반도체디스플레이기술학회지 제18권 제2호(2019년 6월)
Journal of the Semiconductor & Display Technology, Vol. 18, No. 2. June 2019.

머신러닝을 이용한 세금 계정과목 분류

최동빈[*]·조인수[**]·박용범[**†]
[*]단국대학교 컴퓨터학과, [**†]단국대학교 소프트웨어학과

Taxation Analysis Using Machine Learning

Dong-Bin Choi[*], In-su Jo[**] and Yong B. Park[†]**
[*]Dankook University Dept. of Computer Science, [**†]Dankook University Dept. of Sofware Science

ABSTRACT

Data mining techniques can also be used to increase the efficiency of production in the tax sector, which requires professional skills. As tax-related computerization was carried out, large amounts of data were accumulated, creating a good environment for data mining. In this paper, we have developed a system that can help tax accountant who have existing professional abilities by using data mining techniques on accumulated tax related data. The data mining technique used is random forest and improved by using f1-score. Using the implemented system, data accumulated over two years was learned, showing high accuracy at prediction.

Key Words : Random Forest, Data Mining, f1-score, Taxation Analysis, Machine Learning

1. 서 론

데이터 마이닝 기법들의 발전으로 인하여, 많은 분야에서 활용되고 있다. 특히 전산화 되어 많은 데이터가 축적된 환경에서는 전문 능력을 요구하는 분야까지 적용되고 있다.

세무관련 작업에 전산화가 이루어 지면서 관련 데이터가 축적되고 있다. 또한 세무관련 업무는 그 업무의 특성상 전문 능력을 요구한다. 축적된 데이터를 활용하여 전문가를 보조할 수 있는 시스템을 구축하면, 그 생산성을 높일 수 있다.

R.Deepa Lakshmi[1]는 여러 머신 러닝 기법을 활용하여 taxation analysis의 효용성을 보여 주었다. M./s Nss and Co.회사의 365개 회원들의 정보를 토대로 분석하였다. 사용된 기법은 Naive Bayes, SMO, RBF, Bagging, Attribute Selected Classifier, JRip, PART, J48, LAD Tree 기법들이다.

실험 결과는 SMO와 LAD Tree가 정확도면에서 높은 결과를 보여주었다. 데이터의 특성이 다르기에 해당 기법을 그대로 적용하기에는 무리가 있으나, Tree 기법들이 효율적으로 나타나, 본 논문은 Random Forest기법을 차용하였다.

본 논문은 국내 세무관련 회사에서 제공한 데이터를 바탕으로 Random Forest를 사용하여 분류하는 시스템을 구현하였다.

본 논문은 구성은 다음과 같다. 2장에서는 분류기법으로 채택된 Random Forest기법과, 성능 측정에 사용된 f1-score에 관한 설명을 하겠다. 3장에서는 사용된 데이터의 분석을 하며, 4장에서 구현된 시스템과 성능 결과를 기술한다. 5장에 결론으로 마무리한다.

2. 관련 연구

2.1 Random Forest

Breiman[2]이 제안하고 Biau 등[3]에 의해 연구된 random forest기법은 분류 회기 분석등에 사용되는 앙상블기법 중 하나다. 학습 과정에서 구성된 다수의 decision tree로부터 분류 또는 회귀 분석을 출력함으로써 동작한다.

Random forest는 다음과 같은 알고리즘으로 작동한다.

[†]E-mail: ybpark@dankook.ac.kr

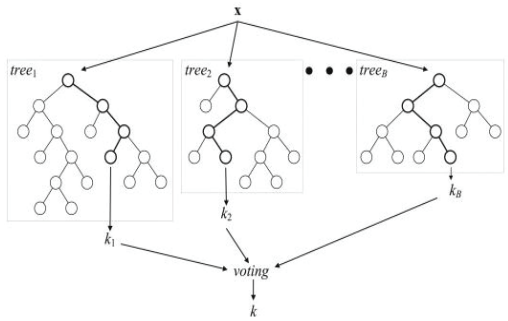

Fig. 1. A general architecture of a random forest[4]

Random Forest(Breiman, 2001)

Setting:

- L : training data set composed of n instances
- L_y : class membership of L
- B : number of classifiers in ensemble

Procedure:

For $b = 1$ *to* B

Step (1) Generate B bootstrap samples $L_1, ..., L_B$ from the original training data set L

Step (2) Grow a random forest tree using a random feature selection from bootstrapped data.
: randomly select $\sqrt{n}$ or $n/3$ predictors at each node and split the data using the best predictors where n is the number of variables in x.

Step (3) Construct train classifiers $C_b(x), b = 1, ..., B$ from each ensemble of trees

Step (4) Aggregate the B train classifiers using simple majority vote.

$$C^*(x) = \arg\max_y \sum_{b=1}^{B} I[C_b(x) = y]$$

Fig. 2. Random Forest algorithm[2]

Random forest는 기존 bagging에 bootstrap sample의 변수에 임의성을 더한 방법으로, 기존 bagging 기법보다 더 다양한 hyperplane을 가지게 되어 예측 및 분류 정확도를 개선시켰다.

Random forest는 DNA 결합 단백질의 식별, 비디오 객체의 분류, 하이퍼 스펙트럼 데이터의 분류, 식생 유형 발생의 예측을 포함한 다양한 연구에 활용되어 좋은 성과를 보인 기법이다[4, 6-12].

2.2 f1-score

1992년 Fourth Message Understanding Conference(MUC-4)에서 소개된 방법으로 기존에 사용된 precision과 recall을 이용한 조화평균이다[5]. 구하는 수식은 다음과 같다.

Precision과 recall을 이용한 조화 평균이기에 f1-score 방식은 불균형 데이터에서 좀더 정확한 척도를 제시해준다[5].

$$\text{Precision} = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

$$\text{Recall} = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

$$\text{F1} = 2 \times \frac{Precision * Recall}{Precision + Recall}$$

Fig. 3. F1-score

3. 데이터형태

3.1 데이터 정보

회사로부터 제공받은 데이터는 거래 종류에 따라 전자세금계산서, 전자계산서, 신용카드, 현금영수증로 구분된다.

그중 가장 많은 feature를 가진 전자세금계산서의 경우 Table 1과 같은 정보를 가지고 있다.

Table 1. Data features

구분	구분	데이터
(매입)세금계산서	기본정보	회사코드 파일명 작성일자 승인번호 발급일자 전송일자
	공급자 정보	공급자사업자등록번호 공급자종사업장번호 공급자상호 공급자대표자명
	수급자 정보	공급받는자사업자등록번호 공급받는자종사업장번호 공급받는자상호 공급받는자대표자명
	금액	합계금액 공급가액 세액 전자세금계산서분류 전자세금계산서종류 발급유형 비고 영수청구구분
	이메일	공급자이메일 공급받는자이메일 1 공급받는자이메일2
	품목	품목일자 품목명 품목규격 품목수량 품목단가 품목공급가액 품목세액 품목비고

신용카드나 현금영수증의 경우 위 feature중 품목에 해당하는 부분이 없다.

제공 받은 정보를 이용하여 얻으려는 계정과목과의 연관성을 파악하기 위해 교차 분석을 실시하여, 필요한 정보만은 추린 결과 Table 2와 같은 필요정보로 축약되었다.

Table 2. Key features

구분	정보
유형	매입 매출
거래종류	전자세금계산서 전자계산서 신용카드 현금영수증
회사(사업자번호)	123-45-67890
업종코드	123456
회사구분	일반 법인 면세 간이
거래처(사업자번호)	123-45-67890
품명	상품명

3.2 데이터 분포

회사에서 제공받은 데이터는 2017년도 2018년도 데이터로 각각 37000개다.

해당 데이터를 계정과목에 따라서 분포그래프로 그려보면 다음과 같다.

각 계정과목당 데이터의 개수의 차이가 매우 심한 불균형 데이터의 형태를 지니고 있음을 알수있다.

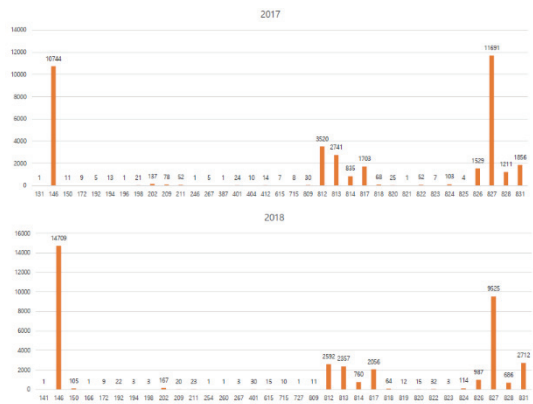

Fig. 4. Data distribution

4. 계정과목 분류 시스템

Table 2의 주요 특징을 이용하여 회사에서 요구한 계정과목 37종 중 하나로 분류하는 시스템을 구현하였다.

구현한 시스템의 전체 개요도는 다음 그림과 같다.

유형과 거래 종류에 따라 먼저 분류한 후 해당 전처리기를 통해서 학습된random forest모델에 입력된다.

주요 feature중 하나인 품명의 경우 그 형태와 종류가 매우

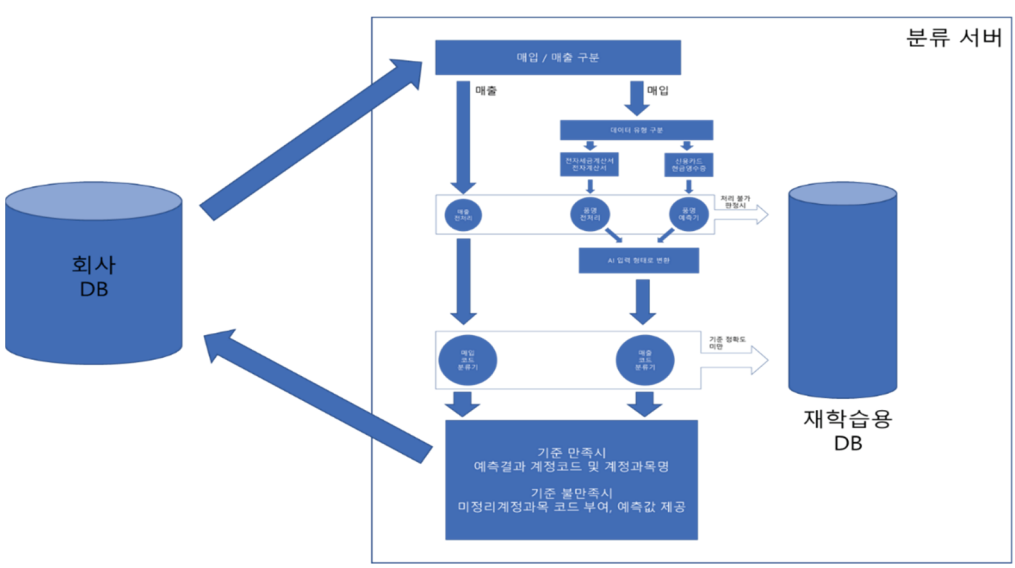

Fig. 5. System structure

다양하여 바로 분류기에 입력될수 없기에 해당 전처리 부분에서 품명에서 키워드를 추출하여 1차 분류를 실행한다.

또한 품명 항목이 없는 신용카드와 현금 영수증의 경우 분류 정확도를 높이기 위해 품명항목을 예측하는 품명 예측기를 사용한다.

품명 예측기도 random forest를 사용하여 학습하였으며, 전자세금계산서와, 전자계산서의 데이터를 이용하였다. 예측결과는 그 확률에 따라서 기준을 만족하지 못할시 재학습용 DB에 해당 데이터가 기록하게 하여, 추후 재학습시 사용하도록 구성하였다.

해당 전처리가 끝난 데이터는 random forest모델에 입력할 수 있는 형태로 변환되며, 학습된 모델에 입력되어 예측을 실행한다.

Random forest에서 나온 결과는 해당 계정과목에 해당할 확률을 기준으로 기준에 만족시 회사에 해당 결과를 알려주며, 만약 기준을 만족하지 못할시 추후 모델을 재학습할 수 있는 DB에 해당 데이터를 기록한다.

5. 결 론

구현된 시스템에 사용된 학습된 모델의 성능은 아래 표와 같다.

Table 3. Train F1-score

	Train	Test
2017	0.98	0.94
2018	0.97	0.94

하지만 세무 데이터의 특성상 항상 동일한 정보가 들어온다는 보장이 없으며, 실제로 2017년도의 데이터와 2018년도의 데이터에서 주어진 정보는 다른 정보가 많았다.

이와 유사한 상황을 가정하여 예측 실험한 결과는 다음 표와 같다.

Table 4. Prediction

Test set / Train set	2017	2018
2017	0.97	0.81
2018	0.79	0.96

기존에 보유한 데이터를 가지고 앞으로 들어오는 데이터를 분류하기엔 낮은 성능으로 보이나, 이는 예측 모델의 개선으로 추후 극복해야할 문제다.

감사의 글

본 연구는 과학기술정보통신부 및 정보통신기획평가원의 대학ICT연구센터지원사업의 연구결과로 수행되었음 (IITP-2019-2017-0-01628)

참고문헌

1. Lakshmi, R. D., & Radha, N. , "Machine Learning Approach for Taxation Analysis using Classification Techniques.," International Journal of Computer Applications, 12(10). , 2011.
2. L.Breiman, Randomforests, MachineLearning45 (2001)5–32.
3. G.Biau, L.Devroye, G.Lugosi, Consistency of random forests and other averaging classifiers, Journal of Machine Learning Research 9 (2008) 2015–2033.
4. Verikas A, Gelzinis A, Bacauskiene M. Mining data with random forests: a survey and results of new tests. Pattern Recognit 2011, 44:330–349.
5. Sasaki Y 2007 The truth of the F-measure Teach. Tutor. Mater. 1–5.
6. T.K. Ho The random subspace method for constructing decision forests IEEE Transactions on Pattern Analysis and Machine Intelligence, 20 (8) (1998), pp. 832-844.
7. G. Nimrod, A. Szilagyi, C. Leslie, N. Ben-Tal Identification of DNA-binding proteins using structural, electrostatic and evolutionary features Journal of Molecular Biology, 387 (4) (2009), pp. 1040-1053.
8. H.T. Chen, T.L. Liu, C.S. Fuh Segmenting highly articulated video objects with weak-prior random forests A. Leonardis, H. Bischof, A. Pinz (Eds.), ECCV 2006, Part IV, Lecture Notes in Computer Science, vol. 3954, Springer-Verlag, Berlin, Heidelberg (2006), pp. 373-385.
9. J. Ham, Y. Chen, M.M. Crawford, J. Ghosh Investigation of the random forest framework for classification of hyperspectral data IEEE Transactions on Geoscience and Remote Sensing, 43 (3) (2005), pp. 492-501.
10. M.M. Crawford, J. Ham, Y. Chen, J. Ghosh Random forests of binary hierarchical classifiers for analysis of hyperspectral data 2003 IEEE Workshop on Advances in Techniques for Analysis of Remotely Sensed Data, IEEE, Greenbelt, MD, USA (2004), pp. 337-345.
11. J. Peters, B. De Baets, N.E.C. Verhoest, R. Samson, S.

Degroeve, P. De Becker, W. Huybrechts Random forests as a tool for ecohydrological distribution modelling Ecological Modelling, 207 (2007), pp. 304-318.
12. J. Peters, N.E.C. Verhoest, R. Samson, M. Van Meirvenne, L. Cockx, B. De Baets Uncertainty propagation in vegetation distribution models based on ensemble classifiers Ecological Modelling, 220 (2009), pp. 791-804.

접수일: 2019년 6월 17일, 심사일: 2019년 6월 22일, 게재확정일: 2019년 6월 24일