank 14		
	Unrated Bank 9		
	Unrated Bank 7		
	Unrated Bank 4		

Table 14 Unrated banks grouping (continued)

Clusters	*Rated and unrated banks*	*Rating*	
Cluster 3	Rated Bank 4	AAA	AAA
	Rated Bank 5	AA+	AA
	Rated Bank 10	AAA	AAA
	Unrated Bank 2		
	Unrated Bank 3		
	Unrated Bank 8		
	Unrated Bank 10		
Cluster 2	Rated Bank 3	AAA	AAA
	Rated Bank 9	AAA	AAA
	Rated Bank 14	AAA	AAA
	Rated Bank 17	AAA	AAA
	Unrated Bank 5		
	Unrated Bank 6		
Cluster 1	Rated Bank 11	A+	A
	Rated Bank 13	A+	A
	Rated Bank 19	A+	A

The unrated banks and their corresponding Euclidean distances to the centroid of the five clusters are given in Table 13. In this table, the lowest values are italicised to highlight their proximity to a cluster.

In the second stage, the already formed clusters (based on the NLM2) are used, and ratings are assigned to 14 unrated banks by looking at the unrated bank's similarity/closeness against a cluster; the details are given in Table 14.

6.4 Discussion

In this study, one of the core objectives is to analyse the rationale behind banks' rating; in other words, analysing the ratings, which is essentially a classification problem. The dataset encompasses 2018–2019 Financial Year data of about 21 rated banks and 14 unrated banks; clustering results of the neighbourhood (NLM2) model have been presented in Table 14. The table mentioned above contains the details on the clusters of rated banks formed using the neighbourhood model based on the $MS(c)$ formula; also, these tables have the details on unrated banks and their respective proximity (based on Euclidean distance) to clusters of rated banks. The following subsections aim to describe each cluster of rated banks' formation and the unrated banks mapped under it (based on Euclidean distance).

Cluster #1

It is to be noted that the cluster consists of three rated banks; it can be observed that the credit ratings of these banks correspond to A+. The cluster consists of weak banks (in the Indian context), but these banks have high payment capacity from a rating standpoint.

The applied algorithm has clearly carved out these banks into a separate cluster. Details on cluster #1 are depicted in Table 15.

Table 15 Cluster #1 – rated banks with no unrated banks close to it

Cluster #1	*Name of the bank*	*Original rating*	*Distance to centroid of cluster #1*
1	Rated Bank 11	A+	0.3800
2	Rated Bank 13	A+	0.2430
3	Rated Bank 19	A+	0.2611

Cluster #2

The cluster consists of banks with credit ratings equal to AAA; the cluster contains very strong banks (from an Indian context) having the highest credit quality. The neighbourhood algorithm has clearly segregated these banks into a unique cluster. Further, as per the results from the second stage of the experiments, two banks – Bank 5 and Bank 6 have close proximity to this cluster #2. This implies that both of these unrated banks have strong potential to be AAA-rated. Information on cluster #2 is provided in Table 16.

Table 16 Cluster #2 – rated banks and closely related unrated banks

Cluster#2	*Name of the bank*	*Original rating*	*Distance to centroid of cluster #2*
1	Rated Bank 3	AAA	0.3977
2	Rated Bank 9	AAA	0.4414
3	Rated Bank 14	AAA	0.54882
4	Rated Bank 17	AAA	0.9643
		Unrated banks	
1	Unrated Bank 5	Unrated	0.8740
2	Unrated Bank 6	Unrated	1.2230

Cluster #3

The cluster consists of three rated banks, where the first two banks have a rating of AAA, and the third bank has a rating of AA+. Regarding the presence of two AAA rated banks in this cluster, the phenomenon could be attributed to the TTC effect in the credit ratings. In other words, due to the unresponsiveness of the rating to transitory fluctuations in the economic cycle, the first two rated banks managed to retain their rating (of AAA). If transitory movements are considered, they might have an AA+ rating (a notch lower than the prime AAA). This cluster mostly contains relatively strong banks (from an Indian rating perspective) having high credit quality. From the results of the second stage of experiments, four unrated banks have close proximity to this cluster #3. This implies that these unrated banks have a strong potential to get a rating of AA+ from the CRA. Details on cluster #3 are depicted in Table 17.

Table 17 Cluster #3 – rated banks and closely related unrated banks

Cluster #3	*Name of the bank*	*Original rating*	*Distance to centroid of cluster #3*
1	Rated Bank 4	AAA	0.1560
2	Rated Bank 10	AAA	0.3213
3	Rated Bank 5	AA+	0.3089
		Unrated banks	
1	Unrated Bank 3	Unrated	0.5809
2	Unrated Bank 8	Unrated	0.5900
3	Unrated Bank 2	Unrated	0.6529
4	Unrated Bank 10	Unrated	1.0637

Cluster #4

This cluster has six rated banks having ratings ranging from AAA to AA. Regarding the inclusion of first two banks (AAA rated) in this cluster, the TTC effect could be considered as a major factor driving this phenomenon. Hence, the two AAA rated banks may not be entitled to AAA rating but rather have a potential AA+ rating if transitory movements are considered.

This cluster consists mostly of good banks (from an Indian rating perspective) having high credit quality. Based on the results of the second stage of experiments, seven banks have close proximity to this cluster #4. This implies that these unrated banks have good potential to get a rating ranging from AA+ to AA from the CRA. Information on cluster #4 is provided in Table 18.

Table 18 Cluster #4 – rated banks and closely related unrated banks

Cluster #4	*Name of the bank*	*Original rating*	*Distance to centroid of cluster #4*
1	Rated Bank 7	AAA	0.2250
2	Rated Bank 12	AAA	0.3456
3	Rated Bank 2	AA+	0.2697
4	Rated Bank 20	AA+	0.2332
5	Rated Bank 18	AA	0.1542
6	Rated Bank 16	AA	0.5531
		Unrated banks	
1	Unrated Bank 4	Unrated	0.5351
2	Unrated Bank 11	Unrated	0.6095
3	Unrated Bank 13	Unrated	0.6498
4	Unrated Bank 7	Unrated	0.6920
5	Unrated Bank 12	Unrated	0.7474
6	Unrated Bank 14	Unrated	0.9922
7	Unrated Bank 9	Unrated	1.1681

Cluster #5

This cluster consists of five rated banks where the ratings of first two banks correspond to AA+ and AA–; the last three banks have a rating of A+. As per the original ratings, this cluster has four relatively weak banks (from and Indian rating standpoint). Regarding the presence of the only AA+ rated bank in this cluster, it can been seen as the influence of TTC effect on the credit ratings. It can be noted that AA+ rated bank shares very similar/comparable values with the cluster #5 centroid. This indicates that the bank's financials are not really reflected on the actual credit rating. If transitory movements are taken into account, the aforementioned bank might have a potential rating of AA or perhaps AA–. The cluster #5 largely contains relatively weak banks (from an Indian rating perspective) having high payment capacity. From the results of the second stage of experiments, one unrated bank, Bank 1, has close proximity to this cluster #5. This implies that this unrated bank has a good chance to get a rating of AA to A+ from the CRA. Details on cluster #5 are depicted in Table 19.

Table 19 Cluster #5 – rated banks and closely related unrated banks

Cluster #5	*Name of the bank*	*Original rating*	*Distance to centroid of cluster #5*
1	Rated Bank 15	AA+	0.2920
2	Rated Bank 1	AA–	0.2362
3	Rated Bank 6	A+	0.2424
4	Rated Bank 8	A+	0.1835
5	Rated Bank 21	A+	0.3283
		Unrated banks	
1	Unrated Bank 1	Unrated	0.5182

Based on the results, it can be concluded that the AHC-NLM2 can exhibit an effective and efficient performance in analysing the rationale behind the original credit ratings assigned by the banks; hence, looking from a PIT approach, the proposed method does a better job in assignment of ratings for banks. Further, the unrated banks' potential ratings have been provided based on the formed clusters using the AHC-NLM2 algorithm.

7 Implications and future scope of the study

Due to the plethora of data and superfast computational capabilities in the current digital age, machine learning algorithms have gained importance in the financial world; there are no finance areas left untouched, including credit ratings. In the credit rating domain, models grounded on statistics are applied widely; logistic regression is a popular technique which is still being used. Our study aims to contribute to the field of application of machine learning in credit ratings, particularly banks' ratings. The NLM models developed for rating the banks are based on 12 key (financial and non-financial) metrics. Even though the study is on the Indian banks, it is opined that the NLM model is agnostic to country or region. The NLM model and the corresponding inputs can be collated for any bank in any country, and the potential ratings of these entities could be predicted. Ultimately, this model and $MS(c)$, the performance measure, could be utilised by multiple parties such as regulatory agencies, investors, and depositors to evaluate the

risk levels of banking entities. Also, this methodology could be used by the unrated banks to find out their potential credit ratings before approaching a rating agency.

One of the research study's limitations is that only quantitative factors were considered inputs to the model; in the future, research studies might be conducted that include subjective parameters like the extent of sovereign support, quality of management, willingness to repay, etc. This algorithm could be used by multiple stakeholders such as regulators, investors, and depositors to assess banks' creditworthiness; further, this method can be applied by the unrated banks themselves to find out their potential credit ratings. In the study, credit ratings of Indian banks were considered; moreover, this algorithm could very well be extended to the banks in other country settings.

8 Conclusions

AHC technique and the usage of the newly devised $MS(c)$ overcome the problems faced by clustering techniques such as inability to separate the clusters and grouping the greater number of clusters. Based on the original NLM, two new variants (NLM1 and NLM2) were developed and models were constructed to analyse the bank ratings; further, a new performance measure called $MS(c)$ was employed to determine the optimal number of groupings in the dataset. The NLM2 model offers better reliability in results when compared against statistical methods. Although NLM1 model did not have the highest $S(c)$ for the bank dataset, it could still perform better than traditional AHC methods like single, complete and weighted. Moreover, the NLM2 has been tested on different class label datasets to ascertain the best method through the $MS(c)$ calculation. In this study, for the unrated banks, Euclidean distance method has been applied to find out the proximity to the newly formed clusters and thereby potential ratings have been provided. CRAs have a tendency to ignore the transient events in the economy. Owing to this, it is less likely for the CRAs to revise the ratings frequently; this produces a need for a rating methodology that considers the current economic state. From the results of the study, it can be concluded that NLM2 meets this need.

References

Baker, H.K. and Mansi, S.A. (2002) 'Assessing credit rating agencies by bond issuers and institutional investors', *Journal of Business Finance and Accounting*, Vol. 29, Nos. 9–10, pp.1367–1398, DOI: 10.1111/1468-5957.00474.

Behr, P. and Güttler, A. (2008) 'The informational content of unsolicited ratings', *Journal of Banking and Finance*, Vol. 32, No. 4, pp.587–599, DOI: 10.1016/j.jbankfin.2007.04.021.

Cantor, R. (2001) 'Moody's investors service response to the consultative paper issued by the Basel Committee on Bank Supervision 'A new capital adequacy framework'', *Journal of Banking and Finance*, Vol. 25, No. 1, pp.171–185, DOI: 10.1016/S0378-4266(00)00121-7.

Chen, W.H. and Shih, J.Y. (2006) 'A study of Taiwan's issuer credit rating systems using support vector machines', *Expert Systems with Applications*, Vol. 30, No. 3, pp.427–435, DOI: 10.1016/j.eswa.2005.10.003.

Gogas, P., Papadimitriou, T. and Agrapetidou, A. (2014) 'Forecasting bank credit ratings', *Journal of Risk Finance*, Vol. 15, No. 2, pp.195–209, DOI: 10.1108/JRF-11-2013-0076.

Han, J., Kamber, M. and Pei, J. (2012) *Data Mining: Concepts and Techniques, Data Mining: Concepts and Techniques*, DOI: 10.1016/C2009-0-61819-5.

Huang, Z. et al. (2004) 'Credit rating analysis with support vector machines and neural networks: a market comparative study', *Decision Support Systems*, Vol. 37, No. 4, pp.543–558, DOI: 10.1016/S0167-9236(03)00086-1.

Lee, Y.C. (2007) 'Application of support vector machines to corporate credit rating prediction', *Expert Systems with Applications*, Vol. 33, No. 1, pp.67–74, DOI: 10.1016/j.eswa.2006.04.018.

Li, R.P. and Mukaidono, M. (1995) 'Maximum-entropy approach to fuzzy clustering', in *IEEE International Conference on Fuzzy Systems*, pp.2227–2232, DOI: 10.1109/fuzzy.1995.409989.

Löffler, G. (2004) 'An anatomy of rating through the cycle', *Journal of Banking and Finance*, Vol. 28, No. 3, pp.695–720, DOI: 10.1016/S0378-4266(03)00041-4.

Michalski, R., Carbonell, J. and Mitchell, T. (1983) *Machine Learning. An Artificial Intelligence Approach*, Vol. 2, DOI: 10.1007/978-3-662-12405-5.

Moore, A.W. (2001) 'K-means and hierarchical clustering', *Statistical Data Mining Tutorials*, pp.1–24 [online] http://www.cs.cmu.edu/afs/cs/user/awm/web/tutorials/kmeans11.pdf (accessed 1 January 2021).

Müllner, D. (2015) 'fastcluster: fast hierarchical, agglomerative clustering routines for R and Python', *Journal of Statistical Software*, DOI: 10.18637/jss.v053.i09.

Murugesan, V.P. and Punniyamoorthy, M. (2020) 'Development of new agglomerative and performance evaluation models for classification', *Neural Computing and Applications*, Vol. 32, No. 7, pp.2589–2600, DOI: 10.1007/s00521-019-04297-4.

Park, S. (1991) 'Bank failure contagion in historical perspective', *Journal of Monetary Economics*, Vol. 28, No. 2, pp.271–286, DOI: 10.1016/0304-3932(91)90054-R.

Rösch, D. (2005) 'An empirical comparison of default risk forecasts from alternative credit rating philosophies', *International Journal of Forecasting*, Vol. 21, No. 1, pp.37–51, DOI: 10.1016/j.ijforecast.2004.04.001.

Viswanathan, H.H.K.K. et al. (2020) 'A modified boosted support vector machine to rate banks', *Benchmarking*, DOI: 10.1108/BIJ-01-2020-0006.

Yoshino, N. et al. (2015) 'SME credit risk analysis using bank lending data: an analysis of Thai SMEs', *SSRN Electronic Journal*, DOI: 10.2139/ssrn.2641712.