Journal of Korea Multimedia Society Vol. 19, No. 2, February 2016(pp. 209-214)
http://dx.doi.org/10.9717/kmms.2016.19.2.209

음성 특징에 따른 파킨슨병 분류를 위한 알고리즘 성능 비교

정 재 우†

Performance Comparison of Algorithm through Classification of Parkinson's Disease According to the Speech Feature

Jae Woo Chung†

ABSTRACT

The purpose of this study was to classify healty persons and Parkinson disease patients from the vocal characteristics of healty persons and the of Parkinson disease patients using Machine Learning algorithms. So, we compared the most widely used algorithms for Machine Learning such as J48 algorithm and REPTree algorithm. In order to evaluate the classification performance of the two algorithms, the results were compared with depending on vocal characteristics. The classification performance of depending on vocal characteristics show 88.72% and 84.62%. The test results showed that the J48 algorithms was superior to REPTree algorithms.

Key words: Data Mining, Weka, Machine Learning, Algorithm, Parkinson's disease

1. 서 론

데이터 마이닝은 "대량의 데이터에서 새롭고 유용한 지식을 창출하는 것"으로서, 데이터 더미에서 일반적인 사실을 의미하는 데이터가 아니라 의사 결정에 도움이 되는 유용한 정보를 포함한 지식을 추출하는 것이므로 '데이터에서 지식을 캐기(knowledge discovery from data: KDD)'라는 용어가 흔히 사용된다. 데이터 마이닝은 데이터 처리, 데이터 요약, 기계학습, 패턴인식, 시각화기술, 통계학, 지식추출기술 등 다양한 분야의 학제적 기술을 필요로 한다. 이러한 데이터 마이닝의 기술 중 기계학습(Machine Learning, 이하 ML)은 데이터 마이닝의 주된 기술적 기반으로서 데이터베이스의 원천 데이터로부터 정보를 뽑아내는 방법론을 제공한다[1]. 즉, 표본 데이터의 과거 경험을 토대로 컴퓨터가 최적의 성능을 갖도록 학습하는 일을 의미한다. 예를 들어 파라미터로 구성된 모델이 있다면, 훈련 데이터를 바탕으로 학습된 모델은 학습 과정에서 만나보지 못한 새로운 데이터로부터 결과를 예측할 수 있다[2].

본 논문에서는 이러한 ML을 기반으로 분류가 이루어지는 알고리즘에 대한 성능 비교를 진행하였다. ML 알고리즘 중에서도 이미 알려진 데이터를 이용하여 훈련하는 과정을 거치는 교사(Supervised) 알고리즘 J48과 REPTree의 성능을 비교하여 어떤 알고리즘이 더 정확한 분류를 제공하는지 확인하여 결과적으로 더 정확한 예측이 가능한 알고리즘을 찾는 것이 본 논문의 목표이다.

두 알고리즘의 성능 비교를 위해 데이터 parkinson.arff을 사용하였다. 파킨슨병 환자는 신체의 운동기

※ Corresponding Author : Jae Woo Chung, Address: (461-713) Eulji Univ., Yangji-dong, Sujeong-gu, Seongnam-si, Gyeonggi-do, Korea, TEL : +82-10-2817-5774, E-mail : wjdwo828@gmail.com

Receipt date : Jan. 22, 2016, Approval date : Jan. 28, 2016
† Dept. of Medical IT Marketing, Eulji University

능이 점진적으로 퇴행할 뿐만 아니라, 병이 진행될수록 심각한 구어장애를 나타내며, 일상적인 사회생활을 하는데도 많은 제약을 받게 된다[3]. 이러한 의학적 근거를 바탕으로, 음성분석기기를 이용하여 측정된 음성 특성과 파킨슨병의 유무에 대해 수집된 데이터를 이용하여 음성 특성에 따른 파킨슨병의 유무에 대한 분류를 진행하였다.

2. 관련연구

2.1 Machine Learning 알고리즘

ML 알고리즘은 비교사(unsupervised), 교사(supervised) 알고리즘으로 나누어서 생각해 볼 수 있다. 먼저 비교사 ML 알고리즘은 주어진 데이터들의 유사성을 기반으로 각 클래스별로 분류를 하며, 자주 사용하는 알고리즘으로는 K-Means[], DBSCAN, AutoClass, Expectation Maximization 등이 있다. 다음으로 교사 ML 알고리즘은 알려진 데이터를 이용하여 모델을 training 한 후, 그것을 기반으로 알려지지 않은 test 데이터를 분류하는 알고리즘을 말하며 자주 사용되는 알고리즘으로는 J48, REPTree, NaiveBayes, BayesNet, Naive Bayes Kernel Estimator 등이 있다[4][9].

본 논문에서는 교사 ML 알고리즘 중 J48 알고리즘과 REPTree 알고리즘을 사용한다.

2.1.1 J48

J48은 C4.5 의사결정 트리를 생성하기 위한 클래스 알고리즘이다[5]. C4.5 알고리즘은 1993년 Quinlan에 의해 제안되었는데, 이것은 ID3 알고리즘과 유사하지만, ID3 알고리즘의 몇 가지 단점(한계)들을 보완한 알고리즘이다[6].

ID3 알고리즘은 엔트로피(Entropy)의 개념을 사용하여 분리기준을 결정하는 반면, C4.5 알고리즘은 엔트로피뿐만 아니라 정보이득(Information Gain)의 개념을 사용하여 분리기준을 결정한다. 엔트로피(Entropy)는 주어진 데이터 집합의 혼잡도를 의미하고, 정보이득(Information Gain)은 어떤 속성을 선택함으로 인해서 데이터를 더 잘 구분하게 되는 것을 의미한다. 이러한 엔트로피(Entropy)와 정보 이득(Information Gain)의 계산식은 식 (1), (2), (3)과 같다.

$$Entropy(S) = -\sum_{i=1}^{m} p_i \log_2(p_i) \quad (1)$$

$$p_i = \frac{freq(C_i, S)}{|S|} \quad (2)$$

식 (1)은 주어진 데이터 세트 S에 대한 엔트로피 값을 구하는 계산식이다. 엔트로피 값은 각 클래스 값의 포함 비율에 로그를 적용하고 다시 값을 가중치로 곱한 값을 모두 더하는 식에 의해 계산된다. $\log_2$ 함수 적용을 통해 마이너스(-) 값이 나타나므로 전체 수식 값에 -를 붙여주어 0에서 1 사이의 값을 갖도록 한다. 가장 혼잡도가 높은 상태의 값이 1이며, 하나의 클래스로만 구성된 상태의 값이 0이다.

$$Gain(A) = I(s_1, s_2, \cdots, s_m) - E(\text{속성}A) \quad (3)$$

식 (3)에서 $I(S_1, S_2, \cdots, S_m)$는 상위 노드의 엔트로피를 의미한다. 그리고 $E(\text{속성}A)$는 A라는 속성을 선택했을 때 하위로 작은 m개의 노드를 나누어진다고 하면 하위 각 노드의 엔트로피를 계산한 후 노드의 속한 레코드의 개수를 가중치로 하여 엔트로피를 평균한 값이다. 원래 노드의 엔트로피를 구하여 A라 하고, 속성 A를 선택한 후의 m개의 하위노드로 나누어진 것에 대한 전체적인 엔트로피를 구하여 B라고 하면, 결국 $Gain(A)$의 의미는 A-B를 의미한다. 이 값이 클수록 정보 이득이 큰 것이고 해당 속성 A가 변별력이 좋다는 것을 의미한다[5].

2.1.2 REPTree

REP(Reduced-Error Pruning) Tree는 training set의 일부를 가지치기(pruning)하기 위하여 남겨두는 방법을 이용하여 생성된 tree를 말한다. 하지만 이 알고리즘은 J48과 달리 training set의 숫자가 줄어든다.

REPTree에서는 해당 노드를 가지치기하면 그 노드 밑의 subtree는 제거되며 노드 자신은 leaf 노드가 된다. 노드를 가지치기했을 때 tree의 성능이 그 이전 tree의 성능보다 나빠지지 않는 경우 노드의 가지치기를 결정한다. 이때, validation set을 사용하여 성능을 측정한다. validation set을 사용함으로써, training set을 통해 우연의 일치로 추가된 노드를 제거하는 효과를 기대할 수 있다. REPTree의 주요 기능은 Table 1과 같고[7], 또한 이러한 REPTree의 기능을 통해 나타나는 효과는 Fig. 1과 같다[8].

Table 1. REPTree Function

1. debug
2. maxDepth
3. minNum
4. minVarianceProp
5. noPruning
6. seed
7. numFolds

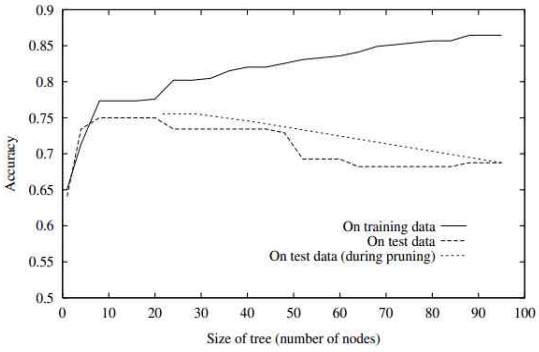

Fig. 1. Effect of REPTree.

3. 음성과 파킨슨병의 관계 분석

파킨슨병은 뇌의 흑질(substantia nigra)에 분포하는 도파민의 신경세포가 점차 소실되어 발생하며 안정 떨림, 경직, 운동 완만(운동느림) 및 자세 불안정성이 특징적으로 나타나는 신경계의 만성 진행성 퇴행성 질환이다. 파킨슨병 환자는 60세 이상에서 인구의 약 1% 정도로 추정된다.

파킨슨병 환자에 관한 연구들을 살펴보면, 구어의 점진적인 퇴행과 더불어 가족, 친구, 직장동료 등과의 의사소통에 어려움을 겪게 됨으로써, 심각한 삶의 질의 저하를 경험하게 된다[3].

본 논문에서는 음성분석기기인 MDVP (Multidimensional Voice Program)을 이용하여 측정된 환자의 음정 (Fo), 음질 (Jitter), Simmer (진폭변화율), RAP (Relative Average Perturbation, 상대평균변동률), PPQ (Pitch Perturbation Quotient, 음도변동률)와 파킨슨병의 유무의 관계가 나타난 데이터를 통해 J48 알고리즘과 REPTree 알고리즘의 분류 성능을 비교한다.

4. 실험

4.1 실험데이터

실험을 위한 도구로써 Waikato 대학교에서 개발된 WEKA v3.6.6을 사용하였고, 사용된 데이터는 MDVP를 이용하여 측정된 195명 환자의 음성과 파킨슨병의 유무에 관련된 사례가 수집된 'parkinson.arff'이다. parkinsons 데이터세트는 Oxford 대학교의 Max Little 교수와 언어 신호를 기록한 콜로라도 주의 음성 국립 센터의 공동 작업으로 만들어졌다.

데이터세트는 파킨슨병을 가진 23명을 포함한 31명의 사람들로부터 얻은 음성 측정으로 구성되어 있다. 테이블의 각 열은 특정한 음성 측정치, 각 행은 개인들로부터 기록된 목소리의 것들과 일치한다. 데이터의 주요 목표는 파킨슨병을 가진 사람들로부터 건강한 사람들을 식별하는 것이다.

세부적인 실험데이터의 구성은 다음과 같다. 환자의 음정(Fo), 음질(Jitter), Simmer(진폭변화율), RAP(상대평균변동률), PPQ(음도변동률), NHR(소음 대 배음 비율), HNR(배음 대 소음 비율)이 기록된 MDVP:Fo(Hz), MDVP:Fhi(Hz), MDVP:Flo(Hz), MDVP:Jitter(%), NHR, HNR, MDVP:RAP, MDVP:PPQ, MDVP:Shimmer(dB)와 같은 9가지 numeric 속성들과 파킨슨병의 유무에 따라 0, 1로 표기하는 numeric 속성인 status로 구성되어있다. 각각의 속성에 대한 세부적인 사항은 Table 2와 같다.

4.2 전처리 과정

1과 0으로 표기된 numeric 속성으로 표기되어 있던 status를 {yes, no}의 nominal 속성으로 이산화(discrete)하였다. 또한 수집된 데이터에 존재할 수 있는 결측치와 이상치는 인접한 사례의 중간 값을 취하거나 전반적인 평균값을 이용한 보정이 가능하기 때문에, 실험데이터 속성 중에서 MDVP:Fo(Hz), MDVP:Fhi(Hz), MDVP:Flo(Hz) 가운데 평균값이 표기된 MDVP:Fo(Hz)만을 사용하였다. 마지막으로 통계적인 결과에 영향이 적거나 안 좋은 영향을 주는 속성인 NHR, HNR과 같은 비관련 특징(irrelevant feature)은 제거하였다.

4.3 실험 결과

실험은 Parkinson 데이터를 기반으로 status 속성

Table 2. Value of Experimental Data

attribute	type	value
status	numeric	continuous from 0 to 1
MDVP:Fo(Hz)	numeric	continuous from 88.333 to 260.105
MDVP:Fhi(Hz)	numeric	continuous from 102.145 to 592.03
MDVP:Flo(Hz)	numeric	continuous from 65.476 to 239.17
MDVP:Jitter(%)	numeric	continuous from 0.002 to 0.033
NHR	numeric	continuous from 0.001 to 0.315
HNR	numeric	continuous from 8.41 to 33.047
MDVP:RAP	numeric	continuous from 0.001 to 0.021
MDVP:PPQ	numeric	continuous from 0.001 to 0.02
MDVP:Shimmer(dB)	numeric	continuous from 0.085 to 1.302

을 대상으로 J48과 REPTree를 사용하였으며, 데이터 분석에는 Cross-Validation의 fold 값을 10으로 수행하였다. Cross-Validation은 전체 집합을 k개로 나눈 뒤 각각 비교하여 전체적으로 특이한 집합이 없는지 확인하는 방식이다.

J48 알고리즘을 통해 나타난 tree는 Fig. 2와 같고, REPTree 알고리즘을 통해 나타난 tree는 Fig. 3과 같다.

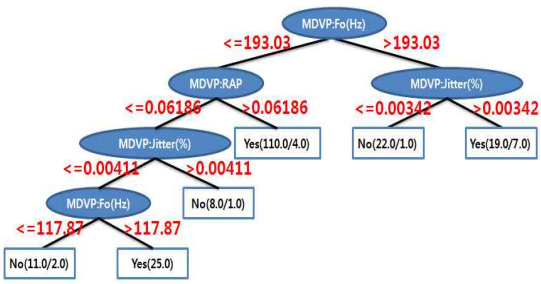

Fig. 2. J48 pruned tree.

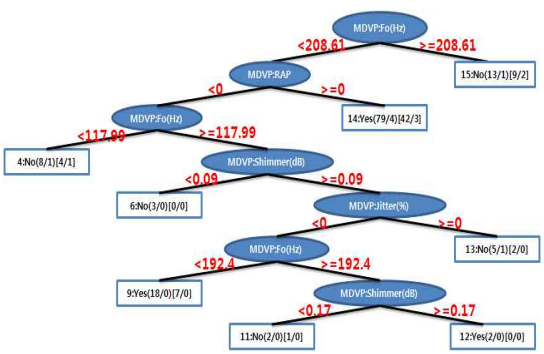

Fig. 3. REPTree.

다음의 Table 3와 표 Table 4는 각각 J48과 REP Tree 알고리즘을 적용하여 생성된 Confusion Matrix를 나타낸다. 세로축은 실제 값이고, 가로축은 분류된 값을 나타내므로, 가로축과 세로축의 값이 동일한 대각선의 경우가 올바르게 분류된 사례의 수이다.

J48과 REPTree 알고리즘으로 분류 실험을 통하여 검증한 전반적인 실험 결과를 모두 통합하여 Table 5, Table 6에 정리하여 나타내었다.

5. 결과 고찰

Table 5의 실험결과를 통하여, J48의 MAE(Mean Absolute Error)가 0.1566, 분류 정확도가 약 88.72%인 것에 비해 REPTree의 MAE가 0.197, 분류 정확도는 약 84.62%로 J48 알고리즘의 분류 정확도가 4% 가량 높은 것을 확인할 수 있었다. 또한 다른 시점에서 동일한 실험을 실행할 경우에 그 실험이 다시 유사하게 재현되는 지를 나타내는 재현율(reproduci-

Table 3. Confusion Matrix base on J48

a	b	classified as
138	9	a = yes
13	35	b = no

Table 4. Confusion Matrix base on REPTree

a	b	classified as
135	12	a = yes
18	30	b = no

Table 5. Result

	J48	REPTree
Correctly Classified Instances	173 (88.72 %)	165 (84.62 %)
Incorrectly Classified Instances	22 (11.28 %)	30 (15.38 %)
Kappa statistic	0.6872	0.5672
Mean absolute error	0.1566	0.197
Root mean squared error	0.3121	0.3404
Relative absolute error	42.0249 %	52.8704 %
Root relative squared error	72.4443 %	79.0147 %
Total Number of Instances	195	195

bility)의 평가 지표로 Kappa statistic을 분석하면 J48의 평균 Kappa는 0.6872, REPTree의 평균 Kappa는 0.5672로 0.10 가량 차이로 J48 알고리즘이 우수한 것을 확인할 수 있었다. 마지막으로 실험에서 나타나는 오차의 제곱을 평균한 값의 제곱근인 RMSE(Root Mean Squared Error)을 확인할 수 있었는데, 이는 통계학의 표준편차와 유사한 의미를 가진다. 실험 결과에서, J48을 적용한 경우에는 평균 RMSE가 0.3121, REPTree의 경우에는 평균 RMSE 0.3404을 보였다.

뿐만 아니라 Table 6에서는 실험결과를 통한 정밀도(Precision)와 재현율(Recall)이 나타나는데, J48의 정밀도(Precision)는 0.885, REPTree의 정밀도(Precision)는 0.841인 것을 확인하였다. 또한 J48의 재현율(Recall)은 0.887, REPTree의 재현율(Recall)은 0.846

Table 6. Accuracy Result

	J48	REPTree
TP Rate	0.887	0.846
FP Rate	0.219	0.303
Precision	0.885	0.841
Recall	0.887	0.846
F-Measure	0.885	0.843
ROC Area	0.879	0.89

Table 7. Accuracy, Precision, Recall and Kappa

	J48	REPTree
Accuracy	88.72%	84.62%
Precision	0.885	0.841
Recall	0.887	0.843
Kappa	0.6872	0.5672

인 것도 확인할 수 있었다.

위와 같은 실험 결과에 대한 분석을 통해 확인한 J48과 REPTree 알고리즘의 정확도, 정밀도, 재현율을 Table 7에 정리하여 나타내었다.

Table 7을 통해 모든 결과가 미세하지만 REP Tree 알고리즘보다 J48 알고리즘이 우수한 것을 확인할 수 있었다. 결국 Table 7은 J48 알고리즘이 REPTree 알고리즘보다 더 우수한 분류 성능을 가진다는 것을 나타낸다.

이 결과를 통해서 새로운 환자의 음성 데이터를 J48 알고리즘의 의사결정트리를 통해 파킨슨병에 대한 예측이 가능하다. 하지만 의료 분야는 높은 확률이 중요시되기 때문에, 90%가 넘는 정확도와 좀 더 우수한 정밀도와 재현율을 위해서 추후에는 실험데이터의 사례를 더 많이 수집하고, 이를 기반으로 다양한 알고리즘을 여러 관점에서 적용할 필요가 있다.

6. 결 론

본 논문에서는 교사 ML 알고리즘으로 적용이 가능한 J48과 REPTree 알고리즘의 성능을 확인하기 위해 동일한 데이터에 대하여 실험하였다. 음성분석 기기를 이용하여 측정된 음성 특성과 파킨슨병의 유무에 대해 수집된 Parkinson 실험데이터를 이용하였으며, 실험 결과를 기반으로 음성 특성에 따른 파킨슨병의 유무에 대한 분류 결과로 두 알고리즘의 성능을 확인하였다. 그 결과 J48의 분류 성공률은 약 88.72%, REPTree의 분류 성공률은 약 84.62%로 J48의 성공률이 약 4% 정도 우수한 것으로 나타났다. 뿐만 아니라 정밀도와 재현율에서도 미세하게 J48의 성능이 우수한 것을 확인하였다. 이를 통해 본 논문

의 목표이자 ML 알고리즘의 최종 목표인 새로운 데이터에 대한 좀 더 정확한 예측을 위해서는 J48 알고리즘을 사용하는 것이 더 우수한 성능을 가진다는 결과를 확인할 수 있었다.

이를 바탕으로 향후 파킨슨병뿐만 아니라 더 다양한 의료분야에 임상 데이터마이닝의 활용이 가능할 것으로 보인다.

REFERENCE

[1] S. Lee and R.W. Park, "Basic Concepts and Principles of Data Mining in Clinical Practice," *Journal of Korean Society of Medical Informatics*, Vol. 15, No. 2, pp. 175-189, 2009.

[2] S.C. Park, M.E. Lee, S.H. Kim, I.S. Na, and Y. Chen, "Machine Learning for Medical Image Analysis," *Korean Institute of Information Scientists and Engineers*, Vol. 39, No. 3, pp. 163-174, 2012.

[3] O. Lee, O. Ran, and D. Ko, "The Effects of Voice and Speech Intelligibility Improvements in Parkinson Disease by Training Loudness and Pitch," *Korean Society of Speech Sciences*, Vol. 8, No. 3, pp. 173-184, 2001.

[4] K. Jung, M. Choi, M. Kim, Y. Won, and J. Hong, "Internet Application Traffic Classification Using Machine Learning Algorithm," *The Committee on Korean Network Operations and Management Review*, Vol. 10, No. 2, pp. 39-52, 2007.

[5] S.Y. Kim, *A Hybrid On-Line Signature Verification System Based on Feature Selection Using Data Mining*, Master's Thesis of Inha University of Information Engineering, 2008.

[6] C4.5Algorithm, http://en.wikipedia.org/ wiki/C4.5_algorithm (accessed Dec, 22, 2015).

[7] S. Kalmegh, "Analysis of WEKA Data Mining Algorithm REPTree, Simple Cart and Random Tree for Classification of Indian News," *International Journal of Innovative Science & Technology,* Vol. 2, No. 2, pp. 438-446, 2015.

[8] Decision Tree Learning, http://jmvidal. cse.sc. edu/talks/decisiontrees/allslides.xml (accessed Dec, 28, 2015).

[9] F. Yi, I. Moon, "K-means based Clustering Method with a Fixed Number of Cluster Members," *Journal of Korea Multimedia Society,* Vol. 17, No. 10, pp. 1160-1170, 2014.

정 재 우

2015년 을지대학교 의료IT마케팅학과 학사

2016년~현재 성균관대학교 전자전기컴퓨터공학과 석사과정

관심분야 : 데이터베이스, 데이터마이닝