이형 데이터 기반의 예측 모델 생성을 위한 데이터 정제 방법론

이은경[1], 유연택[2],이건수[3*]
[1,3*](주)솔루게이트 기업부설연구소
[2]대보정보통신 복지사업단
e-mail: lek@solugate[1], ytyeon@dbcs.co.kr[2], keonsoo@solugate.com[3*]

A Practical Methodology of Preparing Data for Generating Prediction Model using Heterogeneous Data Sources

EunKyung Lee[1], Yeontaek Yoo[2], Keonsoo Lee[3*]
[1,3*]Research and Development Laboratory, Solugate Ltd.
[2]DAEBO Communication Systems Co.

요 약

예측 모델은 어떤 상황이 주어졌을 때, 다음 상황에 대한 예측을 수행하는 시스템으로, 현재 상황을 올바르게 인지하고, 그 인지된 상황을 토대로 미래를 예측할 수 있는 지능을 갖고 있어야 한다. 이러한 예측 모델이 올바르게 동작하기 위해서는 상황을 올바르게 인지하는 기능이 우선되어야 하지만, 원시 데이터로부터 상황을 인지하기 위해서는 원시 데이터를 올바르게 해석하기 위한 데이터 정제 과정이 필요하다. 이에 본 연구에서는 다양한 형태의 원시 데이터를 예측 모델의 유효한 입력 값으로 변환시키기 위한 데이터 정제 방법을 제안한다. 본 방법은 원시 데이터의 형태 정의, 데이터 정규화, 속성 관계 검증, 결측치 보정, 그리고 신뢰도 적용의 5단계로 구성되어 있다.

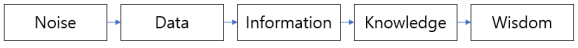
1. 서론

지능형 시스템은 처한 상황에서 스스로 판단하여, 주어진 목표를 달성하는 시스템을 의미한다 [1]. 이러한 동작을 수행하기 위해서, 시스템은 우선 자신의 상황을 인지할 수 있어야 하고, 주어진 목표를 수행하기 위해, 현재 상황에서 어떤 행동을 해야 할 것인지를 판단할 수 있어야 한다. 기존의 전문가 시스템(Expert System)이라고 불리던 초창기 지능형 시스템의 경우, 주어진 상황은 사전에 정의되어 있었으며, 어떤 행동을 선택할 것인지에 대한 규칙들 역시 사전에 정의되어 있다 [2]. 하지만, 규칙을 이용하여 상황을 정의한다는 가정 자체가 프레임 문제(Frame Problem)를 갖고 있기 때문에, 실제 문제에 직접 적용하기에는 어려움이 있다 [3]. 이 문제를 해결하기 위해 다양한 방법들이 연구되었고, 현재는 “불완전한 상황 정보를 토대로 가능성 높은 선택을 하는 것”이 지능형 시스템을 구현하는 기본 목표가 되었다 [4, 5].

상황을 인식함에 있어서 불완전함을 인정하는 것은, 그 상황을 기반으로 내리는 선택 역시 오류의 가능성을 포함하고 있는 것이며, 이 오류의 가능성을 낮추기 위해서는 상황을 인식하기 위한 원시 데이터(Raw Data)를 가능한 정확하게 획득하고 해석할 수 있어야 한다. 본 연구는 지능형 시스템, 즉 예측 모델을 생성하기 위해, 원시 데이터를 정제하는 방법론을 제안한다.

2. 연구 배경

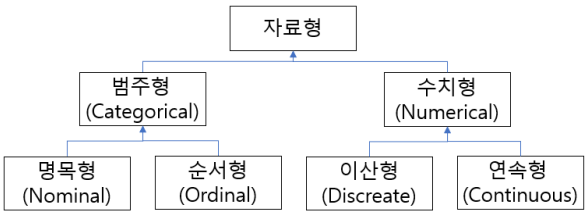
데이터에 대한 정의는 그 데이터를 사용하는 목적, 분야, 데이터의 내용에 따라 달라지지만, 일반적으로는 그림 1과 같은 계층 구조를 갖고 있다 [6]. 사용 의도에 부합되지 않는 기호는 모두 잡음(Noise)들이지만, 의도에 부합되는 내용이 있다면, 그때는 데이터(Data)가 된다. 데이터의 집합으로부터 어떤 패턴이나 새로운 의미가 발견되면, 그것을 정보(Information)이라고 하고, 정보가 쌓이면, 그 분야에서의 지식(Knowledge)를 얻게 된다. 일반적으로 지식은 정보를 체계화 시킨 것으로 정의되고, 이러한 지식은 이해(Understanding)과정을 통해서 지혜(Wisdom)가 된다. 일반적으로 지식을 노하우(know-how)라고 하며, 지혜를 원리를 규명하는 것(know-why)이라고 한다.



(그림 1) 데이터 계층도

본 연구의 주제인 데이터 정제 방법론은 데이터로부터 정보를 추론하기 위해, 데이터를 전처리하는 방법이다. 데이터는 그 데이터가 담고 있는 내용의 형식을 기반으로 다음과 같이 분류할 수 있다. 수치형 데이터는 정수로 표현되는 이산형 데이터와 주어진 범위의 실수로 표현되는 연속형 데이터로 분류될 수 있고, 범주형 데이터의 경우,

각 범주가 순서를 갖는 순서형 데이터와 범주들 사이의 관계가 없는 명목형 데이터로 분류될 수 있다.



(그림 2) 형식에 따른 데이터의 분류

3. 제안 방법

본 연구는 이형 데이터 집합을 이용해 하나의 예측 모델을 만드는 과정에서 데이터의 속성을 정제하기 위한 실천적 방법론을 제안한다. 제안 방법은 다음의 다섯 단계로 구성된다.

첫 번째 단계는 데이터의 타입을 분류하는 것이다. 데이터의 타입은 그 속성의 값을 어떻게 받을 것인지를 결정하는 것으로, 원시 데이터를 그대로 사용할 수도 있고, 원시데이터에 대한 해석을 통해 사용할 수 있다. 가령, 성적을 받을 때, 원 점수로 받을 것인지, 등급으로 받을 것인지, 등수로 받을 것인지를 결정하는 과정이며, 해석의 과정을 통해 원시 데이터가 대체되기 때문에, 이후 다른 해석의 가능성을 남겨두기 위해서는, 원시 데이터를 남겨두고, 해석 데이터를 신규 속성으로 추가하는 방법을 사용할 수 있다.

두 번째 단계는 데이터의 값을 정규화 시키는 것이다. 데이터 값의 종류에 따라 정규화 시키는 목적이 달라지고, 목적에 따라 방법 역시 달라진다. 수치형 데이터의 경우, 정규화는 각 데이터를 서로 비교하기 위해 사용된다. 가령, 국어와 수학 시험 점수로 성적을 준다고 할 때, 국어는 100점 만점, 수학은 200점 만점이라면, 단순 시험 결과를 갖고 비교하면, 수학 점수가 2배의 가중치를 갖고 최종 성적에 반영될 것이다. 이를 방지하기 위해, 각 점수를 0부터 1의 범위로 재조정하여, 두 속성을 동일 가중치로 비교할 수 있게 되며, 대표적인 정규화 공식으로는 [(X-최소값)/(최대값-최소값)]을 사용한다. 동일 속성의 값들을 비교하기 위한 방법으로는 표준점수(Z-score)를 사용할 수 있다. 국어를 50점, 수학을 190점을 맞았다고 할 때, 정규화를 수행하면, 국어는 0.5점, 수학은 0.95점이 되어, 국어보다 수학을 잘한다는 것을 알 수 있지만, 만약, 국어의 평균이 20점이고, 수학은 평균이 180점이라고 하면, 수학보다는 국어를 잘 한다고 할 수 있을 것이다. 표준점수는 [(X-평균)/표준편차]로 구할 수 있으며, 두 번째 과정에서는 표준화 이후 정규화를 통해, 수치형 데이터를 정제할 수 있다. 한편 범주형 데이터의 경우, 각 범주의 의미 수준을 통일한다. 가령, "주택피해"라는 속성을 사용한다고 할 때, 그 값의 범주를 "유실, 매몰, 침수, 소파(小破), 반파(半破), 전파(全破)" 라고 한다면, 의미 수준이 다른 값이 동일 속성의 값으로 사용되고 있는 것이기에, 속성을 2개("피해의 원인"과 "피해의 정도")로 분리해야 한다.

세 번째 단계는 속성들 사이의 의미 관계를 분석하는 것이다. 속성들 사이의 의미 관계를 분석하는 이유는 예측 과정에서 사용 속성이 많다고 결과가 무조건 좋아지지 않으며, 사용 속성이 적다고 결과가 나빠지는 것이 아니기 때문이다. 오히려 속성이 많으면 계산 복잡도가 높아지고, 불필요한 속성이 사용되는 경우, 예측 잡음으로 동작해, 예측 결과를 떨어트리게 된다. 속성이 적으면 계산 복잡도는 높아지지만, 예측을 위한 데이터가 충분하지 않아서, 예측 성능이 나빠지게 된다. 결국 적절한 수의 속성을 사용하는 것이 중요하며, 각각의 속성은 결과를 예측하는데 유의미한 정보를 제공할 수 있어야 하고, 결과와 연관되어 있는 속성들이라면, 그 속성들 사이의 연관성도 존재할 가능성이 있다. 속성들 사이의 연관성이 존재한다면, 연관된 속성의 값이 결측되었을 때, 그 결측된 값을 추론하기 위한 연관 규칙으로 사용할 수 있다. 이 부분은 네 번째 단계에서 보다 상세히 설명하도록 한다. 우선 각각의 속성은 예측 결과와 유의미한 관계를 갖고 있어야 하며, 이를 확인하기 위해서, 속성과 결과의 이항관계를 확인한다. 수치형 데이터의 경우, 상관 계수 분석을 통해서 상관성을 확인할 수 있고, 범주형 데이터의 경우, Apriori 등의 연관규칙분석을 통해 확인할 수 있다. 예측 결과와 상관/연관성을 보이는 속성들을 대상으로 상관/연관 관계를 분석하며, 이때 속성의 수에 따라서, 이항관계뿐 아니라 다항관계에서의 상호 관계를 검증한다. 하지만, 이 경우, n개의 속성에 대해서 2^n의 경우에 대한 분석이 이뤄져야 한다는 단점이 존재하며, 이를 막기 위해, 적당한 휴리스틱 규칙을 적용할 수 있다.

네 번째 단계는 결측치에 대한 보정단계다. 선택을 하는 순간에 필요한 상황 정보는 완벽하지 않은 상태이기 때문에, 비어있는 값들을 어떻게 처리할 것인지에 대한 결정을 내려야 한다. 가장 간단한 방법으로는 불완전한 데이터는 폐기하는 것이지만, 데이터가 충분한 경우에는 가능하지만, 데이터가 충분하지 않은 경우, 모든 결측치가 있는 데이터를 폐기할 때, 실제 사용할 수 있는 데이터가 너무 적어지기 때문에, 비어있는 값을 유추하기 위한 방법을 사용해야 한다. 결측치를 보정하기 위한 과정은 두 단계로 구분할 수 있다. 첫 번째 단계에서는 우선, 그 결측값이 보정이 필요한 값인지에 대한 판단을 수행한다. 가령, "유선전화번호" 속성이 있을 때, 이 값이 비어있다고 해서, 보정이 필요하다고 할 수는 없다. 실제로 유선 번호를 갖고 있지 않을 수 있기 때문에, 보정을 통해서 이 속성값이 채워지면, 그것은 거짓된 값이기 때문이다. 반면, "월별 수입" 속성에서 특정 월의 값이 빠져있다면, 이 값은 보정일 필요할 수도, 필요하지 않을 수도 있다. 만약, 그 달에 무직 상태여서 수입이 없었다면, 이 값을 보정하는 것은 거

짓된 값을 넣는 것이고, 만약, 실제로 값이 빠져있어서, 이전/후의 값의 평균 혹은 작년도 동일 월의 값으로 보정 가능하다면, 보정해야 하는 값이 된다. 보정 과정의 필요성이 확정되면, 두 번째 단계는 값을 보정하는 과정이 된다. 수치형 데이터의 경우, 해당 속성의 평균값, 중간값, 최빈값, 혹은 기 정의된 초기값을 넣는 방법이 있으며, 내삽 혹은 외삽 방법을 사용해서 값을 보정할 수 있다. 범주형 데이터의 경우, 최빈값 혹은 초기값을 사용해 보정할 수 있다. 데이터 타입의 상관없이, 다른 의미적으로 관련된 속성의 값을 참조해서 보정을 수행할 수 있다. 세 번째 단계의 결과로 유관 속성을 알고 있다면, 회귀분석을 통해, 유관 속성 값을 기준으로 결측값을 보정할 수 있다.

다섯 번째 단계는 속성값의 신뢰도를 결정하는 것이다. 일반적으로 획득한 속성의 정보는 참의 진리값을 갖고 있다고 가정되지만, 속성 자체의 신뢰도를 의심해야 하는 가정이 존재하는 경우가 있다. 가령, 1개월에 한번 씩 갱신되는 자료의 경우, 생성된 지 3주가 지난 정보는 당시 생성된 정보에 비해 낮은 신뢰도를 갖게 된다. 만약 그 속성이 "성별"처럼 시간의 흐름에 영향을 받지 않는 고정된 값이라면, 신뢰도의 변화가 없을 것이나, 시간에 따른 가변성을 가진 속성이라면, 상황 변화에 따른 신뢰도 변경 방법을 적용해야 한다 [7]. 또한 네 번째 과정을 통해 보정된 값은 실제 측정된 값에 비해 낮은 신뢰도를 갖고 있어야 하며, 특정 속성 자체의 값의 결측률이 높다면, 그 속성 자체를 신뢰할 수 없을 수도 있다. 신뢰도의 활용은 해당 값의 가중치 적용을 통해 사용할 수 있으며, 예측 모델이 회귀분석을 사용한다고 하면, [속성값 * 속성 가중치 * 신뢰도]의 형태로 표기 가능하다.

4. 예측 모델의 평가

예측 모델의 평가는 예측 정확도를 기반으로 수행되며, 정확도를 계산하기 위한 기본 방법은 혼돈 행렬(Confusion Matrix)를 사용하는 것이다. 혼돈 행렬은 4개의 하위 요소들로 구성되며, 각 요소들은 TP(True Positive), FP(False Positive), TN(True Negative), FN(False Negative)이다. 특히 FP는 "1종 오류(Type 1 Error; 귀무가설 기각 오류)"라고 불린다. 대립가설이 기각되어야 하는데, 채택이 되어서 발생한 오류다. .FN은 "2종 오류(Type 2 Error; 귀무가설 채택 오류)"라고 불린다. 대립가설을 채택해야 하는데, 기각했기 때문에 발생한 오류다. FN과 FP를 구분하는 까닭은 예측 결과에 대한 확신이 없을 때, 전략적으로 오류의 종류를 선택할 수 있기 때문이다. 가령, 암을 예측함에 있어서, 정상을 암으로 오진하는 경우와, 암을 정상이라고 오진하는 경우를 비교하면, 후자의 경우 예측 실패의 비용이 더 크기 때문에, 이런 경우 1종 오류를 전략적으로 선택할 수 있다. 혼돈 행렬로부터 획득 가능한 대표적인 5개의 속성과 각각의 계산식은 표 1에 기술되어 있다.

<표 1> 혼돈행렬을 이용한 정확도 계산식

Recall	TP/(TP+FN)
Precision	TP/(TP+FP)
Specificity	TN/(TN+FP)
NPV	TN/(TN+FN)
F-measure	(2*(Precision*Recall))/(Precision + Recall)

예측 모델의 정확도를 계산함에 있어서, 일반적으로는 정답을 맞추는 것이 중요한 경우, Recall과 Precision을 주로 보고, 이 둘의 조화평균인 F-measure 값을 사용한다. 하지만, FP와 FN을 모두 낮출 수 있는 이상적인 상황이 아닌 경우, 또한 정답과 오답 중 어느 것을 맞추는 것이 중요한가에 따라서, 차선책을 결정해야 하는 경우가 있다. 이때, 어떤 속성을 포기할 것인지를 판단하는 것은, 대상 분야의 전문가의 책임이 수반된 판단이 된다.

최종 예측 결과가 기준에 미치지 못하는 경우는 크게 두 가지 원인을 꼽을 수 있다. 첫 번째 원인은, 데이터를 활용하는 학습 방법이 데이터와 맞지 않기 때문일 수 있다. 기계 학습을 위한 다양한 방법들이 있으며, 각 방법은 나름의 장점과 단점을 갖고 있다. 획득한 데이터에 적합하지 않은 기계 학습 방법을 사용한다면, 데이터가 아무리 잘 정제되어 있다고 해도, 좋은 결과를 낼 수 없을 것이다. 두 번째 원인은 데이터가 제대로 정제되지 않았기 때문이다. 첫 번째 단계에서 데이터의 타입 정의가 잘못 되었을 수 있고, 세 번째 단계에서, 잘못된 연관 관계가 만들어졌을 수 있고, 네 번째 단계에서 결측치를 잘못 보정되었을 수 있으며, 다섯 번째 단계에서 신뢰할 수 없는 값을 높은 가중치로 사용했을 수 있다. 두 번째 단계인 정규화 과정에서는 결과에 영향을 미칠 정도의 오류가 발생할 가능성이 낮다. 적용한 기계학습의 방법이 적합하다면, 제안한 데이터 정제 방법을 역으로 수정해가면서 재적용해 최종 결과를 개선할 수 있다. 역으로 적용하는 까닭은 제안 방법은 각 단계가 이전 단계의 결과를 기준으로 적용되기에, 마지막 단계부터 수정하는 것이 오류 수정 효율을 높일 수 있기 때문이다.

5. 결론

구술이 서말이어도 꿰어야 보배라는 말은 데이터 과학 분에에서 많이 사용되는 격언이지만, 이 격언이 적용되기 위해서는 구술을 준비해야 하는 과정이 선결되어야 한다. 깨진 구슬이거나, 구슬에 돌이 섞여 있다면, 아무리 숙련된 전문가가 꿴다고 해도, 보배를 만들어낼 수 없다. 본 연구에서는 구슬을 수집하기 위한 과정, 즉 데이터 정제 과정에서 적용할 수 있는 다선 단계로 구성된 실천 방법론을 제안하였고, 이후 데이터를 이용한 예측 결과를 토대로 정제 결과를 평가하고, 데이터를 재사용하기 위한 방법을 제안하였다.

Acknowledgement

본 연구는 문화체육관과부 및 한국콘텐츠진흥원의 2019년도 문화기술연구개발 지원사업으로 수행되었음

참고문헌

[1] Negnevitsky, M. "Artificial intelligence: a guide to intelligent systems" Pearson education. 2005.

[2] Abraham, A. “Rule-Based Expert Systems”, in Handbook of Measuring System Design, American Cancer Society. 2005.

[3] Harnad, S. "Problems, problems: the frame problem as a symptom of the symbol grounding problem". Psycoloquy, 4(34). 1993.

[4] Hammami, R., Temponi, C., and Frein, Y. "A scenario-based stochastic model for supplier selection in global context with multiple buyers, currency fluctuation uncertainties, and price discounts". European Journal of Operational Research, 233(1), 159-170. 2014.

[5] Truong, B. A., Lee, Y. K., and Lee, S. Y. "Modeling and reasoning about uncertainty in context-aware systems". In IEEE International Conference on e-Business Engineering (ICEBE'05) (pp. 102-109). IEEE. 2005.

[6] Jifa, G. “Data, Information, Knowledge, Wisdom and Meta-Synthesis of Wisdom-Comment on Wisdom Global and Wisdom Cities”, Procedia Computer Science, vol 17, pp 713 - 719, 1 2013.

[7] Dubois, D., Lang, J., and Prade, H. "A possibilistic assumption-based truth maintenance system with uncertain justifications, and its application to belief revision". In Truth Maintenance Workshop (pp. 87-106). Springer, Berlin, Heidelberg. 1990.