Expert Systems With Applications 165 (2021) 113940

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Sequence-based clustering applied to long-term credit risk assessment

Richard Le [a], Hyejin Ku [a,*], Doobae Jun [b]

[a] Department of Mathematics and Statistics, York University, 4700 Keele Street, Toronto, Canada
[b] Department of Mathematics and Research Institute of Natural Science, Gyeongsang National University, 501 Jinju-daero, Jinju-si, Republic of Korea

ARTICLE INFO

Keywords:
Credit rating
Clustering
Sequence matrix
Transition matrix
Default behaviour

ABSTRACT

This paper studies the effectiveness of estimating credit rating transition matrices using sequence-based clustering on historical credit rating sequences. The data set used in this study consisted of monthly credit rating sequences from Korean companies from 1986 to 2018. The credit rating sequences were converted to sequence matrices and was clustered using PCA-guided K-means. Representative transition matrices of the resulting clusters were then generated to be used in the classification process. The proposed clustering model is evaluated under the 3 different long-term classification scenarios; 7 class credit rating prediction, credit rating transition direction (upgrade, stay, or downgrade) prediction, and default behaviour prediction. All three classification scenarios produced promising results suggesting that the representative transition matrix of the K clusters better describes future credit rating behaviour than a single transition matrix.

1. Introduction

This paper investigates the effectiveness of using historical credit rating sequences to characterize companies in clustering and the resulting transition matrices for the purpose of credit risk analysis. By better utilizing historical credit rating sequences we can improve the estimation of transition matrices. By using sequence matrices we group firms with similar transition behaviour together, and firms exhibiting any momentum in their transitions would belong to the same cluster.

Credit ratings and their revisions can lead to a number of major decisions and hence, consequences. It is in one's best interest to invest in better forecasting techniques to mitigate any losses dependent on credit ratings. In this paper we will be adapting the general clustering methodology described in Park, Suresh, and Jeong (2008) and apply a transition matrix estimation method to predict future credit behaviours solely from historical credit ratings. Park et al. (2008) developed a sequence representation scheme based on Markov models, enabling sequences of web usage activities to be clustered using vector based distances. This method is known as sequence-based clustering. As far as we can tell, we are the first to study the application of sequence-based clustering using K-means strictly on historical credit rating sequences. The majority of models observed in literature use a snapshot of a companies financial statement and fewer models use a historical sequence of financial statements (Chen, Ribeiro, Vieira, & Chen, 2013).

Markov chains are commonly used in modelling the behaviour of credit rating transitions over time. Jarrow, Lando, and Turnbull (1997) were one of the first to model the term structure of credit risk spreads using Markov chains in both the discrete and continuous time case. They estimated the transition probability matrix from historical data by first estimating the generator matrix from a 1-year estimate of transition probabilities provided from a credit rating agency. The generator matrix can be estimated either implicitly from bond market prices or from historical bond transition rating changes. Thomas, Allen, and Morkel-Kingsbury (2002) extends the Jarrow–Turnbull model by introducing a hidden Markov model for the term structure of credit risk spreads. Kiefer and Larson (2004) tested the effectiveness of using a time-homogeneous Markov model to describe the credit rating transitions of municipal bonds, commercial papers, and sovereign debt. They have found that the time-homogeneous Markov model can adequately describe credit rating transitions of municipal bonds over a period of 5 years and commercial papers over a period of 6 months. Credit rating transitions for sovereign debt is also adequately described by Markov models but this conclusion may be the result of the low number of data samples. Dharmaraja, Pasricha, and Tardelli (2017) introduces a hybrid Markov model where they incorporate the asset value of the firm in the transition probabilities of credit ratings. Sharma, Jadi, and Ward (2018) investigates the financial performance of insurance companies by using credit rating transition matrices under a Markov model, noting that less risky rating grades result in more rating stability.

Studies have shown the promising results that clustering can produce in the context of credit risk and credit rating predictions. In a study by Guo, Zhu, and Shi (2012), they compared their proposed support vector domain description (SVDD) combined with fuzzy clustering

[*] Corresponding author.
E-mail addresses: ler3@yorku.ca (R. Le), hku@mathstat.yorku.ca (H. Ku), dbjun@gnu.ac.kr (D. Jun).

https://doi.org/10.1016/j.eswa.2020.113940
Received 19 February 2020; Received in revised form 26 August 2020; Accepted 27 August 2020
Available online 2 September 2020

model with other kinds of support vector machine learning techniques in the context of corporate credit rating classification. The performance of each model was evaluated based on the hit-ratio, the ratio of the number of correct classifications and the overall number of classifications. The variables used as the input of the model are bond-rating data sets from the Korean and Chinese markets. The variables range from shareholder's equity to cash flow from operating activities. Chen et al. (2013) use a trajectory clustering procedure consisting of two consecutive self-organizing maps (SOM) processes. Their method allows for the visualizations of the bankruptcy trajectories of companies enabling a unique perspective and insight on bankruptcy influences. Their model clusters financial statements containing 29 financial ratios of companies spanning 3 years. Morales, Rodríguez, and Montero (2015) applied different fuzzy classification methods for the use in rating classifications. They use both credit ratings and financial statement ratios in their model. Irmatova (2016) introduces a relative attribute rating model (RELARM) based on relative PCA attributes and K-means clustering. Using 9 financial and economic parameters, their model assigns ratings based off the ranked projections of the cluster centres onto a rating vector. In the case of long-term credit rating prediction, the true rating that a firm receives in the future will not be known until that future date arrives. During this period, new credit rating information may become available. Kuncheva and Sánchez (2008) terms this type of problem as delayed labelling and investigates the effectiveness of online nearest neighbour classifiers for treating delayed labelling problems. Plasse and Adams (2016) developed an online linear discriminant analysis algorithm which was applied to a real world consumer credit data set where delayed label information was introduced synthetically. Montiel, Bifet, and Abdessalem (2017) proposed two over-indebtedness risk prediction frameworks, one of which treats over-indebtedness as a streaming learning problem. Although not done in this study, we may be able to extend our model to consider delayed labelling by treating credit rating sequences as a streaming data problem.

The remainder of this paper is organized into 5 sections. In Section 2 we discuss the theory behind our proposed model. In Section 3 we provide an overview of the three different classification scenarios that our model will undertake. In Section 4 we introduce the data to be used, and describe the specific methods of the experiments. In Section 5 we present the results and discussions of our proposed model. Finally we conclude the study in Section 6.

2. Sequence-based clustering

In this section, we describe the methods for sequence-based clustering. We begin by defining the sequence matrices and their properties. The Sequence matrices will be the main objects that are being clustered. We then describe credit rating transition matrices as these will be used in the classification algorithm. Finally, we go over the K-means clustering method used in this study.

2.1. Sequence matrices for credit ratings

We introduce sequence matrices in order to measure the distance between the historical patterns of credit ratings of firms. Consider an n-state time-homogeneous Markov chain where each state represents a particular credit rating. In order to begin clustering these objects, we utilize the representation of sequence vectors and sequence matrices introduced in Park et al. (2008).

Definition 2.1.1. Let $m \in \mathbb{N}$, $X_1^m, X_2^m, \dots, X_T^m$ be a sequence of random variables, and $\mathcal{S}$ be the state space. Then the sequence vector of length $T \in \mathbb{N}$ is the vector $\mathbf{x}_m(T) = (X_1^m, X_2^m, \dots, X_T^m)$ for some firm m with states $X_i^m \in \mathcal{S}$.

Definition 2.1.2. Let N_{ij} be the number of transitions from state i to j for some firm m with the sequence vector $\mathbf{x}_m(T)$. The corresponding sequence matrix $\mathbf{S}_m$ is then an $n \times n$ matrix whose entries are denoted by

$$S_m(i,j) = \begin{cases} \frac{N_{ij}}{\sum_j N_{ij}} & \text{if } N_{ij} > 0, \\ 0 & \text{if } N_{ij} = 0, \end{cases} \tag{1}$$

and so, the entries represent the relative frequency of transitions from state i to j.

Therefore, given a sequence vector $\mathbf{x}_m(T)$ we can generate the corresponding sequence matrix $\mathbf{S}_m$. This sequence matrix describes the frequency of transitions of the given sequence vector.

In the context of credit ratings, there tends to be few credit rating transitions over the period of T leading to sparse credit rating sequences. An extreme example of this observation would be one where a firm takes only one rating for the entire period of T. Suppose $X_t = 2$ for $t \le T$, then the resulting sequence matrix would contain a single entry at $S(2,2) = 1$ and $S(i,j) = 0$ everywhere else.

Now we present some general properties of sequence matrices when using the Euclidean distance measure for the comparison of different sequence matrices. The Euclidean distance measure is used to measure the distance between the historical patterns of credit ratings for the firms. We consider a time-homogeneous Markov chain X with state space $\mathcal{S} = \{1, 2, \dots, n\}$. For a sequence matrix, the sum of the entries in a nonzero row is 1, i.e., $S_m(i,j) = 0$ for all $1 \le j \le n$ or,

$$\sum_{j=1}^{n} S_m(i,j) = 1.$$

Lemma 2.1.1. *Consider a vector $(a_1, a_2, \dots, a_n)$ that satisfies $\sum_{i=1}^n a_i = 1$.*

(a) The minimum value of $\sum_{i=1}^n a_i^2$ is $\frac{1}{n}$ and it is achieved at $(\frac{1}{n}, \frac{1}{n}, \dots, \frac{1}{n})$.

(b) If $0 \le a_i \le 1$ for all i, then the maximum value of $\sum_{i=1}^n a_i^2$ is 1, that is,

$$\frac{1}{n} \le \sum_{i=1}^{n} a_i^2 \le 1.$$

Proof. By the Cauchy–Schwarz inequality,

$$\left(\sum_{i=1}^{n} a_i^2\right)\left(\sum_{i=1}^{n} 1^2\right) \ge \left(\sum_{i=1}^{n} a_i\right)^2$$

Since $\sum_{i=1}^n a_i = 1$,

$$\sum_{i=1}^{n} a_i^2 \ge \frac{1}{n}$$

where the equality holds when $a_i = \frac{1}{n}$, $i = 1, \dots, n$. For the maximum value, we consider

$$\sum_{i=1}^{n} a_i^2 = \left(\sum_{i=1}^{n} a_i\right)^2 - \sum_{i \ne j}^{n} a_i a_j.$$

Since $0 \le a_i, a_j \le 1$,

$$\sum_{i=1}^{n} a_i^2 \le \left(\sum_{i=1}^{n} a_i\right)^2 = 1. \quad \square$$

Next we consider two sequence matrices $\mathbf{S}_{m1}$ and $\mathbf{S}_{m2}$ that represent Markov chains X^{m_1} and X^{m_2} where each state refers to credit ratings of firm m_1 and firm m_2. Let $\mathcal{S}_1$ be the set of states where Markov chain X^{m_1} has ever visited for $t < T$, i.e.,

$$\begin{aligned} \mathcal{S}_1 &= \{i : S_{m_1}(i,j) > 0 \text{ for some } j\} \\ &= \{i : N_{ij} > 0 \text{ for some } j\} \subset \mathcal{S} \end{aligned}$$

Similarly, we let $\mathcal{S}_2 = \{i : S_{m_2}(i,j) > 0 \text{ for some } j\}$. The corresponding sequence matrices $\mathbf{S}_{m1}$ and $\mathbf{S}_{m2}$ for Markov chains X^{m_1} and X^{m_2} have the following property.

Theorem 2.1.1. *Suppose that there is no intersection between $\mathcal{S}_1$ and $\mathcal{S}_2$, i.e., $\mathcal{S}_1 \cap \mathcal{S}_2 = \emptyset$, implying two Markov chains X^{m_1} and X^{m_2} have not visited the same state. Then the Euclidean distance $\|\mathbf{S}_{m1} - \mathbf{S}_{m2}\|$ between $\mathbf{S}_{m1}$ and $\mathbf{S}_{m2}$ is*

$$\sqrt{2} \le \|\mathbf{S}_{m1} - \mathbf{S}_{m2}\| \le \sqrt{n}$$

Proof. Let $k = |\mathcal{S}_1| < n$. Without loss of generality, we may assume that

$$\begin{aligned} \mathcal{S}_1 &= \{1, 2, \ldots, k\} \\ &= \{i : S_{m_1}(i,j) > 0 \text{ for some } j\} \end{aligned}$$

Then[1] the sequence matrix $\mathbf{S}_{m1}$ is of the form

$$\begin{pmatrix} \mathbf{S}'_{m1} & \mathrm{O} \\ \mathrm{O} & \mathrm{O} \end{pmatrix}$$

where $\mathbf{S}'_{m1}$ is a $k \times k$ subsection of $\mathbf{S}_{m1}$ and O is the zero matrix. We have the Euclidean square distance between $\mathbf{S}'_{m1}$ and the zero matrix is

$$\|\mathbf{S}'_{m1}\|^2 \ge \frac{1}{k} + \frac{1}{k} + \cdots + \frac{1}{k} = \frac{k}{k} = 1$$

by using Lemma 2.1.1. Applying the same argument, the Euclidean square distance between $\mathbf{S}_{m2}$ and the zero matrix is greater than or equal to 1. Since $\mathcal{S}_1 \cap \mathcal{S}_2 = \emptyset$, we get

$$\|\mathbf{S}_{m1} - \mathbf{S}_{m2}\|^2 \ge 1 + 1 = 2$$

therefore, $\|\mathbf{S}_{m1} - \mathbf{S}_{m2}\| \ge \sqrt{2}$. On the other hand, by Lemma 2.1.1,

$$\sum_{j=1}^{k} S'_{m1}(i,j)^2 \le 1$$

for each $1 \le i \le k$ where $S'_{m1}(i,j)$ are entries of $\mathbf{S}'_{m1}$. Thus

$$\|\mathbf{S}_{m1}\|^2 = \|\mathbf{S}'_{m1}\|^2 \le 1 + 1 + \cdots + 1 = k.$$

Since $\mathcal{S}_1 \cap \mathcal{S}_2 = \emptyset$, we have $|\mathcal{S}_2| \le n - k$ and $\|\mathbf{S}_{m2}\|^2 \le n - k$. Then

$$\|\mathbf{S}_{m1} - \mathbf{S}_{m2}\|^2 = \|\mathbf{S}_{m1}\|^2 + \|\mathbf{S}_{m2}\|^2 \le k + (n - k) = n.$$

Therefore, $\|\mathbf{S}_{m1} - \mathbf{S}_{m2}\| \le \sqrt{n}$. □

Theorem 2.1.2. *Suppose that two Markov chains X^{m_1} and X^{m_2} have not made the same transition, i.e.,*

$$\{(i,j) : S_{m_1}(i,j) > 0\} \cap \{(i,j) : S_{m_2}(i,j) > 0\} = \emptyset.$$

Then the Euclidean distance $\|\mathbf{S}_{m1} - \mathbf{S}_{m2}\|$ between $\mathbf{S}_{m1}$ and $\mathbf{S}_{m2}$ is

$$\sqrt{2} \le \|\mathbf{S}_{m1} - \mathbf{S}_{m2}\| \le \sqrt{2n}.$$

Proof. The argument is essentially the same as in Theorem 2.1.1. For the upper bound,

$$\sum_{j=1}^{n} S_{m1}(i,j)^2 \le 1$$

for each $1 \le i \le n$, so $\|\mathbf{S}_{m1}\|^2 \le n$. Since $S_{m1}(i,j) \times S_{m2}(i,j) = 0$ for $1 \le i, j \le n$, we get

$$\|\mathbf{S}_{m1} - \mathbf{S}_{m2}\|^2 = \|\mathbf{S}_{m1}\|^2 + \|\mathbf{S}_{m2}\|^2 \le 2n.$$

Thus, $\|\mathbf{S}_{m1} - \mathbf{S}_{m2}\| \le \sqrt{2n}$. □

Definition 2.1.3. For a sequence vector $\mathbf{x}_m(T) = (X_1^m, X_2^m, \ldots, X_T^m)$ of length T for some firm m, it is said to be *ascending* if $X_{t+1}^m \le X_t^m$ for all $t < T$. The sequence vector $\mathbf{x}_m(T)$ is said to be $descending$ if $X_{t+1}^m \ge X_t^m$ for all $t < T$.

Note that there are at most two nonzero entries in each nonzero row for sequence matrices corresponding to ascending or descending sequence vectors. An ascending vector indicates the credit ratings of a firm have been upgraded while a descending vector implies the credit ratings have been downgraded. Credit rating sequences with ascending or descending sequences are examples of sequences that exhibit rating drift behaviour. As noted in D'Amico, Dharmaraja, Manca, and Pasricha (2019), rating drift is more pronounced in downgrades rather than upgrades.

2.2. Transition matrices for credit ratings

Credit rating sequences can be modelled as Markov processes. We consider a discrete n-state time-homogeneous Markov chain. A transition probability is the conditional probability of a stochastic process transitioning to one state given its current state, that is

$$Pr\{X_{t+1} = j | X_t = i\} \tag{2}$$

where $X_t \in \mathcal{S}$ is the credit rating at time $t \in \mathbb{N}$ with state space $\mathcal{S} = \{1, 2, \ldots, n\}$. A Markov chain must satisfy the Markov property. The Markov property is stated as the following

$$Pr\{X_{t+1} = j \mid X_0 = i_0, \ldots, X_t = i_t\} = Pr\{X_{t+1} = j \mid X_t = i_t\}. \tag{3}$$

The conditional probabilities $Pr\{X_{t+1} = j \mid X_t = i\}$ is called the one-step transition probability and can be arranged in a matrix $\mathbf{P}$ called the transition matrix. With $n = |\mathcal{S}|$, the transition matrix satisfies the following properties

$$P(i,j) \ge 0 \quad \forall i, j \le n, \tag{4}$$

$$\sum_{j=1}^{n} P(i,j) = 1 \quad \forall i \le n. \tag{5}$$

The Markov property implies that the transition probabilities only depend on its current state. To calculate the probability of transitions τ steps into the future we use the following theorem (see for instance, Taylor and Karlin (1998))

Theorem 2.2.1. *Let $P^\tau(i,j) = Pr\{X_{t+\tau} = j \mid X_t = i\}$ and $P(i,j)$ represent the entry of a transition matrix $\mathbf{P}$. Then τ-step transition probability $P^\tau(i,j)$ of transitioning from state i to j satisfy*

$$P^\tau(i,j) = \sum_{k=0}^{\infty} P(i,k) P^{(\tau-1)}(k,j) \tag{6}$$

where we define

$$P^{(0)}(i,j) = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{if } i \ne j. \end{cases}$$

Eq. (6) represents matrix multiplication and using transition matrices $\mathbf{P}$ we have the equivalent representation $\mathbf{P}^\tau = \mathbf{P} \times \mathbf{P}^{(\tau-1)}$. Given the assumption of time-homogeneity we can then write more generally

$$\mathbf{P}^\tau = \mathbf{P} \times \mathbf{P} \times \cdots \times \mathbf{P} = (\mathbf{P})^\tau. \tag{7}$$

Therefore, using Eq. (7) one can obtain the transition probabilities for any τ-step transition.

Given the clusters of credit rating sequences formed from the K-means algorithm one can generate transition matrices based on the members of the respective clusters. The industry standard for estimating transition matrices from credit rating sequences is the cohort approach (Christensen, Hansen, & Lando, 2004; Gunnvald, 2014).

Definition 2.2.1 (*Cohort Approach*)**.** Let $n = |\mathcal{S}|$, $\{\mathbf{x}_m(T) \mid m \le M_k \in \mathbb{N}, k = 1, 2, \ldots, K\}$ be the set of credit rating sequences in cluster k with a total of M_k members, $\mathcal{T} = \{t_l \mid 0 \le l \le T$ with $t_l < t_{l+1}\}$ be the set of equally spaced observed time points of the credit ratings $\mathbf{x}_m(T)$ used in constructing the representative transition matrix. Then we define the

[1] This might not be the case when $X_T^{m_1} = j$ and $X_t^{m_1} \ne j$ for all $t < T$. We exclude this sequence here. In this special case, we have a slightly different bound depending on k.

transition period Δt by $t_{l+1} - t_l$. That is, Δt is the shortest period between any two time points. Then the representative Δt-year transition matrix of the kth cluster, $\mathbf{P}_k(\Delta t)$ is an $n \times n$ matrix whose entries are denoted by

$$P_k(i,j) = \frac{\sum_{t_l \in \mathcal{T}} N_{ij}(t_l)}{\sum_{t_l \in \mathcal{T}} N_i(t_l)}, \tag{8}$$

where $N_{ij}(t_l)$ is the number of companies that had transitioned from state i to j in the Δt period, $N_i(t_l)$ is the total number of companies whose current state was i at time t_l.

To generate longer period Δt-year transition matrices one can redefine Δt by $t_{l+2} - t_l$ instead. Consequently, a drawback of the cohort approach is the possibility of completely missing the existence of a particular credit rating in time if we choose to sample points when Δt is large. For example, suppose we have a credit rating sequence $X = (1,5,1,1,1,1,1)$ with times $t_0, t_1, t_2, t_3, t_4, t_5, t_6$ corresponding to the dates 2000, 2001, 2002, 2003, 2004, 2005, 2006. In our example, we will generate our transition matrices by sampling time points t_0, t_2, t_4, and t_6. The observed credit rating sequence used in the estimation of the transition matrix is then $X_{obs} = (1,1,1,1)$. Therefore, the transition to and from credit rating 5 will be completely missed using the estimated 2-year transition matrix.

An alternative approach that captures the intermediate transitions is the duration approach. The duration approach first estimates the transition matrix by taking the matrix exponential of an estimated generator matrix (Gunnvald, 2014; Lando & Skodeberg, 2002). The difference between the cohort and duration approach has been intensively studied by Jafry and Schuermann (2004). They have noted that the cohort approach overestimates default probabilities (the last column of the transition matrix) for less risky rating categories and underestimates default probabilities for the most risky rating categories. By generating a bond portfolio of 400 exposures, they have also concluded that ignoring the efficiency gain in the duration approach is more damaging.

2.3. Clustering algorithm: K-means

In our model, we will be making use of a variant of the K-means algorithm to cluster our data set. The purpose of clustering is to partition the firms into groups that share similar transition behaviours in their respective credit rating sequences. The base K-means algorithm is a popular choice in many applications due to its ease of implementation, simplicity, efficiency, and empirical success (Jain, 2010). Given the assumption that credit rating sequences can be modelled by an n-state time-homogeneous Markov chain, it may be natural to immediately consider clustering transition matrices of the individual firms. Unfortunately, using Euclidean distance to compare transition matrices leads to the mis-clustering of firms who do not experience any transitions in their credit rating sequence but belong to different credit ratings at the same time. For example, let $\mathbf{x}_1 = (2,2,2,2,2)$ and $\mathbf{x}_2 = (5,5,5,5,5)$ then the respective transition matrices $\mathbf{P}'_1$, and $\mathbf{P}'_2$ are

$$\mathbf{P}'_1 = \begin{bmatrix} 1&0&0&0&0&0&0\\ 0&1&0&0&0&0&0\\ 0&0&1&0&0&0&0\\ 0&0&0&1&0&0&0\\ 0&0&0&0&1&0&0\\ 0&0&0&0&0&1&0\\ 0&0&0&0&0&0&1 \end{bmatrix}, \quad \mathbf{P}'_2 = \begin{bmatrix} 1&0&0&0&0&0&0\\ 0&1&0&0&0&0&0\\ 0&0&1&0&0&0&0\\ 0&0&0&1&0&0&0\\ 0&0&0&0&1&0&0\\ 0&0&0&0&0&1&0\\ 0&0&0&0&0&0&1 \end{bmatrix}.$$

As the above matrices are transition matrices they must satisfy properties (4) and (5) and hence the resulting matrices are the identity matrix. Calculating the Euclidean distance we have $\left\|\mathbf{P}'_1 - \mathbf{P}'_2\right\| = 0$ despite being two firms with completely different credit rating sequences. If instead we generated the sequences matrices $\mathbf{S}_1$ and $\mathbf{S}_2$ for $\mathbf{x}_1$ and $\mathbf{x}_2$ respectively

$$\mathbf{S}_1 = \begin{bmatrix} 0&0&0&0&0&0&0\\ 0&1&0&0&0&0&0\\ 0&0&0&0&0&0&0\\ 0&0&0&0&0&0&0\\ 0&0&0&0&0&0&0\\ 0&0&0&0&0&0&0\\ 0&0&0&0&0&0&0 \end{bmatrix}, \quad \mathbf{S}_2 = \begin{bmatrix} 0&0&0&0&0&0&0\\ 0&0&0&0&0&0&0\\ 0&0&0&0&0&0&0\\ 0&0&0&0&0&0&0\\ 0&0&0&0&1&0&0\\ 0&0&0&0&0&0&0\\ 0&0&0&0&0&0&0 \end{bmatrix},$$

using the Euclidean distance between $\mathbf{S}_1$ and $\mathbf{S}_2$ results in $\left\|\mathbf{S}_1 - \mathbf{S}_2\right\| = \sqrt{2}$. Hence, by using sequence matrices, we circumvent this problem as two firms having experienced no transitions over the period T will be considered "far" from each other in Euclidean norm. Therefore, by using the Euclidean distance measure on sequence matrices instead of transition matrices, the differences between the resulting clusters will have a more intuitive interpretation.

In the conventional K-means clustering algorithm the initial cluster centroids are chosen completely randomly. Because the resulting clusters are highly dependent on the initial cluster centroids, the initialization of the cluster centroids is an important question to answer. To improve the initialization of the clustering process we choose to instead use PCA-guided K-means (Xu, Ding, Liu, & Luo, 2015). The idea of PCA-guided K-means is that the optimal solution to the minimization problem lies in space known as the PCA-subspace, a smaller space than the original space. To implement this algorithm, we first cluster our data in the PCA-subspace and then initialize our cluster centroids in the feature space based on the cluster membership in the PCA-subspace. Although the resulting solution is not guaranteed to be the global optimal solution, the resulting solution tends to be better (in terms of within cluster variance) than the solutions obtained by just searching within the full data space. Therefore, using the PCA-guided K-means algorithm we intend to partition M firms into K clusters such that firms in each cluster share similar credit rating transition behaviour.

3. Long-term credit rating model

We will be testing the performance of the model against three different classification scenarios, (1) the prediction of future credit rating, (2) the prediction of the direction of future credit rating transitions, and (3) the classification of risky firms most likely to default. For each classification scenario we first split the entire data set into a training set and a test set. The model is trained using the training set and is evaluated based on its classification performance using the test set. García, Marqués, and Sánchez (2015) highlights the importance of experimental design in credit scoring and bankruptcy prediction. They note that the choice of data splitting method is dependent on the nature of the classifiers and complexity of the problem. In our study, we found that the K-fold cross-validation method suited our goals well. We forgo the use of the single holdout method as this results in our model producing a single set of representative transition matrices. For similar reasons, we forgo the use of leave-one-out cross-validation as the set of representative transition matrices may remain relatively unchanged by removing a single sequence matrix from the training set. By using K-fold cross-validation we test the effectiveness of our model in the case of a variety of different clusters and the predictive power of their representative transition matrices. After clustering the training set we generate the representative transition matrix $\mathbf{P}_k(\Delta t)$ for each cluster. This is done using the cohort approach.

For each scenario we let $\mathbf{x}_m(T)$ be the credit rating sequence of the mth firm from the test set. For each firm m, we generate the sequence matrix $\mathbf{S}_m$ based on $\mathbf{x}_m(T)$. The sequences matrices of the training set are then partitioned into K different clusters. Using the testing set, we assign a single firm to one of the K clusters based on the Euclidean distance between $\mathbf{S}_m$ and the clusters' centroid $\boldsymbol{\mu}_k$. That is, the assigned cluster k^* is chosen by

$$k^* = \text{argmin}_k \left\| \mathbf{S}_m - \boldsymbol{\mu}_k \right\| . \tag{9}$$

After assigning the firm to a cluster, we can estimate the future behaviour of the firm by using the cluster's representative transition matrix.

3.1. Credit rating and transition direction prediction

In the credit rating prediction scenario, we intend to determine the most likely credit rating a firm will take at time t' in the future given the current credit rating X_t at time t. We will consider n class labels for an n-state homogeneous Markov chain, i.e. $\mathcal{S} = \{1, 2, \dots, n\}$. Let τ be the difference $t' - t$, then, using the representative transition matrix $\mathbf{P}_k(\Delta t)$, we calculate the τ-step transition matrix $\mathbf{P}_k^{\tau}(\Delta t)$. Because we assume time-homogeneity, the τ-step transition matrix can be calculated by using Eq. (7), that is

$$\mathbf{P}_k^{\tau}(\Delta t) = (\mathbf{P}_k(\Delta t))^{(t'-t)}. \tag{10}$$

The prediction of the future credit rating $\hat{X}_{t'}^m$ for some firm m is then

$$\hat{X}_{t'}^m = \text{argmax}_j P_k^{\tau}(X_t^m, j). \tag{11}$$

For the evaluation of multi-class classification performance we will be considering the number of true positives (TP), false positives (FP), false negatives (FN), and true negatives (TN). In the binary case we can organize these counts using a confusion matrix

		Predicted class	
		Positive	Negative
Actual class	Positive	TP	FN
	Negative	FP	TN

where our classes are the "Positive" and "Negative" classes. Given the confusion matrix it can then be observed that the count of TP represent the number of correct prediction of the positive class, FP represents the number of incorrect predictions of the positive class, FN represents the number of incorrect predictions of the negative class, and TN represents the number of correct prediction for the negative class. For the multi-class classification scenario with classes A, B, C, and D we can generate the following confusion matrix.

		Predicted class			
		A	B	C	D
Actual class	A	$c_{1,1}$	$c_{1,2}$	$c_{1,3}$	$c_{1,4}$
	B	$c_{2,1}$	$c_{2,2}$	$c_{2,3}$	$c_{2,4}$
	C	$c_{3,1}$	$c_{3,2}$	$c_{3,3}$	$c_{3,4}$
	D	$c_{4,1}$	$c_{4,2}$	$c_{4,3}$	$c_{4,4}$

Then, the number of TP, FP, FN, and TN can be calculated for each individual class in a similar manner to the binary example by treating one of our classes as the positive class and everything else as the negative. Generally, let $\mathcal{L} = \{L_l \mid 1 \leq l \leq L\}$ be the set of class labels, then to calculate the number of TP, FP, FN, and TN for class label $L_l \in \mathcal{L}$ we must consider the confusion matrix for classes "L_l" and "Non-L_l"

		Predicted class	
		L_l	Non-L_l
Actual class	L_l	TP_l	FN_l
	Non-L_l	FP_l	TN_l

where each cell of the above confusion matrix can be calculated by the following equations

$$\text{TP}_l = c_{l,l} \tag{12}$$

$$\text{FP}_l = \sum_{i=1}^{L} c_{i,l}, \text{ for } i \neq l \tag{13}$$

$$\text{FN}_l = \sum_{j=1}^{L} c_{l,j}, \text{ for } j \neq l \tag{14}$$

$$\text{TN}_l = \sum_{i=1}^{L} \sum_{j=1}^{L} c_{i,j} - (\text{TP}_l + \text{FP}_l + \text{FN}_l). \tag{15}$$

To evaluate the wellness of the estimates made in the multi-class classification scenario we will be using the average accuracy, denoted by AA, and the micro-averaged $F1$-measure, denoted by $F1_\mu$ (Sokolova & Lapalme, 2009).

$$AA = \frac{\sum_{l=1}^{n} \frac{\text{TP}_l + \text{TN}_l}{\text{TP}_l + \text{FN}_l + \text{FP}_l + \text{TN}_l}}{L} \tag{16}$$

$$F1_\mu = 2 \frac{Pr_\mu \cdot Re_\mu}{Pr_\mu + Re_\mu} \tag{17}$$

where

$$Pr_\mu = \frac{\sum_{l=1}^{L} \text{TP}_l}{\sum_{l=1}^{L} (\text{TP}_l + \text{FP}_l)} \tag{18}$$

$$Re_\mu = \frac{\sum_{l=1}^{L} \text{TP}_l}{\sum_{l=1}^{L} (\text{TP}_l + \text{FN}_l)} \tag{19}$$

where Pr_μ and Re_μ are the micro-averaged precision and recall respectively. It should also be noted that when using micro-averaging for multi-class classification the micro-averaged recall, micro-average precision, and micro-average $F1$-score are equal to each other. Micro-averaging is used instead of macro-averaging (that is, averaging the precision, recall, and $F1$ across the L classes respectively) because macro-averaging weights each class's precision, recall, and $F1$-score equally across the classes while micro-averaging takes into consideration the size of each of the classes for the respective measure. This prevents the smaller classes from over contributing in the averaging of the $F1$-score (Sokolova & Lapalme, 2009).

When using K-fold cross-validation, a total confusion matrix is calculated by summing up the K confusion matrices that were generated at each fold. This total confusion matrix is then used to calculate $\text{TP}_l, \text{FP}_l, \text{FN}_l$, and TN_l as this is the most unbiased method in computing the $F1$-measure when there is a high class imbalance (Forman & Scholz, 2010).

In the transition direction prediction scenario, we intend to determine which direction a firm's credit rating will move in by time t' in the future, given the current credit rating X_t^m at time t. We define this set of class labels as $\mathcal{L} = \{-1, 0, 1\}$ where -1, 0, and 1 represent the downgrade, stay, and upgrade classes respectively. Using the representative transition matrix $\mathbf{P}_k(\Delta t)$, we calculate the τ-step transition matrix $\mathbf{P}_k^{\tau}(\Delta t)$. The prediction of the direction that firm m's credit rating will change at time t' is then estimated by $\hat{d}_m(t, t')$ where

$$\hat{d}_m(t, t') = \begin{cases} 1 & \text{if} \quad Pu = \max(Pu, Ps, Pd) \\ 0 & \text{if} \quad Ps = \max(Pu, Ps, Pd) \\ -1 & \text{if} \quad Pd = \max(Pu, Ps, Pd) \end{cases} \tag{20}$$

where

$$Pu = \sum_{j < X_t^m} P_k^{\tau}(X_t^m, j)$$

$$Ps = P_k^{\tau}(X_t^m, X_t^m)$$

$$Pd = \sum_{X_t^m < j} P_k^{\tau}(X_t^m, j).$$

The predicted estimate of the change in the direction of the credit rating $\hat{d}_m(t,t')$ is then compared to the true change in direction $d_m(t,t')$, calculated by

$$d_m(t,t') = \begin{cases} 1 & \text{if} \quad X^m_{t'} < X^m_t \\ 0 & \text{if} \quad X^m_{t'} = X^m_t \\ -1 & \text{if} \quad X^m_{t'} > X^m_t. \end{cases} \tag{21}$$

To evaluate the wellness of the estimates made using the test set, we again calculate a total confusion matrix from the K-fold cross-validation and calculate the average accuracy using Eq. (16) and micro-average $F1$-score using Eq. (17).

3.2. Prediction of default behaviour

In the default behaviour prediction scenario, we intend to determine whether a firm will be in default within τ years based on their current credit rating. We do so by checking the probability of default of a firm against an appropriate threshold enabling the classification of the firm's default behaviour. That is, whether the firm will be in default or not within τ years. The class labels that we will consider are binary. We define the class label set as $\mathcal{L} = \{1, 0\}$ where 1 represent a firm having defaulted within τ years and 0 for a firm not defaulting within τ years.

Using the representative transition matrix $\mathbf{P}_k(\Delta t)$ of the kth cluster we calculate the probability of defaulting within the next τ years. To calculate the probability of default within τ time steps we first define the following

Definition 3.2.1. Given state space $\mathcal{S}$ and $n = |\mathcal{S}|$, let $\mathbf{P}_k(\Delta t)$ be the $n \times n$ representative Δt-year transition matrix of cluster k. Then $\mathbf{Q}_k$ is the $(n-1) \times (n-1)$ subsection of $\mathbf{P}_k(\Delta t)$ containing entries $P_k(i,j)$ for $1 \leq i, j < n$ and $\mathbf{R}_k$ is a vector of size $n-1$ containing entries $P_k(i,j)$ for $1 \leq i < n$ and $j = n$.

Given the subsections $\mathbf{Q}_k$ and $\mathbf{R}_k$ as defined above, we then denote $\mathbf{r}^\tau_k$ as the $(n-1) \times 1$ vector whose entries are the probability of default within τ years for a firm assigned to cluster k and is calculated by

$$\mathbf{r}^\tau_k = (\mathbf{I} + \mathbf{Q}_k + \mathbf{Q}^2_k + \cdots + \mathbf{Q}^{(\tau-1)}_k)\mathbf{R_k} \tag{22}$$

where $\mathbf{I}$ is the identity matrix. The entries of $\mathbf{r}^\tau_k$ are denoted by $r^\tau_k(i)$ for $1 \leq i < n$. Given a firm m that was assigned to cluster k, the probability of default within τ years based on the firm's current credit rating X^m_t is then $r^\tau_k(X^m_t)$. Once a firm has been assigned a probability of default, we will refer to the associated probability as a "risk score". At every step of the K-fold cross-validation process we assign all the firms in each of the fold's respective test set a risk score. By the Kth fold of the cross-validation process, all of the firms in the data set will be assigned a risk score.

To measure the quality of the classification of the firms' default state, we will assess our model based on two different evaluation measures. The first evaluation is done using the measure Somers' Delta (Somers' D). The measure Somers' D is an asymmetric measure of association between an independent (x) and dependent variable (y) (Somers, 1962; Trueck & Rachev, 2009). Somers' D measures this association between independent and dependent variables by considering the number of concordant pairs, the number of discordant pairs, and the number of tied pairs on the dependent variable.

Definition 3.2.2 (*Somers' D*)**.** Let C be the number of concordant pairs, D be the number of discordant pairs, and Y_0 be the number of tied pairs on the dependent variable. A pair (x_i, y_i) and (x_j, y_j) is concordant when both $x_i > x_j$ and $y_i > y_j$. A pair is discordant when $x_i > x_j$ and $y_i < y_j$. A pair is tied on the dependent variable when $y_i = y_j$. Somers' D is then calculated as

$$d_{yx} = \frac{C - D}{C + D + Y_0} \tag{23}$$

so the value of d_{yx} ranges from -1 to 1.

The operational interpretation of Somers' D is the measure of the proportionate excess of concordant over discordant pairs among the number of pairs not tied on the independent variable. Somers' D can be applied in two types of applications (Newson, 2006). To measure the effect of the independent variable on the dependent variable, treating d_{yx} as a measure of "effect size", or, to measure the performance of the independent variable as a predictor of the dependent variable, treating d_{yx} as "predictor performance indicator". Given the context of credit risk we let the estimated risk score be the independent variable and the true default status (whether a firm has indeed defaulted within the next τ time step) as the dependent variable.

The second evaluation is done by setting a threshold value θ and then classifying a firm as being in default or not based on whether their risk score exceeds the chosen threshold value. This comparison and classification is done using threshold values from a discretized interval ranging from 0 and 1. Hence, for each θ chosen, the mth firm can be assigned a predicted default behaviour y_m based on $r^\tau_k(X^m_t)$

$$\hat{y}_m = \begin{cases} 1 & \text{if} \quad r^\tau_k(X^m_t) > \theta, \\ 0 & \text{if} \quad r^\tau_k(X^m_t) \leq \theta. \end{cases} \tag{24}$$

A convenient way of displaying how the performance of a classification model is by using a receiver operator characteristic (ROC) curve. A ROC curve is a plot of the true positive rate, TPR (also known as the recall) against the false positive rate, FPR. To construct this curve we select a threshold θ, estimate $\hat{y}_m$ using Eq. (24), generate a confusion matrix, and then calculate the FPR and TPR by

$$FPR = \frac{\text{FP}}{\text{FP} + \text{TN}} \tag{25}$$

and

$$TPR = Re = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{26}$$

This process is done for all thresholds from 0 to 1. At the same time the precision and $F1$-score value can be calculated from the confusion matrix at every threshold.

In practice, it is more useful to choose an optimal threshold or "cut-off" point for binary classification. By choosing a threshold we set the rate for Type I and Type II errors. In the context of credit risk, a Type I error can result in opportunity costs and lost potential profits from lost interest income, while a Type II error can result in the lost interest and principle through defaults (Trueck & Rachev, 2009). Liu (2002) calculates the optimal threshold by taking the line tangent to the ROC curve. This tangent line has a slope that is proportional to the ratio of "good" and "bad" cases, and inversely proportional to the cost ratio of the Type I and Type II errors.

In general it is difficult to use costs to evaluate models as different institution have different cost and pay-off structures and so, it would be challenging to present a single cost function and provide a general framework for optimal decision making of a financial institution (Trueck & Rachev, 2009). Instead, we will be using the methods described in Sanchez (2016) to determine the optimal threshold in the worst-case scenario for the purpose of model evaluation. By using game theory and treating the classifier and "nature" as players, we choose the optimal threshold at the point where the ROC curve and the descending diagonal line (i.e. the line $TPR = 1 - FPR$) intersects.

4. Data and experimental methods

In this section we describe the data and methods used in this study. First, we present the data, its characteristics, and how the data was processed before classification. Next, we define the model parameters and present the algorithm used to evaluate the three classification scenarios described in Section 3.

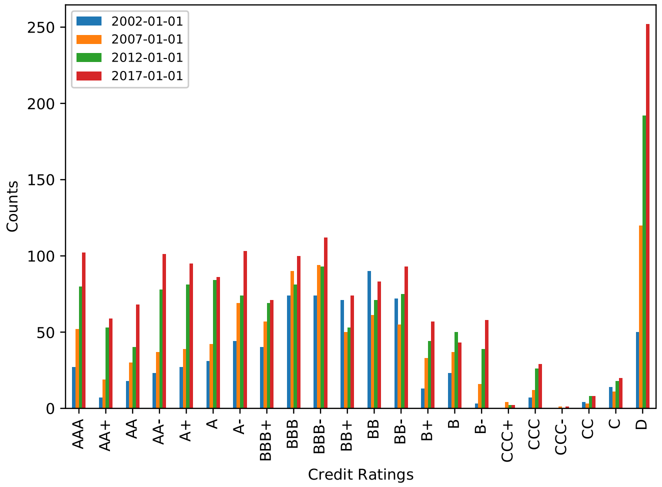


Fig. 1. The frequency distribution of the 22 class credit ratings for the years 2002, 2007, 2012, 2017.

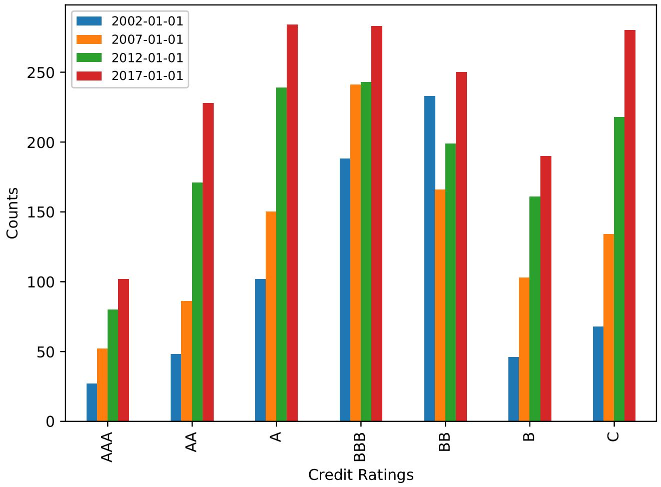


Fig. 2. The frequency distribution of the 7 class credit ratings for the years 2002, 2007, 2012, 2017.

Table 1
The 7 aggregated classes.

New ratings	Old ratings
AAA	{AAA}
AA	{AA+, AA, AA−}
A	{A+, A, A−}
BBB	{BBB+, BBB, BBB−}
BB	{BB+, BB, BB−}
B	{B+, B, B−, CCC+, CCC, CCC−}
C	{CC, C, D}

Table 2
The number of valid firms used in the 5-fold cross-validation based on the input date.

Input date	Number of valid firms
2000-01-01	542
2001-01-01	590
2002-01-01	712
2003-01-01	752

Table 3
The class labels for each classification scenario.

Scenario	Class label set ($\mathcal{L}$)
Credit rating prediction	{1,2,3,4,5,6,7}
Transition direction	{−1, 0, 1}
Default	{1, 0}

4.1. Data

The data set we will be using was collected and provided by National Information & Credit Evaluation Inc., a major bond-rating company in Korea. The data set consisted of monthly corporate credit ratings from 1986-09-01 to 2018-09-01 for 1899 firms in Korean indices such as the KOSDAQ and KOSPI. Firms in this data set can take any rating from the following set of 22 credit ratings

{AAA, AA+, AA, AA-, A+, A, A-, BBB+, BBB, BBB-,
BB+, BB, BB-, B+, B, B-, CCC+, CCC, CCC-, CC, C, D}.

The firms that take the "D" rating are considered to be in default. Some firms were "closed" after some time and are considered to be in default. Firms that were missing credit rating sequences, made for sale, or was merged with another firm were removed from the data set. After pruning the data set, there are 1648 firms remaining in the data set. The distribution of the remaining 1648 firms' credit rating classes for selected dates can be found in Fig. 1.

From Fig. 1, it can be observed that there are very few samples in credit class CCC+, CCC, CCC-, CC, and C for the selected dates. We mitigate the negative effects of this imbalanced data set by combining similar categories together reducing the number of credit rating classes from 22 to 7 classes. Doing so will minimize the number of classes that contain low instances of that minority class. The particular mapping of old classes to new classes can be found in Table 1.

The distribution of the new aggregated classes can be found in Fig. 2. By reducing the number of classes to 7 we diminish the degree of imbalance that was present in the data set. Something to note is that the distribution of credit ratings appear to change dramatically from year to year. This is in part due to the fact that a number of firms were not rated or did not exist at that time. For example, only 712 firms were rated on 2002-01-01 where as 1617 firms were rated on 2017-01-01. For the credit rating prediction and transition direction classification scenarios, we will be using the relabelled credit rating sequences as outlined in Table 1. For the default prediction classification scenario we will move the old ratings CC and C to the new B rating group leaving the last class to represent default by containing exclusively D ratings.

4.2. Experimental method

We treat the credit rating sequences as a Markov process with a state space $\mathcal{S} = \{1, 2, 3, 4, 5, 6, 7\}$ with the numbers 1 representing the least risky credit class AAA and 7 representing the most risky credit class C. The total number of clusters K was set to 15.

For each classification scenario, we set the input date $t \in \{2000, 2001, 2002, 2003\}$. The input date represents the initial point in time we will begin making our prediction from. For quarterly transition matrices we set Δt to be 0.25. The predictions will be made τ years into the future where $\tau \in \{5, 10, 15\}$. Credit rating sequences with fewer than 5 years of credit rating data will not be used and excluded from the analysis. Therefore, of the remaining 1648 firms number of valid firms was reduced to the amounts indicated in Table 2. Each firm is assigned a sequence matrix $\mathbf{S}_m$ based on its credit rating sequence $\mathbf{x}_m(T)$ as described in Section 2.1. Using 5-fold cross-validation we split the number of valid firms into two groups, a training set, and test set. 15 clusters are then generated based off the training set using PCA-guided K-means. A representative Δt-transition matrix $\mathbf{P}_k(\Delta t)$ was estimated for each cluster where $\Delta t = 0.25$. The transition matrices are generated as described in Section 2.2. Given $\mathbf{P}_k(\Delta t)$ we can then calculate $\mathbf{P}_k^{\tau}(\Delta t)$ where $\tau \in \{5, 10, 15\}$ using Eq. (10).

Each firm from the test set is then assigned to a cluster and assigned a predicted class from a class label set based on the classification

Table 4
Results from the credit rating prediction scenario. Results under the column label (C) and (B) represent the results from clustering and the benchmark model respectively.

τ	Input date	Predicted date	(C) AA	(B) AA	(C) $F1_{\mu}$	(B) $F1_{\mu}$
15	2000-01-01	2015-01-01	**0.8993 (0.0036)**	0.8585 (0.0027)	**0.6475 (0.0126)**	0.5047 (0.0095)
	2001-01-01	2016-01-01	**0.9013 (0.0033)**	0.8705 (0.0011)	**0.6546 (0.0115)**	0.5467 (0.0038)
	2002-01-01	2017-01-01	**0.9148 (0.0028)**	0.8712 (0.0020)	**0.7018 (0.0097)**	0.5491 (0.0070)
	2003-01-01	2018-01-01	**0.9178 (0.0026)**	0.8816 (0.0016)	**0.7125 (0.0090)**	0.5855 (0.0058)
10	2000-01-01	2010-01-01	**0.9012 (0.0029)**	0.8751 (0.0001)	**0.6542 (0.0102)**	0.5627 (0.0004)
	2001-01-01	2011-01-01	**0.9039 (0.0026)**	0.8833 (0.0001)	**0.6636 (0.0092)**	0.5915 (0.0003)
	2002-01-01	2012-01-01	**0.9126 (0.0021)**	0.8748 (0.0002)	**0.6942 (0.0075)**	0.5617 (0.0007)
	2003-01-01	2013-01-01	**0.9202 (0.0022)**	0.8860 (0.0003)	**0.7205 (0.0076)**	0.6009 (0.0011)
5	2000-01-01	2005-01-01	**0.9207 (0.0024)**	0.9135 (0.0001)	**0.7225 (0.0084)**	0.6974 (0.0003)
	2001-01-01	2006-01-01	**0.9183 (0.0023)**	0.9138 (0.0000)	**0.7140 (0.0081)**	0.6983 (0.0000)
	2002-01-01	2007-01-01	**0.9244 (0.0021)**	0.9097 (0.0000)	**0.7355 (0.0075)**	0.6840 (0.0000)
	2003-01-01	2008-01-01	**0.9264 (0.0018)**	0.9179 (0.0000)	**0.7425 (0.0062)**	0.7128 (0.0000)

scenario as shown in Table 3. This process is done for all the test sets at each fold. The result is that each valid firm is assigned to a predicted class by the end of the 5-fold cross-validation process. The effectiveness of our clustering model against the benchmark model will be based on the performance measures described in Section 3. The benchmark model uses a single representative transition matrix $\mathbf{P}(\Delta t)$ estimated from all of the credit rating sequences, in the absence of clustering. This single transition matrix is then used for classification purposes. The results in Section 5 were calculated by averaging 1000 shuffled 5-fold cross-validation results.

The algorithm used in the different classification scenarios can be grouped into 2 main algorithms. For the credit rating prediction and transition direction prediction classification scenarios:

1. Initialize the number of clusters K, $\mathcal{S} = \{1, 2, 3, 4, 5, 6, 7\}$, $n = |\mathcal{S}|$, t, Δt, and τ. Under the credit rating prediction scenario set the class label set as $\mathcal{L} = \{1, 2, 3, 4, 5, 6, 7\}$. Under the transition direction prediction scenario set the class label set as $\mathcal{L} = \{-1, 0, 1\}$.
2. From the data set of credit rating sequences, generate each firm's corresponding sequence matrices.
3. Begin 5-fold cross-validation process and split the data set into a training and test set:
    (a) Cluster in PCA subspace.
    (b) Set initial centroids based on cluster membership from clustering in the PCA subspace.
    (c) Cluster in full data space.
    (d) Generate the representative transition matrices $\mathbf{P}_k(\Delta t)$ for each cluster.
    (e) For all firms in the test set, assign to a cluster based on the firm's $\mathbf{S}_m$.
    (f) Under the credit rating prediction scenario, estimate the future credit rating $\hat{X}^m_{t'}$ based on $\mathbf{P}^{\tau}_k(\Delta t)$. Under the transition direction prediction scenario, find $\hat{d}_m(t, t')$ based on $\mathbf{P}^{\tau}_k(\Delta t)$.
    (g) Repeat step 3 for all 5 folds.
4. Generate a confusion matrix based on the predicted and actual classes.
5. Evaluate the performance based on Section 3.1. End.

The algorithm for the default prediction scenario is as follows:

1. Initialize the number of clusters K, $\mathcal{S} = \{1, 2, 3, 4, 5, 6, 7\}$, $n = |\mathcal{S}|$, t, Δt, and τ. Set the class label set as $\mathcal{L} = \{1, 0\}$.
2. From the data set of credit rating sequences, generate each firm's corresponding sequence matrices.
3. Begin 5-fold cross-validation process and split the data set into a training and test set:
    (a) Cluster in PCA subspace.
    (b) Set initial centroids based on cluster membership from clustering in the PCA subspace.
    (c) Cluster in full data space.
    (d) Generate the representative transition matrices $\mathbf{P}_k(\Delta t)$ for each cluster.
    (e) Use Definition 3.2.1 to generate the default probabilities $\mathbf{r}^{\tau}_k$ for each cluster based on $\mathbf{P}_k(\Delta t)$.
    (f) For all firms in the test set, assign to a cluster based on the firm's $\mathbf{S}_m$.
    (g) Based on the current rating of test set firms at time t, use the calculated $\mathbf{r}^{\tau}_k$ to assign a risk score $r^{\tau}_k(X^m_t)$ to every test firm.
    (h) Repeat step 3 for all 5 folds.
4. Proceed to steps 5 to calculate the Somers' D or step 6 to predict default behaviour.
5. Estimate d_{yx} for every firm. End.
6. Compare the threshold and the risk scores to estimate the default behaviour, $\hat{y}_m$.
7. Generate a confusion matrix based on the predicted and actual classes.
8. Evaluate the performance based on Sections Section 3.2. End.

5. Results and discussions

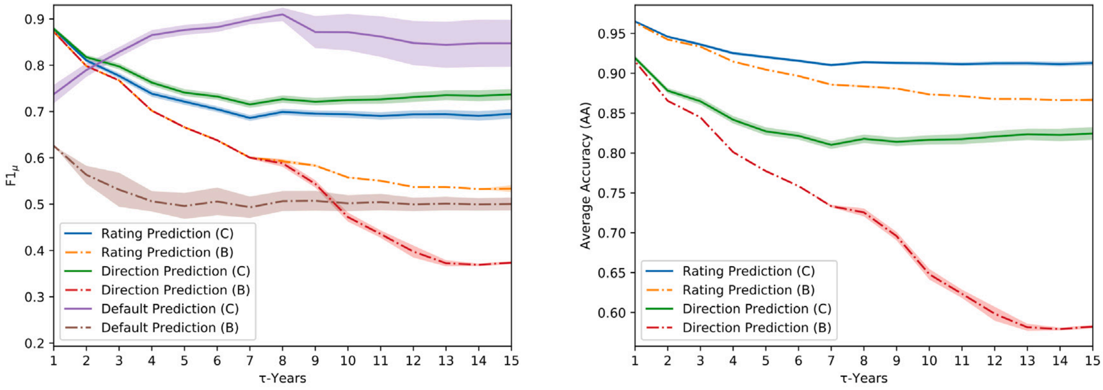
The brackets beside the performance measures in the following tables are the standard deviation of the respective measures. The low standard deviation for the benchmark model is the result of the diagonally dominant matrices produced by the benchmark model. From Table 4, the clustering model appears to outperform the benchmark model in terms of both the AA and the $F1_{\mu}$-measure in the credit rating prediction scenario. Predictions were made using representative transition matrices where $\tau \in \{5, 10, 15\}$. Due to the imbalanced nature of the data set, the significance of the micro-averaged $F1$-measure should have higher precedence over the averaged accuracy. It is a more accurate representation of the model's performance as it takes into account the size of the individual classes in $\mathcal{S}$.

Similar to the credit rating prediction scenario, the results in Table 5 show that the clustering model outperforms the benchmark model. This is observed for both the AA and the $F1_{\mu}$-measure when making predictions using representative transition matrices with $\tau \in \{5, 10, 15\}$.

The main diagonal of the transition matrices estimated from credit ratings tend to be diagonally dominant (Jafry & Schuermann, 2004). The main diagonal represents the probability that a firm maintains its current credit rating after a transition period. This observation can be commonly found in the benchmark model as it aggregates all of the training data before estimating the transition matrix. Because the transition matrix is used during the classification process, a consistently

Table 5
Results from the transition direction prediction scenario. Results under the column label (C) and (B) represent the results from clustering and the benchmark model respectively.

τ	Input date	Predicted date	(C) AA	(B) AA	(C) $F1_{\mu}$	(B) $F1_{\mu}$
15	2000-01-01	2015-01-01	**0.7992 (0.0089)**	0.5671 (0.0053)	**0.6987 (0.0133)**	0.3506 (0.0080)
	2001-01-01	2016-01-01	**0.8041 (0.0089)**	0.5769 (0.0077)	**0.7062 (0.0134)**	0.3654 (0.0115)
	2002-01-01	2017-01-01	**0.8293 (0.0084)**	0.5746 (0.0044)	**0.7440 (0.0125)**	0.3619 (0.0066)
	2003-01-01	2018-01-01	**0.8330 (0.0077)**	0.5899 (0.0054)	**0.7495 (0.0116)**	0.3848 (0.0081)
10	2000-01-01	2010-01-01	**0.7945 (0.0076)**	0.6371 (0.0040)	**0.6917 (0.0114)**	0.4557 (0.0060)
	2001-01-01	2011-01-01	**0.7962 (0.0065)**	0.6647 (0.0078)	**0.6943 (0.0097)**	0.4970 (0.0116)
	2002-01-01	2012-01-01	**0.8152 (0.0065)**	0.6543 (0.0054)	**0.7228 (0.0097)**	0.4814 (0.0080)
	2003-01-01	2013-01-01	**0.8270 (0.0066)**	0.6772 (0.0059)	**0.7406 (0.0099)**	0.5158 (0.0088)
5	2000-01-01	2005-01-01	**0.8246 (0.0060)**	0.7983 (0.0000)	**0.7370 (0.0091)**	0.6974 (0.0000)
	2001-01-01	2006-01-01	**0.8189 (0.0063)**	0.7989 (0.0002)	**0.7284 (0.0095)**	0.6983 (0.0003)
	2002-01-01	2007-01-01	**0.8338 (0.0052)**	0.7893 (0.0000)	**0.7507 (0.0079)**	0.6840 (0.0000)
	2003-01-01	2008-01-01	**0.8330 (0.0046)**	0.8085 (0.0000)	**0.7495 (0.0069)**	0.7128 (0.0000)



(a) $F1$-measure results.

(b) AA measure results.

Fig. 3. The results from the credit rating, rating transition direction, and default prediction scenarios. The labels (C) and (B) represent the clustering and benchmark models respectively.

Table 6
Results of the Somers' D for the default behaviour prediction scenario. Results under the column label (C) and (B) represent the results from clustering and the benchmark model respectively.

τ	Input date	Predicted date	(C) d_{yx}	(B) d_{yx}
15	2000-01-01	2015-01-01	**0.2836 (0.0210)**	0.1859 (0.0031)
	2001-01-01	2016-01-01	**0.2850 (0.0196)**	0.1918 (0.0029)
	2002-01-01	2017-01-01	**0.3470 (0.0194)**	0.2137 (0.0028)
	2003-01-01	2018-01-01	**0.3506 (0.0191)**	0.2263 (0.0023)
10	2000-01-01	2010-01-01	**0.2649 (0.0188)**	0.1863 (0.0028)
	2001-01-01	2011-01-01	**0.2658 (0.0175)**	0.1943 (0.0026)
	2002-01-01	2012-01-01	**0.3206 (0.0181)**	0.2051 (0.0028)
	2003-01-01	2013-01-01	**0.3392 (0.0184)**	0.2223 (0.0023)
5	2000-01-01	2005-01-01	**0.2374 (0.0165)**	0.1724 (0.0026)
	2001-01-01	2006-01-01	**0.2353 (0.0150)**	0.1792 (0.0022)
	2002-01-01	2007-01-01	**0.2790 (0.0144)**	0.2017 (0.0024)
	2003-01-01	2008-01-01	**0.2855 (0.0147)**	0.2058 (0.0019)

diagonally dominated transition matrix leads to similar predictions across the 1000 runs. Although the clustering model uses the same methods in generating the transition matrices, the results are more accurate. The difference is that the clustering model partition firms with similar transition behaviours together and generates a representative transition matrix from this collection of firms. In other words, it can be said that the representative transition matrices of each cluster is custom-tailored to the distinct behaviour of each group of firms. Hence, the predictions are made on a test firm using a representative transition matrix that best characterizes it.

Judging from the values in Table 6, both the clustering and benchmark models demonstrate some ability in producing risk scores that function well as a predictor of the dependent variable (the default behaviour of the firms) as both models produce $d_{yx} > 0$. However, despite both models performing well, it can be observed that the clustering model outperform the benchmark model by producing "more effective" risk scores.

From Table 7, it can be observed that the clustering model outperforms the benchmark model in all performance measures. Treating the false positive rate FPR as a measure on the Type I error and the false negative rate as a measure of the Type II error, we find that the clustering model has an overall lower proportion of FP and FN.

Plotting the $F1_{\mu}$ and AA while varying the value of τ we can determine the effectiveness of each model for different prediction horizons. For practical purposes we used $\tau \in [1, 15]$, that is, prediction horizons ranging from one to fifteen years into the future. The transparent area in the plots of Fig. 3, represent the standard deviation of the respective results of each classification scenario. From Fig. 3 it is obvious that the clustering model outperforms the benchmark model for the majority of the tested τ values in terms of both AA and $F1_{\mu}$ for all classification scenarios. As we vary τ there are two observations that can be made. The first observation: For decreasing values of τ the degree by which the clustering model outperform the benchmark model also decreases. It is common for firms to maintain their current rating across shorter time periods, increasing the performance of the benchmark model as τ decreases. With longer time periods, firms are more likely to change ratings, and so the resulting benchmark transition matrix from Eq. (10) with large τ does a poor job in catching all the different behaviours of every firm. The second observation: For increasing values of τ, both the clustering and benchmark model performance decrease for the credit

Table 7
Results using the worst-case scenario thresholds for the default behaviour prediction scenario. Results under the column label (C) and (B) represent the results from clustering and the benchmark model respectively.

τ	Input date	(C) Re	(B) Re	(C) Pr	(B) Pr	(C) $F1_{\mu}$	(B) $F1_{\mu}$
15	2000-01-01	**0.9102 (0.0269)**	0.7408 (0.0460)	**0.6583 (0.0873)**	0.3445 (0.0230)	**0.7608 (0.0613)**	0.4681 (0.0170)
	2001-01-01	**0.9179 (0.0225)**	0.7486 (0.0421)	**0.6812 (0.0937)**	0.3526 (0.0261)	**0.7783 (0.0633)**	0.4774 (0.0188)
	2002-01-01	**0.9378 (0.0131)**	0.7325 (0.0519)	**0.7856 (0.0793)**	0.3941 (0.0272)	**0.8528 (0.0484)**	0.5087 (0.0150)
	2003-01-01	**0.9420 (0.0133)**	0.7628 (0.0354)	**0.7953 (0.0763)**	0.4283 (0.0354)	**0.8605 (0.0464)**	0.5465 (0.0222)
10	2000-01-01	**0.9275 (0.0283)**	0.7655 (0.0410)	**0.6839 (0.0836)**	0.3487 (0.0281)	**0.7841 (0.0594)**	0.4767 (0.0220)
	2001-01-01	**0.9412 (0.0212)**	0.7836 (0.0388)	**0.7289 (0.0843)**	0.3742 (0.0382)	**0.8186 (0.0564)**	0.5027 (0.0284)
	2002-01-01	**0.9425 (0.0129)**	0.7428 (0.0501)	**0.7777 (0.0634)**	0.3876 (0.0342)	**0.8508 (0.0398)**	0.5014 (0.0189)
	2003-01-01	**0.9415 (0.0129)**	0.7685 (0.0332)	**0.7853 (0.0690)**	0.4358 (0.0389)	**0.8546 (0.0428)**	0.5492 (0.0232)
5	2000-01-01	**0.9411 (0.0309)**	0.7833 (0.0340)	**0.6864 (0.0554)**	0.3351 (0.0383)	**0.7923 (0.0421)**	0.4658 (0.0324)
	2001-01-01	**0.9595 (0.0237)**	0.8111 (0.0303)	**0.7565 (0.0275)**	0.3731 (0.0555)	**0.8456 (0.0200)**	0.5058 (0.0450)
	2002-01-01	**0.9619 (0.0211)**	0.7838 (0.0393)	**0.8033 (0.0119)**	0.3958 (0.0345)	**0.8753 (0.0109)**	0.5132 (0.0251)
	2003-01-01	**0.9598 (0.0100)**	0.8088 (0.0220)	**0.7985 (0.0119)**	0.4380 (0.0313)	**0.8717 (0.0080)**	0.5583 (0.0246)

Table 8
The confusion matrices using input date 2001-01-01 under the credit rating prediction scenario.

(a) The clustering model using $\tau = 15$.

		Predicted class						
		AAA	AA	A	BBB	BB	B	C
Actual class	AAA	19.3	6.7	0.0	0.0	0.0	0.0	0.0
	AA	4.1	44.0	8.9	23.7	3.2	0.0	2.0
	A	0.2	8.9	55.8	35.5	14.9	0.8	3.8
	BBB	0.0	0.1	13.9	100.5	13.8	1.0	9.6
	BB	0.0	0.1	3.3	18.3	69.5	6.9	1.8
	B	0.0	0.0	1.0	3.5	4.3	14.3	9.0
	C	0.6	0.4	0.6	1.7	0.9	0.4	82.4

(b) The benchmark model using $\tau = 15$.

		Predicted class						
		AAA	AA	A	BBB	BB	B	C
Actual class	AAA	15.0	7.0	4.0	0.0	0.0	0.0	0.0
	AA	1.0	26.7	26.2	23.1	7.0	0.0	2.0
	A	0.0	5.0	45.8	47.2	18.0	0.0	4.0
	BBB	0.0	1.0	6.0	99.0	24.0	0.0	9.0
	BB	0.0	0.0	2.0	12.0	83.0	0.0	3.0
	B	0.0	0.0	0.0	7.0	11.0	0.0	14.0
	C	0.0	1.0	2.0	8.0	23.0	0.0	53.0

(c) The clustering model using $\tau = 5$.

		Predicted class						
		AAA	AA	A	BBB	BB	B	C
Actual class	AAA	16.5	4.3	0.2	0.0	0.0	0.0	0.0
	AA	1.3	35.3	8.3	1.7	0.0	0.0	0.3
	A	0.0	7.7	64.2	40.0	5.7	1.0	3.3
	BBB	0.0	1.0	7.7	124.0	32.4	0.4	7.5
	BB	0.0	0.9	1.4	10.4	96.8	1.4	7.1
	B	0.0	0.0	0.1	2.4	8.6	11.9	6.9
	C	0.0	0.0	0.0	2.4	0.8	3.3	72.4

(d) The benchmark model using $\tau = 5$.

		Predicted class						
		AAA	AA	A	BBB	BB	B	C
Actual class	AAA	16.0	4.0	1.0	0.0	0.0	0.0	0.0
	AA	0.0	31.0	14.0	2.0	0.0	0.0	0.0
	A	0.0	4.0	65.0	43.0	6.0	1.0	3.0
	BBB	0.0	1.0	4.0	132.0	31.0	3.0	2.0
	BB	0.0	1.0	2.0	11.0	103.0	0.0	1.0
	B	0.0	0.0	0.0	4.0	11.0	15.0	0.0
	C	0.0	0.0	0.0	4.0	15.0	10.0	50.0

Table 9
The confusion matrices using input date 2001-01-01 under the transition direction prediction scenario.

(a) The clustering model using $\tau = 15$.

		Predicted class		
		−1	0	1
Actual class	−1	65.1	21.9	4.0
	0	16.0	261.7	43.3
	1	11.4	76.8	89.9

(b) The benchmark model using $\tau = 15$.

		Predicted class		
		−1	0	1
Actual class	−1	39.6	38.0	13.4
	0	127.7	160.7	32.6
	1	61.8	100.8	15.4

(c) The clustering model using $\tau = 5$.

		Predicted class		
		−1	0	1
Actual class	−1	34.4	32.3	0.3
	0	17.6	380.5	13.9
	1	3.6	92.4	15.1

(d) The benchmark model using $\tau = 5$.

		Predicted class		
		−1	0	1
Actual class	−1	0.0	67.0	0.0
	0	0.0	412.0	0.0
	1	0.0	111.0	0.0

rating prediction and transition direction prediction scenarios. The rate at which the performance deteriorates, however, is higher in the benchmark model. The clustering model's performance decreases but then levels out eventually for $\tau \in [1, 15]$ in all classification scenarios. In the default prediction scenario the clustering model performance actually increases while the benchmark model performance decreases with increasing values of τ. It can be stated that the clustering model's performance is more consistent over $\tau \in [1, 15]$. The confusion matrices

Table 10
The confusion matrices using input date 2001-01-01 under the default behaviour prediction scenario.

(a) The clustering model using $\tau = 15$.

		Predicted class	
		1	0
Actual class	1	82.7	7.3
	0	40.3	459.7

(b) The benchmark model using $\tau = 15$.

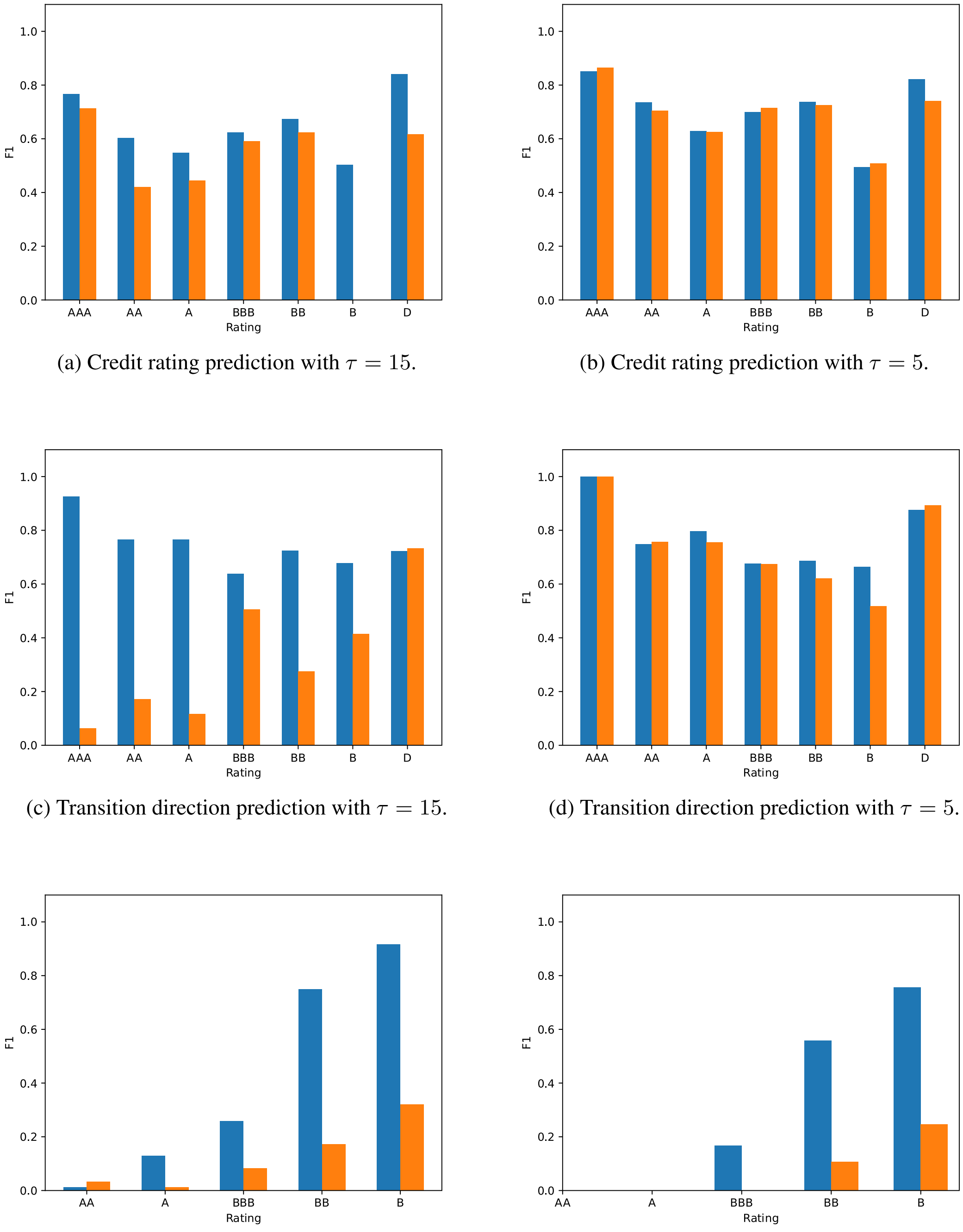
		Predicted class	
		1	0
Actual class	1	68.0	22.0
	0	128.6	371.4

(c) The clustering model using $\tau = 5$.

		Predicted class	
		1	0
Actual class	1	65.0	3.0
	0	22.6	499.4

(d) The benchmark model using $\tau = 5$.

		Predicted class	
		1	0
Actual class	1	54.4	13.6
	0	90.8	431.2



(a) Credit rating prediction with $\tau = 15$.

(b) Credit rating prediction with $\tau = 5$.

(c) Transition direction prediction with $\tau = 15$.

(d) Transition direction prediction with $\tau = 5$.

(e) Default behaviour prediction with $\tau = 15$.

(f) Default behaviour prediction with $\tau = 5$.

Fig. 4. The $F1$-measure for each credit rating under the three classification scenarios. The left (blue) and right (orange) bar graphs are the results using the clustering and benchmark models respectively.

of each classification scenario and a breakdown of the performance of the model across each credit rating in terms of the $F1$ is presented in Appendix A and Appendix B, repsectively.

6. Conclusions

The changes in the credit rating of a firm can have a substantial impact on bond pricing, valuation of credit derivatives, and management decisions of companies. In this paper, we adapted the sequence-based clustering technique used in web-usage mining to improve transition matrix estimation methods in the context of credit risk. The clustering algorithm used was the PCA-guided K-means algorithm and the analysis is in part possible due to the convenient cluster-ready representation of sequence matrices. Some properties of sequence matrices were presented which can prove to be beneficial for future development and implementation in models that intend to utilize this sequence matrix representation. Credit rating prediction, credit rating transition direction prediction, and default behaviour prediction were the three classification scenarios that were used to test the performance of the clustering model.

The clustering model was compared against the benchmark model where clustering was absent. The results suggest that by clustering the sequence matrices of firms, the overall predictive power of the representative transition matrices is greater than just using a single transition matrix. The performance of the models in the credit rating prediction and transition direction prediction classification scenarios were evaluated in terms of the average accuracy and micro-averaged $F1$-score. The performance of the models in the default behaviour prediction classification scenario were evaluated in terms of the recall, precision, and $F1_{\mu}$-score. The worst-case scenario threshold provides a suitable means of evaluating our model against the benchmark model.

The following are some potential extensions of our work. The clustering algorithm used the Euclidean distance measure to distinguish similarity between different sequence matrices of firms. One can develop and incorporate a distance measure that considers transition risks. In the context of credit risk, the Euclidean distance measure is not able to distinguish clusters by their "riskiness". Another extension could include using more sophisticated models making use of other features of financial companies after clustering based off their sequence matrices. One can also attempt to address the issue that in reality, a time-homogeneous Markov chain does not fully model credit rating sequences. This can be addressed by incorporating features into the model that take into account the non-Markovian effects of credit ratings such as rating drift.

CRediT authorship contribution statement

Richard Le: Investigation, Writing - original draft, Visualization. **Hyejin Ku:** Conceptualization, Methodology, Supervision. **Doobae Jun:** Data curation, Formal analysis.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This research was partly supported by Natural Sciences and Engineering Research Council of Canada, Discovery grant 504316. The work of Doobae Jun was supported by Basic Science Research Program through the NRF of Korea funded by the Ministry of Science (2018R1D1A1B07051450).

Appendices

The following section contains supplementary information to improve understanding of the results presented in this paper. In Appendix A, we present the confusion matrices for each classification scenario to highlight the sample size and balance of the classes. In Appendix B, we present the performance of the clustering model against the benchmark model across each credit rating using the $F1$-score for each classification scenario. The input date used for the results in the following section was set to 2001-01-01. The following results are generated by averaging over 1000 5-fold cross-validation runs.

Appendix A. Confusion matrices

See Tables 8–10.

Appendix B. Performance measures by credit rating

See Fig. 4.

References

Chen, N., Ribeiro, B., Vieira, A., & Chen, A. (2013). Clustering and visualization of bankruptcy trajectory using self-organizing map. *Expert Systems with Applications*, *40*(1), 385–393. http://dx.doi.org/10.1016/j.eswa.2012.07.047, http://www.sciencedirect.com/science/article/pii/S0957417412009025.

Christensen, J. H., Hansen, E., & Lando, D. (2004). Confidence sets for continuous-time rating transition probabilities. *Journal of Banking & Finance*, *28*(11), 2575–2602, Recent Research on Credit Ratings. http://dx.doi.org/10.1016/j.jbankfin.2004.06.003. http://www.sciencedirect.com/science/article/pii/S0378426604001025.

D'Amico, G., Dharmaraja, S., Manca, R., & Pasricha, P. (2019). A review of non-Markovian models for the dynamics of credit ratings. *Reports on Economics and Finance*, *5*(1), 15–33.

Dharmaraja, S., Pasricha, P., & Tardelli, P. (2017). Markov Chain model with catastrophe to determine mean time to default of credit risky assets. *Journal of Statistical Physics*, *169*(4), 876–888. http://dx.doi.org/10.1007/s10955-017-1890-z.

Forman, G., & Scholz, M. (2010). Apples-to-apples in cross-validation studies: pitfalls in classifier performance measurement. *ACM SIGKDD Explorations Newsletter*, *12*(1), 49–57.

García, V., Marqués, A. I., & Sánchez, J. S. (2015). An insight into the experimental design for credit risk and corporate bankruptcy prediction systems. *Journal of Intelligent Information Systems*, *44*(1), 159–189. http://dx.doi.org/10.1007/s10844-014-0333-4, https://doi.org/10.1007/s10844-014-0333-4.

Gunnvald, R. (2014). *Estimating probability of default using rating migrations in discrete and continuous time* (Master's thesis), KTH, Mathematical Statistics, Available at http://www.diva-portal.org/smash/record.jsf?pid=diva2%3A747996&dswid=7601.

Guo, X., Zhu, Z., & Shi, J. (2012). A corporate credit rating model using support vector domain combined with fuzzy clustering algorithm. *Mathematical Problems in Engineering*, *2012*.

Irmatova, E. (2016). RELARM: A rating model based on relative PCA attributes and k-means clustering. arXiv preprint arXiv:1608.06416.

Jafry, Y., & Schuermann, T. (2004). Measurement, estimation and comparison of credit migration matrices. *Journal of Banking & Finance*, *28*(11), 2603–2639. http://dx.doi.org/10.1016/j.jbankfin.2004.06.004, Recent Research on Credit Ratings http://www.sciencedirect.com/science/article/pii/S0378426604001037.

Jain, A. K. (2010). Data clustering: 50 years beyond k-means. *Pattern Recognition Letters*, *31*(8), 651–666, Award winning papers from the 19th International Conference on Pattern Recognition (ICPR). http://dx.doi.org/10.1016/j.patrec.2009.09.011. http://www.sciencedirect.com/science/article/pii/S0167865509002323.

Jarrow, R. A., Lando, D., & Turnbull, S. M. (1997). A Markov model for the term structure of credit risk spreads. *Review of Financial Studies*, *10*(2), 481–523, http://www.jstor.org/stable/2962353.

Kiefer, N. M., & Larson, C. (2004). *Testing simple Markov structures for credit rating transitions*. Comptroller of the Currency.

Kuncheva, L. I., & Sánchez, J. (2008). Nearest neighbour classifiers for streaming data with delayed labelling. In *2008 Eighth IEEE International conference on data mining* (pp. 869–874). IEEE.

Lando, D., & Skodeberg, T. M. (2002). Analyzing rating transitions and rating drift with continuous observations. *Journal of Banking & Finance*, *26*(2–3), 423–444, Retrieved from https://EconPapers.repec.org/RePEc:eee:jbfina:v:26:y:2002:i:2-3:p:423-444.

Liu, Y. (2002). The evaluation of classification models for credit scoring. Available at https://pdfs.semanticscholar.org/a62c/581a334e155cd6867a54d329b3db81ef2034.pdf.

Montiel, J., Bifet, A., & Abdessalem, T. (2017). Predicting over-indebtedness on batch and streaming data. In *2017 IEEE international conference on big data (big data)* (pp. 1504–1513). IEEE.

Morales, M. H., Rodríguez, J. T., & Montero, J. (2015). *Credit rating using fuzzy algorithms*.

Newson, R. (2006). Confidence intervals for rank statistics: Somers' D and extensions. *The Stata Journal*, *6*(3), 309–334. http://dx.doi.org/10.1177/1536867X0600600302.

Park, S., Suresh, N. C., & Jeong, B.-K. (2008). Sequence-based clustering for web usage mining: A new experimental framework and ANN-enhanced K-means algorithm. *Data & Knowledge Engineering*, *65*(3), 512–543. http://dx.doi.org/10.1016/j.datak.2008.01.002, http://www.sciencedirect.com/science/article/pii/S0169023X08000104.

Plasse, J., & Adams, N. (2016). Handling delayed labels in temporally evolving data streams. In *2016 IEEE international conference on big data (big data)* (pp. 2416–2424).

Sanchez, I. E. (2016). Optimal threshold estimation for binary classifiers using game theory. *F1000Research*, *5*.

Sharma, A., Jadi, D. M., & Ward, D. (2018). Evaluating financial performance of insurance companies using rating transition matrices. *The Journal of Economic Asymmetries*, *18*, Article e00102.

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, *45*(4), 427–437. http://dx.doi.org/10.1016/j.ipm.2009.03.002, http://www.sciencedirect.com/science/article/pii/S0306457309000259.

Somers, R. H. (1962). A new asymmetric measure of association for ordinal variables. *American Sociological Review*, *27*(6), 799–811, http://www.jstor.org/stable/2090408.

Taylor, H. M., & Karlin, S. (1998). *An introduction to stochastic modeling* (3 ed.). Academic Press.

Thomas, L. C., Allen, D. E., & Morkel-Kingsbury, N. (2002). A hidden Markov chain model for the term structure of bond credit risk spreads. *International Review of Financial Analysis*, *11*(3), 311–329, Credit Derivatives. http://dx.doi.org/10.1016/S1057-5219(02)00078-9. http://www.sciencedirect.com/science/article/pii/S1057521902000789.

Trueck, S., & Rachev, S. (2009). *Academic press advanced finance*, *Rating based modeling of credit risk: Theory and application of migration matrices*. Elsevier Science, https://books.google.ca/books?id=C8mxdgm_K8EC.

Xu, Q., Ding, C., Liu, J., & Luo, B. (2015). PCA-Guided search for K-means. *Pattern Recognition Letters*, *54*, 50–55. http://dx.doi.org/10.1016/j.patrec.2014.11.017, http://www.sciencedirect.com/science/article/pii/S0167865514003675.