JKIICE Journal of the Korea Institute of Information and Communication Engineering

한국정보통신학회논문지 Vol. 24, No. 7: 827~833, Jul. 2020

빅 데이터 기반의 체납 수용가 예측 모델

정재안[1] · 이규환[1] · 정회경[2*]

Prediction Model for Unpaid Customers Using Big Data

Jaean Jeong[1] · Kyouhwan Lee[1] · Hoekyung Jung[2*]

[1]Graduate Student, Department of Computer Engineering, Paichai University, Daejeon 35345, Korea
[2*]Professor, Department of Computer Engineering, Paichai University, Daejeon 35345, Korea

요 약

본 논문에서는 지자체의 요금 체납을 줄이기 위해 특정 지자체를 대상으로 검침원의 면담 등을 통해 지방상수도 통합정보시스템에서 체납에 영향을 미치는 내부 데이터 요소를 찾았다. 또한 국가 통계 데이터 중에서 체납에 영향을 미치는 후보 데이터를 도출하였다. 독립변수가 종속변수에 미치는 영향도는 정보이득이라는 데이터 집합에서 종속변수에 대한 무질서도를 조사하여 표본 데이터를 수집하였다. 그리고 빅 데이터 분석 알고리즘인 의사결정트리와 로지스틱 회귀기법 중 어느 알고리즘이 더 높은 예측율을 나타내는지 n-fold cross-validation 방법을 사용하여 평가하였다. 이를 통해 지자체의 데이터를 기초로 알고리즘의 성능을 비교한 결과 의사결정트리가 로지스틱회귀보다 더 정확한 수용가 납부 패턴을 찾을 수 있음을 확인하였다. 머신러닝을 이용한 분석 알고리즘 모델 개발의 과정에서는 알고리즘의 정확성 향상을 위해 의사결정트리의 복잡성과 정확성에 직접적인 영향을 주는 최소 데이터 개수와 최대 순도라는 두 개의 환경변수의 최적값을 도출하였다.

ABSTRACT

In this paper, to reduce the unpaid rate of local governments, the internal data elements affecting the arrears in Water-INFOS are searched through interviews with meter readers in certain local governments. Candidate data affecting arrears from national statistical data were derived. The influence of the independent variable on the dependent variable was sampled by examining the disorder of the dependent variable in the data set called information gain. We also evaluated the higher prediction rates of decision tree and logistic regression using n-fold cross-validation. The results confirmed that the decision tree can find more accurate customer payment patterns than logistic regression. In the process of developing an analysis algorithm model using machine learning, the optimal values of two environmental variables, the minimum number of data and the maximum purity, which directly affect the complexity and accuracy of the decision tree, are derived to improve the accuracy of the algorithm.

키워드 : 빅데이터 분석, 요금 체납, 의사결정트리, 지방상수도

Keywords : Big data analysis, Unpaid, Decision tree, Local waterworks

Received 6 March 2020, **Revised** 11 April 2020, **Accepted** 22 April 2020

*** Corresponding Author** Hoekyung Jung(E-mail:hkjung@pcu.ac.kr, Tel:+82-42-520-5640)
Professor, Department of Computer Engineering, Paichai University, Daejeon 35345, Korea

Open Access http://doi.org/10.6109/jkiice.2020.24.7.827 print ISSN: 2234-4772 online ISSN: 2288-4165

ⓒThis is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License(http://creativecommons.org/li-censes/by-nc/3.0/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.
Copyright Ⓒ The Korea Institute of Information and Communication Engineering.

Ⅰ. 서 론

수용가의 수도요금이 장기간 체납되면서 발생하는 결손금을 줄이기 위해 체납 수용가가 발생하지 않도록 하려는 다양한 노력과 함께 체납 수용가를 사전에 예측하여 체납이 발생하지 않도록 하는 사전 관리의 필요성이 대두되었다.

본 논문에서는 수용가의 체납 발생에 영향을 줄 것으로 예상되는 데이터를 K사가 운영하는 데이터와 정부가 운영하는 국가통계포털로부터 수집하였다. 지방상수도통합정보시스템은 지방상수도 통합 정보를 관리하고 있어 고객정보, 민원정보, 검침정보, 고지정보 및 수납정보 등 데이터 항목을 확보할 수 있었고, 국가통계포탈로부터는 전국주택가격 동향 조사 데이터, 국세통계 체납액 정리 데이터 등을 확보할 수 있었다. 이렇게 수집한 데이터들에 대해서는 정제 등의 데이터 전처리를 통해 체납예측 모델을 개발하는데 사용하고, 모델 전산화 및 시각화 서비스를 개발한 후 예측모델의 검증을 위해 K사가 운영하고 있는 지자체를 대상으로 시범운영을 통해 예측도를 검증하였다.

Ⅱ. 빅데이터 분석 기법

본 장에서는 빅 데이터 분석서비스 제공을 위해 빅 데이터 분석 기법을 기술하고 연구 분석한다. 빅 데이터 분석을 위에 사용되는 기법들은 다양하지만 주로 데이터 마이닝 기법이 사용된다. 인터넷을 통한 다양한 글과 SNS상에서 생산되는 다양한 텍스트를 분석하기 위한 텍스트 마이닝, SNS상에서 어떻게 사회 기능이 이루어지는가에 대한 관점을 분석하기 위해 개인과 연결된 사회 조직간의 관계를 규명하는 기법인 소셜 분석을 사용한다. 유사한 그룹을 묶어 새로운 지식과 가치를 얻는 방법인 클러스터 분석, 특정 항목에 특화된 공간데이터 분석 등 데이터 분석 목적에 따라 다양한 기법들이 지속적으로 발전하고 있다[1].

기존의 데이터 마이닝 기법들은 일반적으로 정형화되어 빅 데이터에 적용되고 있고, 예측하고자 하는 변수의 특성을 고려하여 k-평균법, 로지스틱 회귀, 선형 회귀분석, 계층적 군집화, 주성분 분석, 신경망, 서포트 벡터 머신 등 다양한 방법론들이 사용된다. 분석 목적이 반응변수 예측인지, 반응변수에 대한 영향력의 해석인지에 따라 모형 적절성, 해석의 난이도를 따져 적절한 데이터 마이닝 기법을 사용하는 것이 좋다[2].

텍스트 마이닝은 비구조화 형태의 대규모 텍스트 집합에서 정보를 추출해내는 과정으로 기존에는 직접 설문한 결과나 네티즌 반응 등을 살펴 얻을 수 있던 정보였으나, 데이터의 양이 급증함에 따라 자동으로 텍스트의 종류, 주관성, 주요 단어 등을 분석하는 자연어처리가 필요하다[3-6]. 물동량 분석에서는 기존의 군집분석, 회귀분석, 시계열분석 등의 한계를 개선하기 위해 비정형 데이터 수집과 흐름 분석을 통한 물동량 예측 연구도 진행되고 있다.

Ⅲ. 데이터 구축

본 장에서는 상수도 요금 체납 수용가를 예측하는 모델을 개발하기 위해 필요한 데이터 기능 및 데이터 확보 방안, 데이터 수집환경 구성을 통한 데이터 수집 및 데이터 전처리 방법 등을 소개한다.

3.1. 분석 데이터 정의 및 독립변수 선정기준

수용가의 납부 형태인 납기 내, 납기 후, 체납 등에 영향을 미치는 것으로 판단되는 요금조정, 요금감면, 수납, 민원 및 수도미터 정보 등 지방상수도통합정보시스템 보유 데이터와 국세 체납액, 물가지수 등 국가통계 제공 데이터를 분석에 활용할 분석 데이터로 정의하였다. 또한 독립변수 선정기준은 종속변수인 체납여부에 영향을 미칠 수 있는 데이터인 체납예측 시나리오에서 제시된 데이터, 지자체 검침원의 체납수용가 검침 경험에 대한 인터뷰를 통해 제시된 데이터와 수도 요금에 영향을 미치는 수용가의 지역적 특성을 담은 데이터와 관련 있는 내부 데이터를 독립변수로 선정하였다.

3.2. 데이터 수집 환경 구성

내·외부에서 수집한 데이터를 한 곳으로 모아서 빅데이터 분석을 위한 시스템은 메인 서버(Name node) 1대와 데이터를 분산 저장하는 데이터 서버(Data node) 2대로 구성하였다.

메인 서버는 빅데이터의 주요 기능인 분산파일 시스템에서 디렉터리, 파일명, 파일블록 등 데이터의 속성

값인 메타 정보를 관리하는 마스터(Master) 역할을 하며, 슬레이브(Slave)에 해당하는 데이터 서버에게 데이터 입출력과 관련된 작업 지시 등 총괄 역할을 수행한다. 데이터 서버는 다양한 소스에서 발생하는 각종 데이터를 분산파일 시스템에 의해 메인 서버에서 블록 단위로 쪼개고 이렇게 분산된 데이터를 데이터 서버의 로컬 디스크에 저장 및 복제를 통해 데이터 입출력에 대한 처리와 데이터 백업에 대한 역할을 수행한다.

빅데이터 분석은 대용량 데이터를 기반으로 진행되므로 이러한 대용량 데이터를 저장하기 위해 하둡의 분산파일 시스템을 사용하여 마스터 1대와 슬레이브 2대의 서버로 분석 서버를 구성하였다.

3.3. 데이터 수집

지자체에서 당월 요금계산이 끝나는 시점에 빅 데이터 분석용 DB로 전송하도록 Ankus 스케쥴러를 설정하여 데이터를 자동으로 수집하도록 하고, 외부 데이터는 국가통계포탈에서 CSV 파일 형태로 다운받아 빅 데이터 분석용 DB에 저장하였다. 추가적으로 통계 데이터 업데이트가 이뤄진 경우 CSV 파일을 새로 받아서 분석용 DB에 저장되도록 하였다.

3.4. 데이터 전처리

3.4.1. 데이터 정제

데이터 활용도의 정밀도를 높이기 위해 이상값 검출, 데이터 중복 검출, 데이터의 차이를 추출하는 여러 기법들이 있다[7,8]. 분석 기초 데이터들 중에서 활용 범위를 벗어나는 이상 값 또는 값이 null로 되어 있는 결측 값이 존재하는지 여부를 확인한 후 해당 항목은 분석 기초 데이터에서 제외하였다.

3.4.2. 영향도 조사

분석하고자 하는 데이터가 종속변수인 수용가 납부유형을 얼마나 잘 분별할 수 있는지, 즉 결과에 미치는 영향이 얼마나 되는지를 조사하였다. 독립변수가 종속변수에 미치는 영향도를 정보이득(Information Gain)이라는 데이터 집합에서 종속변수에 대한 무질서도를 통해서 조사하였다.

표 1은 변수에 대한 엔트로피 예로 독립변수인 고지방법 중 일반지로의 경우는 수용가 납부유형이 납기 내, 납기 후, 체납으로 나눠져 있고, 자동이체는 전부 납기 내의 값을 보이는 상황에서 정보이득 산정식을 구하는 예를 보여주고 있다. 엔트로피 수치가 높을수록 데이터 집합의 무질서도는 높은 것으로 판단한다.

Table. 1 Example entropy for a variable

Notice type (Independent variable)	Payment mtd. (Dependent variable)	Entropy
Giro	Payment after due date	Entropy(notice type, giro)= $\left(-\frac{1}{6}log_2\frac{1}{6}\right)+\left(-\frac{2}{6}log_2\frac{2}{6}\right)+\left(-\frac{3}{6}log_2\frac{3}{6}\right)=0.91$
Giro	Payment after due date	
Giro	Be in arrears	
Giro	Paid by the due date	
Giro	Be in arrears	
Giro	Be in arrears	
Direct debit	Paid by the due date	Entropy(notice type, E. trans) = $\left(-\frac{6}{6}log_2\frac{6}{6}\right)+\left(-\frac{0}{6}log_2\frac{0}{6}\right)+\left(-\frac{0}{6}log_2\frac{0}{6}\right)=0.00$
Direct debit	Paid by the due date	
Direct debit	Paid by the due date	
Direct debit	Paid by the due date	
Direct debit	Paid by the due date	
Direct debit	Paid by the due date	

3.4.3. 표본 추출

종속변수의 유형 중 특정 값의 빈도가 높을수록 결과에 미치는 영향이 큰 이유로 최소 빈도를 기준으로 분석 데이터의 표본을 추출하였다. 그림 1은 각 조정년월별 종속변수의 분포를 나타낸다.

3.4.4. 데이터 저장

DB에 저장되어 있는 데이터는 Ankus의 DB2DB 기능을 사용하여 메인 서버의 DB에 저장하였다. 다운받은 CSV 파일은 ssh에서 명령어 입력을 통해 DB에 적재하였다. 그림 2는 데이터를 CUBRID DB에 저장하기 위해 설정된 화면이다.

Ⅳ. 모델 구축

본 장에서는 모델 구축의 내용을 기술한다.

4.1. 알고리즘 비교

빅데이터 분석 알고리즘인 의사결정트리(Decision Tree)와 로지스틱 회귀(Logistic Regression) 기법 중 어

느 알고리즘이 더 높은 예측율을 나타내는지 사전 검증 후 사용할 알고리즘으로 선정하였다.

Paid by the due date

month	2015	2016	2017
1	363,725	388,210	408,428
2	351,220	380,693	408,610
3	364,286	381,879	420,979
4	363,164	389,736	418,043
5	366,885	394,017	423,934
6	372,215	397,915	-
7	376,882	402,498	-
8	380,187	407,102	-
9	374,514	404,946	-
10	381,879	405,304	-
11	377,782	406,472	-
12	388,299	414,565	-

Payment after due date

month	2015	2016	2017
1	19,010	23,412	26,626
2	26,813	29,984	27,175
3	21,963	22,340	20,791
4	23,859	24,070	21,045
5	23,309	24,430	18,135
6	22,810	23,382	-
7	21,311	22,752	-
8	21,618	21,445	-
9	25,590	23,194	-
10	22,340	25,015	-
11	27,353	25,982	-
12	22,331	20,993	-

Be in arrears

month	2015	2016	2017
1	57,646	56,179	61,441
2	63,876	58,645	63,865
3	56,910	55,369	58,910
4	58,212	58,124	63,728
5	55,981	55,145	61,938
6	53,355	54,685	-
7	53,039	54,406	-
8	51,963	52,889	-
9	55,780	55,129	-
10	55,369	57,017	-
11	56,156	56,743	-
12	53,411	56,444	-

Fig. 1 Distribution of Dependent Variables by Adjusted Month

4.1.1. 의사결정트리

데이터로부터 추출한 패턴을 그림 2와 같이 변수 명을 의미하는 노드와 변수 값을 의미하는 링크로 구성하여 표현한다. 부모 노드가 없는 최상위 루트 노드부터 자식 노드가 없는 최하위 단말 노드까지의 경로를 IF-THEN의 규칙으로 표현할 수 있어 분석 결과 패턴의 가독성이 높다. 종속변수를 최상위 루트로 하여 정보 이득 값이 높은 순서로 상위 노드에 위치하며, 데이터 분석 결과 최하위 노드에 나오는 규칙이 체납 예상 수용가의 납부 유형과 예상 율을 보여주는 방식이다.

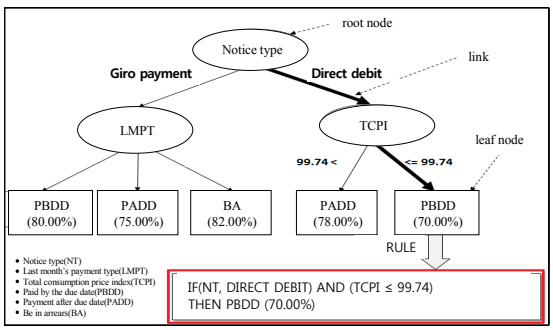

Fig. 2 Decision Tree Example

4.1.2. 로지스틱 회귀

종속변수와 독립변수간의 관계를 일차다항식 식 (1)으로 표현하며, 각 독립변수의 계수는 종속변수에 미치는 영향을 의미한다. 논문에서는 납기 후 회귀식의 전월 수납방법건수가 수용가가 요금을 납기 후에 내는데 긍정적 영향을 준다는 의미로 분석하는 활용하였다.

$$y = w_0x_0 + w_1x_1 + \cdot \cdot \cdot \ w_nx_n + \alpha \qquad (1)$$

4.2. 알고리즘 평가

두 가지 알고리즘의 분석 결과에 대한 평가를 수행하기 위하여 본 논문에서는 n-fold cross-validation 방법을 사용하였고, 검증 결과를 기초하여 분석에 사용할 알고리즘 선정 및 분석 결과 신뢰도를 확인하였다.

전체 실험 데이터를 n개의 집합으로 분할한 후 n개 중 한 개의 집합들에 대해 생성된 모델을 평가할 때 이용하는 테스트 데이터로 사용하고, 나머지 집합들에 대해 모델을 생성할 때 이용하는 학습 데이터로 사용하였다. 각 집합이 전부 테스트 데이터로 사용될 수 있도록 총 n번의 실험을 실시하였고, 그림 3은 n-fold cross-validation의 예 중 5개 집합으로 실험하는 5-fold cross-validation 방법의 처리 과정을 보여주고 있다.

Table. 2 Algorithm Performance Comparison Result

	Decision hierarchy Tree	Logistic regression
Experiment data 1	76.29%	75.65%
Experiment data 2	76.32%	75.61%
Experiment data 3	76.29%	75.63%
Experiment data 4	76.18%	75.69%
Experiment data 5	76.39%	75.69%
Mean (±standard deviation)	76.294% (±0.075)	75.654% (±0.035)

지자체 데이터를 기초로 하여 실시한 알고리즘의 성능 비교 결과 표 2와 같이 의사결정트리(76.294%, ±0.075)가 로지스틱회귀(75.654%, ±0.035)보다 더 정확한 수용가 납부 패턴을 찾을 수 있음을 확인하였다.

n-fold cross-validation을 5회 수행한 결과에 따른 두 알고리즘의 평균값 간의 차이(0.64%)가 유의미한지 검증하기 위하여 t-test를 실시하였다. t-test는 하나의 귀무가설을 수립하고 p-value라는 귀무가설이 맞다는 전제하에 표본으로 뽑은 통계량이 우연히 나타날 수도 있다는 유의확률 기법에 따라 p-value < 0.05 인 경우는 귀무가설을 거짓으로 판명하고 해당 귀무가설을 기각하는 방식으로 검증하였다.

두 알고리즘의 평균값 간의 차이를 검증하기 위하여 "두 알고리즘의 평균값은 차이가 없다"라는 귀무가설을 수립하고 p-value를 구한 결과 p-value는 1.395e-07의

값을 보이고, 이는 p-value < 0.05 인 경우로 귀무가설을 거짓으로 판명하고 두 알고리즘의 평균값에는 차이가 있는 유의미한 결과라는 검증이 가능하다는 결론을 얻었다.

두 가지 알고리즘의 평균값은 유의미한 차이가 있음을 확인하고 그 차이를 바탕으로 더 높은 예측율을 보인 의사결정트리 알고리즘을 분석 알고리즘 모델로 선정하였다.

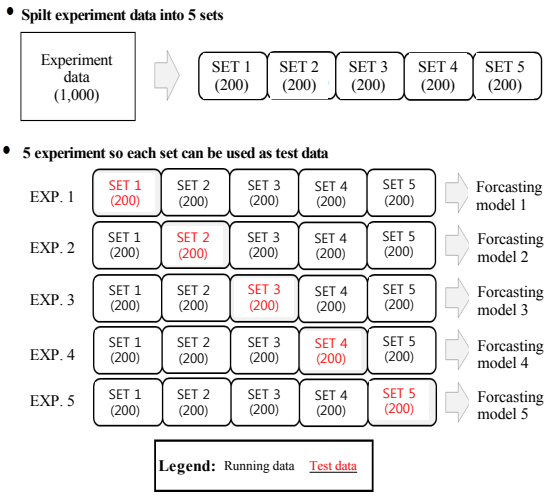

Fig. 3 5-fold Cross-validation Example

4.3. 모델 구현

4.3.1. 분석 알고리즘 모델링

정제 과정을 거친 데이터를 기초로 하여 머신러닝을 통한 분석 알고리즘 모델을 개발하는 과정으로 의사결정트리 알고리즘의 환경변수 설정을 통해 복잡성과 정확성 향상을 기대할 수 있다.

의사결정트리의 복잡성과 정확성에 직접적인 영향을 주는 "최소 데이터 개수"와 "최대 순도" 두 개의 환경변수의 최적 값을 도출하였다. 최소 데이터 개수는 의사결정트리가 생성될 때 노드에 포함된 데이터 개수의 최소값을 정하는 변수로 너무 적은 데이터 개수를 가지고 자식 노드를 생성할 경우 규칙의 정확성이 떨어지므로 최소 값을 정해서 더 이상 자식노드 다시 말해 새로운 규칙을 생성하지 않도록 설정하였다. 최대 순도는 의사결정트리가 생성될 때 노드에 포함된 종속변수 값 중 가장 많은 빈도가 최대 순도 보다 클 경우 신규 생성한 자식 노드의 결과가 부모노드의 결과와 큰 차이가 없으므로 더 이상 자식노드 다시 말해 새로운 규칙을 생성하지 않도록 설정하였다.

표 3은 지자체의 107,274개의 데이터에 대해 최소 데이터 개수와 최대 순도의 변화에 따라 생성된 의사결정트리의 정확성과 규칙 개수를 보인다. 표 3에서 보듯이 최소 데이터 개수는 적을수록, 최대 순도는 클수록 생성되는 규칙의 수가 많아지는 것을 볼 수 있다. 하지만 생성된 규칙의 개수가 많아지면 생성된 의사결정트리의 정확성이 낮아지는 현상도 발생한다. 최소 데이터 개수에서 100으로 설정한 그룹의 예측율 평균이 가장 높아서 최소 데이터 개수는 100개로 선정하였으며, 최대 순도는 0.25와 0.10의 평균 차이가 유의미한지 검정하기 위해 앞에서 설명한 t-test를 통해 76.31%와 76.29% 차이의 의미를 확인하였다. 두 평균의 차이 0.20%는 의미가 없다는 귀무가설을 수립하고 실시한 t-test 결과 도출된 p-value는 0.6921로 0.05보다 크므로 두 평균간 정확성 차이는 없다고 검증되어 규칙을 보다 적게 생성한 최대 순도 0.10을 알고리즘 환경변수로 선정하였다.

Table. 3 Impact on Minimum Data Count and Maximum Purity

Specific local gov.(107,274EA)		Minimum data(%, EA)		
		100	200	300
Maximum purity	0.35	76.29/91	76.31/60	76.13/42
	0.25	76.31/78	76.27/54	76.11/38
	0.10	76.29/62	76.26/48	76.05/28

4.3.2. 모델 개발

표 4는 지자체의 데이터를 기초로 하여 Ankus 분석 결과로 나온 규칙을 앞에서 설명한 n-fold cross-validation을 5회 수행하여 도출한 당월 납부유형 규칙의 결과를 보인다. 예시에서 규칙은 앞에서 수행한 의사결정트리 알고리즘이 분석 데이터를 기초로 하여 각 변수의 조합을 통해 체납 확률을 예측한 규칙으로 규칙 조건은 (1) 현재 체납이 있는 수용가이며, (2) 일반지로 방식으로 고지서를 받고, (3) 인터넷을 통한 납부 실적이 없고, (4) 최근 6개월간 체납 건수가 4회를 초과한 수용가를 의미한다.

가능성은 n-fold cross-validation을 실행 결과 체납으로 예측된 수용가 중에서 실제 과거 수납 형태가 체납인 수용가의 %를 의미한다. 실험 2, 3, 5의 경우는 체납-규

칙1이 아닌 다른 규칙이 가장 높은 가능성을 보인 상황으로 규칙1에서는 해당 없음의 의미로 ‘ - ’로 표시했다. 예측 분석 모델은 최근 1년의 데이터를 분석하여 새로운 예측 모델로 갱신을 하며, N지자체 요금 데이터를 매월 특정시기에 자동으로 수집할 수 있는 스케줄러를 구성하여 데이터를 입력되는 데이터의 분석 결과로 제시하였다.

Table. 4 Sample Payment Type Forecast Rule

Rule	Condition of Rule	Possibility (%)	Exp. num.
Arrear Rule 1	(1) Number of last month’s arrears > 0 AND (2) Notice type = Giro AND (3) Number of last month’s payment method_ eGiro <=0 AND (4) Recent arrear‘s ratio > 0.67	92.00	1
		-	2
		-	3
		92.13	4
		-	5

4.4. 분석결과 시각화

본 논문에서는 Ankus를 기반으로 작업을 수행하였다. 그림 4는 분석결과에 대한 상세 내용을 보여 주고 있으며, 수용가는 예측율이 가장 높았던 하나의 규칙에 의해 결과를 예측하게 되며, 해당 규칙은 전체 데이터 중 차지하는 비율과 해당 규칙 내 납부 유형의 분포를 보여준다.

그림 4는 지자체의 107,274개의 데이터를 의사결정 트리 알고리즘으로 예측 모델 생성 결과 도출한 규칙을 표시한 내용을 보여주고 있다.

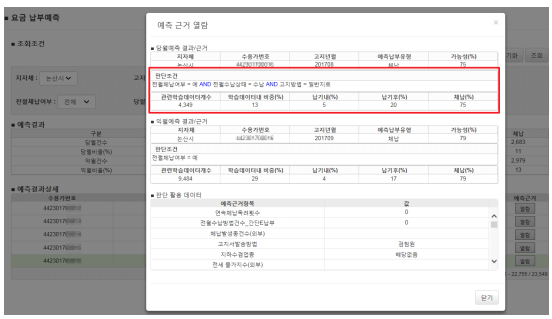

Fig. 4 Analysis Result Screen

4.5. 분석결과 검증

개발된 모델에 대한 검증을 위해 분석 프로세스, 시각화 등의 결과물을 검증하였다. 그림 5에 소비자 지불 패턴 화면을 보이며, 논문에서 개발한 요금 체납 수용가 사전 예측 결과와 실제 납부 현황의 비교 결과는 표 5와 같으며, 빅 데이터 분석에 사용할 기초 데이터 다양화와 분석 모델을 적용할 지자체의 확대를 통해 예측 정확도를 더 높일 수 있도록 해야 할 필요성이 있는 것으로 분석되었다.

Fig. 5 Consumer Payment Pattern Screen

Table. 5 Predecition result and actual payment form comparion result.

Timing of notice	Arrears prediction (A)	Payment (B)	Accuracy (B/A)
2017. 7	2,685	1,719	64.02%
2017. 8	2,683	1,713	63.85%
2017. 9	3,267	1,872	57.30%

Ⅴ. 결 론

본 논문은 지자체의 상수도 체납요금을 줄이는 방안으로 지자체를 대상으로 요금고지 시 체납수용가를 사전 예측하여 예상 체납수용가들을 대상으로 다양한 알림 활동을 통해 체납율을 줄이는 목적으로 연구하였다.

개발된 분석서비스는 지자체를 대상으로 실제 업무에 적용하여 운영한 결과 기대한 예측치 보다 낮게 나왔다.

본 서비스를 운영하면서 체납에 영향을 미치는 변수가 새로 도출될 수 있고 개발된 모델의 보완도 필요하다고 판단된다. 결과를 지속적으로 활용하기 위해서는 데이터 수집·가공·분석기법 추가, 예측모델 보완, 데이터 가시화 방법 등 분석서비스 추가가 필요하고, 데이터의 수집 종류 및 분석대상 데이터의 범위를 확대하여 기존

분석 결과와 결합함으로써 분석 범위 확장이 필요하다. 또한 초기 데이터와 추가 데이터의 구조적인 이질성, 데이터 항목간의 의미적인 이질성 및 데이터 값의 이질성 등을 식별하여야 하며, 분석에 영향을 미치는 데이터들이 있는지 검토와 함께 랜덤 포레스트 방법 등 지속적으로 모델을 개선해 나갈 필요가 있다.

ACKNOWLEDGEMENT

This work was supported by the research grant of Pai Chai University in 2020.

REFERENCES

[1] J. K. Hong, "Analysis of Sales Volume by Products According to Temperature Change Using Big Data Analysis," *The Korea Journal of BigData*, vol. 4 no.2. pp. 85-91, 2019.

[2] S. H. Back, "Sales Volume Prediction Mode for Temperrature Change using Big Data Analysis," *The Korea Journal of BigData*, vol. 4 no.1, pp. 29-38, 2019.

[3] D. S. Lee, "The Trends of Next Generation Cyber Security," *Journal of the Korea Institute of Information and Communication Engineering*, vol. 23, no. 11, pp. 1478-1481, Nov. 2019.

[4] J. P. Yu, "A Model of Predictive Movie 10 Million Spectators through Big Data Analysys," *The Korea Journal of BigData*, vol.3, no.1, pp.63-71, 2018.

[5] D. J.Park and W.S. Kim, "Improvement of the Paralled Importation Logistics Process Using Big Data," vol. 17, no. 4, pp. 267-273, Dec. 2019.

[6] Y. C. Choung, "Detection of redundant data in big data environment," *ITPM*, vol. 11, no.3, pp. 1227-1232, 2019.

[7] K. S. Choi, "K-SuperCast: A big data based GDP forecasing model," *Journal of the Korea Data & Information Science Society*, vol.30, no.4, pp. 723-743, 2019.

[8] J. M. Jo, "Effectiveness of Normalization Pre-Processing of Big Data to the Machine Learning Performance," *The Journal of the Korea Institute of Electronic Communication Science*, vol.14, no.03, pp. 547-552, 2019.

정재안(Jaean Jeong)

1986년 전북대학교 전산통계(이학사)
2020년 배재대학교 컴퓨터공학(공학석사)
2020년 ~ 현재 배재대학교 컴퓨터공학(박사과정)
1993년 ~ 현재 한국수자원공사 (수자원연구소, 정보관리처, 소프트웨어설계처 등 근무)
※관심분야 : 데이터 수집 및 공유, 데이터 모델링, 빅 데이터 아키텍쳐 설계 및 분석

이규환(Kyouhwan Lee)

1990년 충남대학교 계산통계학과(이학사)
2019년 배재대학교 컴퓨터공학과(공학석사)
2020년 ~ 현재 배재대학교 컴퓨터공학과 (박사과정)
※관심분야 : 빅데이터, 인공지능, ERP, XML, IoT

정회경(Hoekyung Jung)

1985년 광운대학교 컴퓨터공학과(공학사)
1987년 광운대학교 컴퓨터공학과(공학석사)
1993년 광운대학교 컴퓨터공학과(공학박사)
1994년 ~ 현재 배재대학교 컴퓨터공학과 교수
※관심분야 : 멀티미디어 문서정보처리, USN, IoT, Machine learning, Bigdata, Embedded system