http://www.jsebs.org
ISSN: 2288-3908

The Journal of Society for e-Business Studies
Vol.24, No.1, February 2019, pp.91-103
https://doi.org/10.7838/jsebs.2019.24.1.091

가변 마코프 모델을 활용한 매출 채권 연령 분석

Analysis of Accounts Receivable Aging Using Variable Order Markov Model

강윤철(Yuncheol Kang)*, 강민지(Minji Kang)**, 정광헌(Kwanghun Chung)***

초 록

기업 입장에서 앞으로 있을 현금흐름에 대한 예측이란, 향후 발생할 수 있는 유동성(현금 부족) 위험을 미리 파악할 수 있다는 점과 미래의 투자계획을 세우는데 중요한 자료가 될 수 있다는 점에서 중요한 의의를 지닌다. 그러나 기업 간 거래에서 매출 채권 형태로 발생하는 거래 유형은 다른 유형의 거래와는 달리 채무 이행 불확실성이 존재하며, 이로 인해 정확한 현금흐름 예측을 어렵게 한다. 본 연구에서는 추계적 분석 기법의 하나인 가변 마코프 기법(Variable Order Markov model)을 활용하여 기업 간에 발생 할 수 있는 매출 채권과 관련한 현금흐름 동향을 예측한다. 구체적으로는, PST(Probabilistic Suffix Tree)라는 가변 마코프 기법을 활용하여, 지난 과거의 매출 채권 발행 및 수금 내역을 바탕으로 해당 매출 채권들의 기대 연령 예측 연구를 수행하였다. 본 연구결과를 통해, 기존의 다른 기법들과 대비하여 가변 마코프 기법을 활용 시, 평균 12.5% 이상의 정확도를 보여주고 있음을 밝혔다.

ABSTRACT

An accurate prediction on near-future cash flows plays an important role for a company to attenuate the shortage risk of cash flow by preparing a plan for future investment in advance. Unfortunately, there exists a high level of uncertainty in the types of transactions that occur in the form of receivables in inter-company transactions, unlike other types of transactions, thereby making the prediction of cash flows difficult. In this study, we analyze the trend of cash flow related to account receivables that may arise between firms, by using a stochastic approach. In particular, we utilize Variable Order Markov (VOM) model to predict how future cash flows will change based on cash flow history. As a result of this study, we show that the average accuracy of the VOM model increases about 12.5% or more compared with that of other existing techniques.

키워드 : 현금흐름예측, 매출채권 연령관리, 가변마코프모델
Cash Flow forecasting, Account Receivable Aging, Variable Order Markov Model

---

This work was supported by 2017 Hongik University Research Fund.

* First Author, Department of Industrial Engineering, College of Engineering, Hongik University (yckang@hongik.ac.kr)
** Co-Author, Department of Industrial Engineering, College of Engineering, Hongik University (jaellen@naver.com)
*** Corresponding Author, College of Business Administration, Hongik University(khchung@hongik.ac.kr)

Received: 2019-01-11, Review completed: 2019-01-24, Accepted: 2019-01-29

1. 서 론

기업에 있어서의 안정적인 현금흐름은 기업을 유지하고 운영하는데 필수적이며, 정확한 현금흐름 예측은 거의 모든 유형의 기업에서 중요하게 다뤄지고 있다. 중소기업의 경우 신용 위험과 가용자원 제약 등으로 인해 높은 수준의 재정관리 비용이 발생하고 있으며, 중견 혹은 대기업의 경우에도 다양한 사업 전개 및 확장 전략에 따라 효율적인 자본 배분을 위해 앞으로의 현금흐름 예측이 필요한 경우가 많다. 특히, 기업 간에 일어나는 현금흐름은 내재되어 있는 특유의 불확실성으로 인해 사전에 제대로 파악되지 못할 수 있으며, 현금흐름 예측의 실패로 인해 예기치 못한 현금 유동성 위기를 겪을 위험성이 항상 존재하고 있다.

이러한 관점에서 매출 채권 연령 예측 분야(Forecasting accounts receivable aging)는 고객 회사의 채무 이행 지연 혹은 불이행과 같은, 이른 바 “악성 채무” 가능성을 사전에 파악할 수 있다는 점에서 주목을 받아 왔다. 매출 채권이란 기업이 상품을 판매하는 과정에서 발행한 채권으로, 아직 회수가 되지 않은 매출 채권의 발행 이후 기간(즉, 채권 연령) 정보를 포함하고 있다. 만약 사전에 해당 채권의 향후 회수 가능성을 가늠할 수 있다면, 해당 정보를 활용하여 회사의 현금흐름 유동성 위험에 선제적으로 대응할 수 있을 것이다.

한편, 현금의 흐름은 거시경제적 요인, 고객의 지불 행위, 특정 공급 사슬망의 고유 특성 등 다양하고 복합적이며 불확실성을 갖는 요인들에 의해 영향을 받을 수 있으며, 이러한 이유로 현금흐름은 일반적으로 추계적 과정(Stochastic process)으로 간주된다. 현금흐름 분석과 밀접한 연관을 갖고 있는 매출채권 연령 예측 분야에서는 이와 같은 추계적 과정 이론들을 적극적으로 분석에 활용하고 있다. 매출 채권 연령을 예측하기 위해 사용되어 온 대표적인 알고리즘으로, Corcoran[4]이 제안한 모델과 Pate-Cornell et al.[5]이 제안한 모델 두 가지를 꼽을 수 있다. Corcoran 모델은 채권 연령(Account Receivable 연령이하 AR 연령)에서 현금흐름을 예측하기 위해 지수 평활 접근법의 개념을 사용한다. 반면 Pate-Cornell et al.[5]은 추계적 과정(Stochastic process)을 활용하여 단기 현금 유동 예측을 수행한다. 최근에는 지수 평활 접근법(즉, Corcoran 모델)과 추계적 과정(즉, Pate-Cornell 모델)을 통합하여 현금흐름을 예측하는 시도(SFA 모델[7])가 있었다. SFA(Stochastic financial analytics)에서는 마코프 모델(Markov model)을 이용하여 모든 고객들의 지불 행위를 표현하고, 여기에 베이즈 기법(Bayes’ Rule)을 적용하여 개별적인 송장 수준에서 고객의 각 지불 행위 패턴을 파악하는 방식으로 AR 연령을 예측하였다. 특히, 이 모델은 기존의 두 모델이 갖고 있는 장점만을 취합하여 통합한 모델로 기존의 모델과 비교하여 상대적으로 높은 수준의 정확도를 갖고 있는 것으로 보인다.

현금흐름을 예측한다는 것은 시계열 상에서 나타날 수 있는 앞으로의 현금유입값을 예측하는 것과 같다고 볼 수 있다. 미래의 현금 유입값들이 과거에 연속적으로 있어왔던 현금유입값들과 관계가 있다고 가정한다면, 기존의 마코프 모델은 예측을 하는 시점의 바로 직전 시점의 상태(즉, first-order 마코프 모델)에 의해서만 상태값, 즉 앞으로의 현금유입값을 예측하는 경우라 볼 수 있다. 실제로 마코프 성질(Markov Property)라 불리우는 이 특성은 예측

바로 이전의 값이 결국 기존의 history 정보를 함축적으로 대표할 수 있는 값이라는 가정을 전제로 두고 있다[2].

본 연구에서는 위와 같은 기존의 마코프 모델 가정에서, 바로 이전의 상태 값만이 아닌 더 이전 차수의 정보까지 활용하는 고 차수(high-order) 마코프 모델 개념을 활용하게 되면 기존의 방법론들보다 더 높은 예측 정확도를 가질 것이라는 가정에서 시작하고자 한다. 고 차수 마코프 모델을 사용한다는 의미는 이전에 발생했던 정보뿐만이 아니라 각 과거 정보 간에 내재 되어 있을 수 있는 특정 시퀀스 패턴을 추가로 활용한다는 의미를 내포하고 있다. 다만, 상태의 개수 및 차수가 증가함에 따라 소요되는 계산복잡도(Computational Complexity)가 지수적으로 증가하기 때문에, 순수한 고 차수 마코프 모델이 실질적으로 활용되기 어려운 측면이 존재한다. 이러한 문제를 해결하기 위해 가변 차수 마코프 모델(Variable Order Markov model: 이하 VOM)이 종종 활용되고 있다. VOM은 고 차수 마코프 모델의 일종이나 다른 순수한 의미에서의 고 차수 마코프 모델과는 달리, 다음 상태를 예측하는데 필요한 과거 상태의 개수가 일정하지 않으며 상태의 특성 및 기존정보의 흐름 패턴에 따라 각기 다른 차수를 가질 수 있다. 만약 예전 정보를 구체적으로 활용할 필요가 있을 경우에만 차수를 증가시키고 그렇지 않은 경우에는 차수를 감소시킬 수 있다면, 전체 소요되는 계산 복잡도를 줄일 수 있고 실질적인 활용성을 높이는데 장점을 가져다 줄 것이다.

VOM은 일반적으로 데이터 기반 학습을 통해 구성되며, 지금까지의 문헌을 살펴보면 여러 가지 VOM 모델들이 제안되어 왔다. 최초의 VOM 모델인 Ziv and Lempel[9], 탈출과 배제 알고리즘을 사용하는 Prediction by Partial Match[3], 이진화 변수를 이용하는 Context Tree Weighting Method[8]과 각 접미사(suffix)들의 확률을 계산하는 Probabilistic Suffix Tree(이하 PST)[6]가 있다. 이들 중 PST는 가변 차수 마코프 모델을 풀기 위해 가장 널리 알려진 알고리즘 중 하나로 문서에서 생략된 문장들을 예측하거나, 생물학 분야의 DNA 서열 분석 등 주로 주어진 서열정보를 통해 앞으로 나타날 서열을 예측하는데 주로 사용되어 왔으며, 기존의 은닉 마코프 모델(Hidden Markov model)보다 일반적으로 우수한 성능을 보여준다고 알려져 있다[1]. 특히, 적은 데이터만으로도 PST 모델 생성이 가능한 점이 장점이며, 이는 많은 양의 데이터를 필요로 하는 최근의 딥러닝 모델(예를 들면 RNN)과 대비되는 특징이라 할 수 있다. 본 논문에서는 VOM 알고리즘 중 하나인 PST 알고리즘이 기존에 주로 적용되어 왔던 생물학 분야를 넘어 현금흐름을 예측하는, 이른바 기업 어플리케이션에서는 어떻게 적용될 수 있는지 분석하고 기존 모델을 수정하여 해당 어플리케이션에 적용될 수 있는 방안에 대해 논하고자 한다. 특히, 본 연구에서 제시하는 결과가 기존의 모델들과 비교했을 때, 예측 정확도 측면에서 어떠한 장점이 있는지 살펴보고자 한다.

본 논문의 구성은 다음과 같다. 제2장에서는 기존의 PST 알고리즘에 대해 설명하고, 제3장에서는 본 연구에서 고려하는 문제 및 모델을 정의하며, 기존의 PST 알고리즘이 본 연구의 문제에 적용되기 위해 어떻게 바뀌어야 되는지 설명한다. 제4장에서는 실제 데이터를 활용하여, 제시된 모델이 기존의 알고리즘과 정확도 측면

에서 어떠한 특징이 있는지 실험을 통해 밝히고 결과에 대해 논의한다. 결론과 추후 연구에 대한 부분은 마지막 5장에 기술하였다.

2. 가변 마코프 모델

본 연구에서는 가변 마코프 모델의 하나인 PST를 활용하여 현금흐름 예측을 하고자 한다. 이를 위해 제2장에서는 주어진 시퀀스 데이터를 활용하여 어떻게 PST 모델이 생성되는지 설명한다. PST 모델 생성 절차는 다음과 같다. 우선 가능한 상태 요소들로 이루어진 집합(Set)을 Ω라고 하자. 주어진 시퀀스 데이터를 q^m이라 정의하고, 해당 데이터는 m 개의 상태 요소값(중복허용)들로 정의되었다고 가정한다. 이때 $q^m = \langle \omega_1, \omega_2, ..., \omega_m \rangle$, $\omega_i \in \Omega$와 같이 상태 요소들의 시퀀스 다중집합으로 정의할 수 있다. 다음으로 어떤 시퀀스 s^n이 존재하여, $s^n = \langle \omega_j, \omega_{j+1}, ..., \omega_{j+(n-1)} \rangle$, $\omega_i \in \Omega$를 만족한다고 하였을 때, s^n은 q^m의 서브시퀀스(subsequence)라 볼 수 있다.

q^m의 서브시퀀스로 가능한 모든 경우의 시퀀스로 이루어진 집합(Set)을 Q, s^n이 q^m의 서브시퀀스로 나타난 횟수를 $|s^n|$이라고 한다면, 샘플 시퀀스에서 특정 시퀀스 s^n을 관찰할 확률은 다음과 같이 추정할 수 있다.

$$p(\widehat{s^n}) \cong \frac{|s^n|}{|Q|}$$

여기서 $p(\widehat{s^n})$의 합은 1이며 이를 확률 분포로 활용한다고 했을 때, 주어진 시퀀스 s^n 이후에 나타나는 문자 ω의 조건부 확률을 정의한다. 이 확률은 시퀀스 s^n직후에 문자가 나타나는 횟수로 계산한다. 시퀀스 s^n 이후에 오는 경우의 수는 $\sum_{\omega \in \Omega} x_{s^n \omega}$로 표현되며, ω가 문자열 s 이후에 나올 조건부 확률은 다음과 같이 정의한다.

$$p(\widehat{\omega|s}) = \frac{x_{s^n \omega}}{\sum_{\omega \in \Omega} x_{s^n \omega}}$$

위의 계산방법을 이용하여 연속적인 문자열 이후에 나타날 수 있는 문자의 확률을 계산할 수 있다. 보다 자세한 수리적 접근 방법은 Bejerano and Yona[6]의 PST 알고리즘 설명파트를 참고할 수 있다.

3. 현금흐름 예측을 위한 PST 모델

이 장에서는 앞서 살펴본 PST 알고리즘을 현금흐름을 예측하기 위한 PST 모델로 활용하기 위한 방안에 대해 설명한다.

3.1 상태 정의

고객의 지불 행위는 현재 시점에서 채권의 연령 기간에 따라 <Table 1>과 같이 분류한다. 일반적으로, 기업 회계 관점에서는 30일(약 1개월) 기준으로 채권의 연령을 분류하므로, 30일 단위로 기준 상태를 구분하는 게 자연스러운 가정이라 볼 수 있다. 만약 현재 시점에서 30일이 아직 지나지 않은 채권이 지불 완료 될 경우 s_0 상태로 간주한다. 그리고 만기 후 120일이 지난 채권(s_5)은 모두 악성채무(Bad Debt)로 분류한다.

〈Table 1〉 Definition of a Single Element of a State

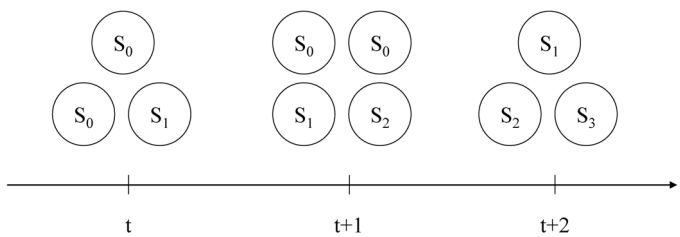
Element	Description
s_0	Paid before due-date
s_1	Paid 1 day to 30 days after due-date
s_2	Paid 31 days to 60 days after due-date
s_3	Paid 61 days to 90 days after due-date
s_4	Paid 91 days to 120 days after due-date
s_5	Paid/Non-paid 121 days after due-date

<Table 1>을 이용하여 본 연구에서 사용될 PST의 상태를 정의한다. 본 연구의 목적은 현재 채권의 연령을 근거로 다음 시점에서 채권의 상태를 예측하여 이를 토대로 현금흐름을 예측하고자 하는 것이다. 문제는 한 기업 입장에서 현재 유동중인 채권들은 복수개가 될 수 있으며, 심지어 한 기업으로부터도 동시에 여러 개의 채권을 발행 중에 있을 수 있다. 즉, 현재 시점에서 분석해야 하는 채권 연령값들은 다수로 이루어진 집합으로 구성될 수 있으며 이를 하나의 마코프 모델 상태값으로 표현하기 위해서는 마코프 모델 상태 값이 <Table 1>에서 정의한 단일상태값들이 복합적으로 표현되어야 한다. 이 경우 복수개의 단일 상태값들은 중복을 허용하는 다중집합(Multiset) 개념을 활용하여 표현할 수 있다. 예를 들어, 현재 시점에서 채권 연령 현황 상태가 만기 전 지불 완료인 경우 2건, 만기 후 1일~30일 이내 1건, 그리고 만기 후 30~60일 이내 1건일 경우가 발생할 수 있으며, 이 경우 다중집합 개념을 활용한다면 $\{s_0, s_0, s_1, s_2\}$로 표현된다.

<Figure 1>에 나타나는 상태 순서의 경우, 다중집합 개념을 사용하여 t시점에 $\{s_0, s_0, s_1\}$, t+1시점에는 $\{s_0, s_0, s_1, s_2\}$, 다음 t+2시점에는 $\{s_1, s_2, s_3\}$로 상태값들을 표현할 수 있다.

3.2 상태 간 전이 모델

본 연구에서 각 고객의 지불 패턴은 이전의 지불패턴과 완전히 독립적이지 않고 어느 정도 서로 연관성을 가지고 있다고 가정한다. 이 가정하에서는 미래의 현금 유입 패턴을 분석하기 위해서 과거에 나타났던 매출 채권 연령 패턴 분석이 필요할 것이다. 일반적인 마코프 특성에 의하면 현재 상태는 직전의 상태에만 영향을 받는다. 그러나, 본 연구에서 고려하는 문제는 현재 상태에 영향을 주는 과거 패턴(예시-이전 2단계, 3단계 전 등)들을 직전의 상태 하나만 고려하는 게 아닌, 여러 개의 과거 기록과 그들 간의 패턴정보를 활용하기를 원한다. 현금흐름 예측 관점에서 볼 때, 이는 다음과 같이 설명될 수 있다. 만약 고객이 그리 오래되지 않은 과거에 한 번이라도 악성채무였던 적이 있었다면



〈Figure 1〉 Example of a State Sequence

(즉, 지불 시점이 아주 늦은 적이 있거나 지불 불이행이 있었다면), 비록 바로 직전에 채권 지불이 일찍 이루어졌다 하더라도, 그 고객은 상대적으로 “성실한” 다른 고객과 비교했을 때 “악성 채무” 상황에 다시 직면할 가능성이 더 높다고 볼 수 있다는 것이 본 연구의 핵심 가정이다.

문제는 우리가 정의한 상태가 multiset 속성을 지니고 있으며, 해당 상태에 속한 각각의 단일 요소 상태가 전 시점의 어떤 개별 요소 상태와 연결되어 있었는지 추적할 방법이 없다는 것이다. 이러한 문제로 인해, 다중집합의 상태 값 간의 전이(transition) 형태를 파악하는 게 어려워진다. 따라서, 각 다중집합 요소들 사이에 있을 수 있는 모든 경우의 수(즉, cartesian product)를 고려하고 각 경우의 수는 동일한 가중치로 발생할 수 있다고 가정하는 게 자연스럽다. 이는 불확실성이 있는 경우에 각각의 대안을 동일한 가중치를 주어 평가하는 라플라스 기법과 원리가 유사하다. 하지만 cartesian product 연산 특성으로 인해 고려하고자 하는 기간(즉, PST의 최대 확장 가능 차수)이 길어질수록 전체 집합 크기가 기하급수적으로 증가한다. 이를 방지하기 위해 본 연구에서는 PST의 최대 차수를 합리적인 수준에서 제한하는 방식을 사용할 수 있다.

3.3 현금흐름 예측을 위한 PST 모델

데이터가 주어졌을 때 PST 모델은 해당 데이터로부터 학습을 통해 구축된다. 학습 절차는 다음과 같다. 우선 학습에 필요한 파라미터를 세팅한다. 주요 파라미터는 <Table 2>와 같이 정리할 수 있다. 첫째, P_{min}은 데이터에서 해당 문자열을 PST에 포함시키기 위한 최소 확률을 일컫는다. 다음으로 h는 해당 문자 하나를 PST 노드에 추가시키기 위해 사용되는 threshold 값이다. L은 PST 노드 중 가장 긴 문자열의 길이를 뜻한다. 마지막으로 d는 노드로 성립될 수 있는 최소 확률 값을 나타낸다.

학습 알고리즘은 다음과 같이 주어진다. 먼저 PST 초기화 및 위에서 주어진 주요 파라미터들을 적절히 세팅한다. 다음으로 multiset M을 정의하고 여기에 데이터에서 나타나는 상태들의 cartesian product 연산 결과를 저장한다. 이때, M에 포함된 중복 값을 모두 소거하고 유일한 값으로만 이루어진 집합을 S로 정의한다. 이후 S에 있는 모든 요소가 제거될 때까지 다음 알고리즘을 반복한다.

단계 1: 집합 S에서 요소 s(즉, 샘플 문자열)를 추출

단계 2: k는 문자열 s에서 가장 첫 번째 문자를 제외한 문자열로 정의한다.

〈Table 2〉 Descriptions on Main Parameters for PST

Main parameters	Description
P_{min}	Minimum probability to include the corresponding string sequence into the PST
h	Threshold value to include a single character added to the current node of the PST
L	The length of the node whose name is the longest string compared to other nodes
d	Minimum probability to be defined as a node in PST

단계 3: 만약 다음 세 가지 조건을 모두 만족하면, 현재 PST에 해당 s를 s가 속할 수 있는 가장 긴 문자열 노드에 연결되는 새로운 노드로 추가한다.

조건 1) $P(\omega|s) \geq d$,
조건 2) $P(\omega|s)/P(\omega|k)$ 값이 h보다 크거나 1/h보다 작은 경우,
조건 3) 후보노드 ωs 의 길이가 L보다 작은 경우.

추가로 위 알고리즘에서 원활한 학습이 일어날 수 있도록 매 반복시마다 이뤄지는 확률 계산에서 평활지수(smoothing factor)를 반영하여 노드가 갑자기 늘어나는 현상을 방지할 수도 있다.

3.4 PST 예측 모델 설명

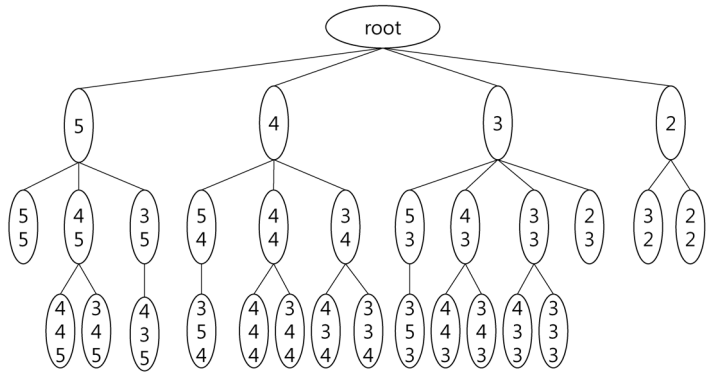
특정 고객과의 거래내역을 이용하여 만든 PST를 <Figure 2>에 도식화 하였고, 각 노드의 확률 값들을 <Table 3>에 정리하였다. PST 확률 값을 활용하여 다음 상태값을 예측하는 순서는 다음과 같이 정리될 수 있다.

1) 현재까지의 상태값 시퀀스 데이터를 추출한다. 예를 들면, 현재 시점으로 5개월 이전 시점부터 분석을 한다고 가정하고, 그 때부터 기록된 상태값들로 이루어진 시퀀스가 2-2-4-3-5라고 가정한다(즉, 5개월 및 4개월 전 상태 2, 3개월 전 상태 4, 2개월 전 상태 3, 1개월 전 상태 5).
2) 변수를 선언하고, 해당 시퀀스의 가장 접미부분(suffix)에 해당되는 노드 값을 해당 변수에 추가한다. 즉, 시퀀스 2-2-4-3-5에서는 5가 가장 후미에 위치한 노드에 해당된다.
3) 다음으로 이제까지 찾은 접미부분의 노드(즉 현재 변수에 저장되어 있는 시퀀스이름)가 PST에 실제로 존재하는지 찾아본 후, 만약 존재하는 경우에는 해당 접미부분 노드를 기준으로 한 차수 이전의 상태값을 현재 변수의 시퀀스 값 왼편에 추가한다. 위 예제에서는 현재 변수에 저장된 시퀀스 값 5가 PST에 존재하므로, 한 차수 이전의 상태 값인 3을 현재 변수 값인 5 왼편에 추가한다. 즉, 결과적으로 3-5 라는 시퀀스가 변수에 저장이 된다.
4) 다음 단계 2로 다시 돌아가서 3-5에 해당되는 노드가 PST에 있는지 확인한다. 만약 있다면, 단계 3을 다시 한 번 수행하여 변수에 저장되어 있는 시퀀스를 늘려 나간다. 이 작업을 변수에 있는 시퀀스가 PST 내 노드 값에 존재하지 않는 경우가 될 때까지 반복한다. 만약 찾지 못하면 다음 단계로 진행한다.
5) 현재 변수에 있는 시퀀스 명에 해당하는 확률 값을 찾아낸다. 이 예시에서 사용된 '2-2-4-3-5' 시퀀스의 확률은 '4-3-5' 노드의 확률을 따르게 됨을 알 수 있으며, <Table 4>의 결과를 참고하여, 다음 상태가 2일 확률은 0, 3일 확률은 0.325651, 4일 확률은 0.662825, 그리고 5일 확률은 0.011729가 된다.

참고로, <Figure 2>의 PST의 경우 L값이 3일 때의 PST 결과와 4 이상일 때의 PST 결과가 서로 동일하게 도출이 되었다. 이는 현재의 현금흐름 예측을 하는데 필요한 VOM 차수는 3차로 충분하다는 의미로 해석될 수 있다.

〈Table 3〉 Probabilities of PST Shown in 〈Figure 3〉

Node	P(2|node)	P(3|node)	P(4|node)	P(5|node)
Root	0.009656	0.539856	0.433179	0.017617
2	0.147146	0.852956	0	0
22	0	1	0	0
32	0	1	0	0
3	0.009226	0.551141	0.425242	0.0147
23	0	1	0	0
33	0	0.548369	0.435955	0.01588
333	0	0.552284	0.430762	0.017159
433	0	0.451383	0.525123	0.023699
43	0	0.488315	0.495357	0.016533
343	0	0.509237	0.472495	0.018473
443	0	0.472357	0.507339	0.020509
53	0	0.314677	0.685425	0
353	0	0.382417	0.617686	0
4	0	0.456796	0.521176	0.022233
34	0	0.461419	0.520504	0.018282
334	0	0.479767	0.501245	0.019193
434	0	0.425059	0.552824	0.022322
44	0	0.437836	0.540047	0.022322
344	0	0.451908	0.523505	0.024791
444	0	0.438345	0.536572	0.025289
54	0	0.292207	0.707895	0
354	0	0.292207	0.707895	0
5	0	0.386566	0.607402	0.006237
35	0	0.366972	0.625937	0.007296
435	0	0.325651	0.662825	0.011729
445	0	0.363701	0.627311	0.009192
55	0	0	1	0



〈Figure 2〉 Example of PST

4. 실험 결과

본 장에서는 앞서 제시한 수정된 PST 알고리즘을 활용하여 현금 유동성 예측을 한 결과에 대해 설명하고, 기존의 다른 기법들과 비교하여 성능이 어떻게 되는지 검증하기로 한다.

4.1 PST 실험 절차

본 연구에서 제안한 PST 알고리즘을 테스트하기 위해 2010년에서 2012년까지 미국 중서부(mid-west) 지역에 위치한 어느 한 중소 제조업체의 실제 매출채권 데이터를 사용하였다.

Date	Due Date	Paid At	Amount	Days until payment	Payment Sequence State
...	...	...	...	...	...
2011-03-31	2011-04-30	2011-05-31	13,003.00	61	3
2011-03-31	2011-03-31	2011-05-31	100.00	61	3
2011-04-13	2011-05-13	2011-05-31	1,929.00	48	2
...	...	...	...	...	...

〈Figure 3〉 Example of Sample Account Receivable Data

해당 제조업체의 경우 다양한 종류의 제품을 생산하여 여러 업체에 납품하고 있으며, 납품 시 대금지불과 관련한 정보들(예시-납품 후 몇 일 이내로 대금이 지불되어야 하는지에 대한 정보 등)에 대한 데이터들을 <Figure 3>과 같이 엑셀형태로 관리하고 있다. PST 실험 절차는 다음과 같다. 첫 번째로, 연령 정보(단위: 일)를 이용하여 각 시퀀스 상태를 결정한다(<Figure 2>에서 가장 우측 열 값). 다음으로, 데이터 집합을 두 개의 이벤트 기준으로 분리한다. 하나는 특정 시점 기준으로 해당 시점까지 지불 완료된 채권에 관한 것이고, 다른 하나는 그 시점까지 지불 완료되지 않은 채권에 관한 것이다. 이러한 이벤트를 각각 “지불(paid)”과 “미지불(non-paid)”이라 명명한다. 이 기준에 따라 구분된 데이터 집합을 학습 및 검증용 데이터로 나누고, 학습용 데이터를 이용하여 지불과 미지불에 대한 PST 모델을 각각 생성한다. 마지막으로, 생성된 PST 모델들과 검증용 데이터를 사용하여 미래의 예상 유입 금액을 예측한다. 여기서 특정 시점 t에서의 예상 유입 금액 계산은 Bayes' Rule을 이용하여 계산하며 수식은 다음과 같다.

$$P(paid|\,T=t) = \frac{P(T=t|paid)\ \bullet\ P(paid)}{P(T=t|paid)\ \bullet\ P(paid) + P(T=t|nonpaid)\ \bullet\ P(nonpaid)}$$

4.2 결과 및 분석

해당 제조업체와 거래중인 총 8개 고객사들과의 과거 거래 데이터를 이용하여 PST 모델을 만들고 이를 Corcoran 모델(method 1), SFA 모델(method 2), 및 Pate-Cornell 모델(method 3)과 비교 분석을 수행하였다. PST 모델 학습 시 사용된 파라미터 값으로, $P_{min} = 0.001$, $h = 1.05$, $L = 3$, $d = 0.0001$이 사용되었다. 예측 정확도는 검증용 데이터 값과 각 알고리즘이 도출한 예측값의 차이, 즉 Percentage Error로 정의하였다(<Table 4> 참조). 전체적인 결과를 봤을 때 Mean Absolute Percentage Error (MAPE) 값은 PST가 가장 작았다. 특히 PST 알고리즘이 기존에 제시되었던 다른 기법들보다 평균적으로 12.51% 이상의 높은 정확도를 보여주고 있다. 가장 오차가 작게 나타난 기법도 PST이며, 오차율 2% 이내의 정확도를 보여주고 있다. 다만 모든 경우에 있어서 PST가 다른 기법에 비해 항상 낮은 오차율을 보여주는 것은

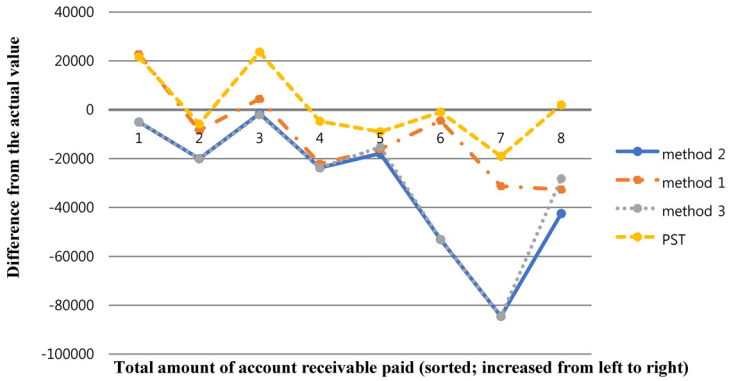
아니다. 예를 들면, Site 4번의 경우 기법 2 혹은 3에 비해 오차폭이 상당히 큼을 알 수 있으며, Site 5번의 경우에도 4번만큼은 아니나 큰 오차를 보여주고 있다.

〈Table 4〉 Comparison Results with other Algorithms

Site #	Method 1	Method 2	Method 3	PST
1	5.66%	67.35%	67.35%	1.24%
2	27.74%	36.12%	23.98%	-1.72%
3	32.65%	88.34%	88.34%	19.90%
4	-135.91%	29.95%	29.95%	-129.26%
5	-11.32%	3.99%	5.28%	-60.32%
6	34.80%	82.01%	82.01%	23.94%
7	45.80%	49.23%	49.23%	9.80%
8	30.28%	33.62%	28.54%	16.92%
MAPE	40.52%	48.83%	48.83%	32.89%

<Figure 4>에서는 기준 축(그래프에서 y = 0인 축)과 각 알고리즘이 예측한 결과값이 얼마나 차이가 발생하는지 해당 차이 값을 그래프로 표현하였다. 여기서 가로 축은 현금유입량 크기이다(즉, 실제 현금유입량이 작은 사이트에서 큰 사이트로 정렬하여 표기하였다). 세로 축은 예측값에서 실제값을 뺀 값을 표현하였다. <Figure 4>로부터 현금유입량 크기와 각 알고리즘의 예측 정확도 사이에 특정한 패턴이 있는지 개괄적으로 살펴볼 수 있다. <Figure 4>에서 보는 바와 같이, PST를 제외한 다른 알고리즘들은 실제 현금유입량 값이 커지면 커질수록 예측 오차 절대값이 커지며, 대부분 기준 축(y = 0) 하위에 위치하는, 즉under-estimation(과소예측)을 하는 경향이 나타나고 있다. 그러나, PST의 경우 특별히 over-estimate 혹은 under-estimate하는 경향은 나타나지 않으며, 현금유입량 크기와 관계없이 전 구간에 걸쳐서 고른 오차값 분포를 보여주고 있다. 특히 큰 규모의 현금유입량이 발생한 경우 큰 오차를 보여주는 다른 방법들과는 달리, PST는 비교적 높은 정확도를 보여주고 있음을 <Figure 4>에서 확인할 수 있다.

본 실험에서 사용한 해당 업체의 고객사 수가 제한적인 관계로 많은 수의 검증 및 비교분석은 수행하지 못하였으나, 실험 결과만을 놓고 보았을 때 기존에 사용되어 왔던 알고리즘들과 비교하여 PST의 정확도가 크게 떨어지지 않음을 알 수 있다. 직관적으로 본다면 기법 2와



〈Figure 4〉 Difference between the Acutual Value and the Forecasted Value

3의 경우 내부적으로 마코프 모델을 사용하나, 1차 모델에 그치며 이로 인해 과거에 있을 수 있는 특정 패턴 분석에는 제한적일 수 있다. 이러한 과거의 특정 패턴 정보를 PST에서는 적극 반영하여 예측 값 형성에 큰 영향을 줄 수 있다. 즉, 마코프 모델을 사용하는 관점에서 놓고 보았을 때, VOM 모델의 한 종류인 PST 모델과 1차 마코프 모델을 사용하는 다른 기법들과는 활용하는 정보량 자체가 다르며, 이러한 이유로 PST가 일반적으로 우수한 성능을 보이게 되는 것이라 할 수 있다. 다만 PST의 경우 오차 값의 편차가 다른 기법보다 상대적으로 크게 발생하며(최소 1.24%, 최대 129.26%), 이로 인해 예측의 안정성이 다소 떨어지는 특성도 역시 발견하였다. 이는 순수하게 마코프 모델만을 사용하여 예측을 시도하면서 나타나는 결과로 해석될 수 있고, 만약 지수평활법과 같은 과거 데이터와의 가중치 합산을 통해 보정한다면 더 나은 안정성을 확보 할 수 있을 것이라 예측된다.

5. 결 론

대부분의 기업에서는 발행된 매출채권이 언제 회수되는지 정확히 알 수 없는 상황으로 인해 예기치 못한 현금 유동성 위험을 겪을 수 있다. 이 경우 기존의 매출채권 회수 패턴을 분석하여 불확실성을 줄이는 일이 필요하다. 본 연구에서는 VOM 모델 중 하나인 PST를 활용하여 매출 채권으로부터 유입되는 현금흐름 예측을 수행하는 방법을 제안하였다. 일반적으로 PST는 생물학 및 의학 분야에서 사용되어 왔던 시퀀스 분석 툴이나, 매출채권 회수 금액 예측과 같은 새로운 분야에도 적용될 수 있음을 본 연구에서 보여주었다. 특히, 제안한 예측방법의 실험결과를 기존에 사용되어 오던 현금흐름 예측기법들과 비교했을 때, 비교적 높은 정확도로 현금유입값을 예측하였다. 다만 예측 안정성 측면에서 PST를 활용하는 기법이 타 기법과 비교하였을 때 다소 떨어질 수 있음을 실험을 통해 밝혔다. 추후에, 더 많은 데이터를 이용하여 PST의 정확도 검증이 필요해 보이며, 이론적으로 어떠한 부분이 정확도의 차이를 가져다주는 지에 대한 구조적 분석 역시 필요하다. 또한 변동성이 큰 PST의 특성을 어떻게 안정화시킬 수 있을지에 대한 연구 역시 추가되어 해당 알고리즘을 보완할 수 있다면, 매출 채권 회수 패턴분석이 필요한 각 기업에서 해당 기법이 실질적으로 유용하게 활용될 수 있을 것이다.

References

[1] Bejerano, G. and Yona, G., “Variations on probabilistic suffix trees: statistical modeling and prediction of protein families,” Bioinformatics, Vol. 17, No. 1, pp. 23-43, 2001.

[2] Choe, H. and Shim, J., “Experimental Study on Random Walk Music Recommendation Considering Users’ Listening Preference Behaviors,” The Journal of Society for e-Business Studies, Vol. 22, No. 3, pp. 75-85, 2017.

[3] Cleary, J. and Witten, I., “Data com-

pression using adaptive coding and partial string matching," IEEE Transactions on Communications, Vol. 32, No. 4, pp. 396-402, 1984.

[4] Corcoran, A. W., "The use of exponentially-smoothed transition matrices to improve forecasting of cash flows from accounts receivable," Management Science, Vol. 24, No. 7, pp. 732-739, 1978.

[5] Pate-Cornell, M. E., Tagaras, G., and Eisenhardt, K. M., "Dynamic optimization of cash flow management decisions: a stochastic model," IEEE Transactions on Engineering Management, Vol. 37, No. 3, pp. 203-212, 1990.

[6] Ron, D., Singer, Y., and Tishby, N., "The power of amnesia: Learning probabilistic automata with variable memory length," Machine Learning, Vol. 25, No. 2-3, pp. 117-149, 1996.

[7] Tangsucheeva, R. and Prabhu, V., "Stochastic financial analytics for cash flow forecasting," International Journal of Production Economics, Vol. 158, pp. 65-76, 2014.

[8] Willems F. M., Shtarkov, Y. M., and Tjalkens, T. J., "The context-tree weighting method: basic properties," IEEE Transactions on Information Theory, Vol. 41, No. 3, pp. 653-664, 1995.

[9] Ziv, J. and Lempel, A., "A universal algorithm for sequential data compression," IEEE Transactions on Information Theory, Vol. 23, No. 3, pp. 337-343, 1977.

저 자 소 개

강윤철	(E-mail: yckang@hongik.ac.kr)
2002년	KAIST 산업공학과 (학사)
2004년	서울대학교 산업공학과 (석사)
2014년	Pennsylvania State University, Industrial and Manufacturing Engineering, (Ph. D.)
2004년~2007년	LG CNS
2007년~2008년	서울대학교 자동화 연구소
2014년~2016년	Research Associate, Pennsylvania State University
2016년~현재	홍익대학교 산업공학과 조교수
관심분야	Decision-making under uncertainties

강민지	(E-mail: jaellen@naver.com)
2019년 예정	홍익대학교 정보컴퓨터공학부 산업공학과 (학사)
관심분야	Financial Engineering, Data Analytics

정광헌	(Email: khchung@hongik.ac.kr)
1997년	서울대학교 산업공학과 (학사)
1999년	서울대학교 산업공학과 (석사)
2010년	University of Florida, Industrial & Systems Engineering (Ph.D.)
1999년~2004년	삼성 SDS
2010년~2011년	Post-Doctoral Fellow, Center for Operations Research and Econometrics(CORE), Belgium
2011년~현재	홍익대학교 경영학부 조교수
관심분야	Optimization, SCM, Energy, Healthcare Service