Proceedings of KFIS Autumn Conference 2006 Vol. 16, No. 2.

k-NN 기법을 이용한 학습자 데이터의 노이즈 선별 방법

Noise-Reduction of Student's Learning Data using k-NN Method

윤태복[1], 이지형[2], 정영모[3], 차현진[4], 박선희[5], 김용세[6]

[13456] 성균관대학교 창의적 설계추론 지적 교육 시스템 연구단
E-mail: (tbyoon, ewmhb, lois6934, seonhp2, yskim)@skku.edu
[2] 성균관대학교 정보통신공학부
E-mail: jhlee@ece.skku.ac.kr

요 약

사용자 모델링을 위해서는 사용자의 성향 및 행위 등의 다양한 정보를 수집하여 분석에 이용한다. 하지만 사용자(인간)로 부터 얻은 데이터는 기계나 환경에서 수집된 데이터 보다 패턴을 찾기 힘들어 모델링하기 어렵다. 그 이유는 사용자는 사용자의 현재 상태와 상황에 따라 다양한 결과를 보이며, 일관성을 유지 하지 않는 경우가 있기 때문이다. 사용자 모델링을 위해서는 분산되어 있는 데이터에서 노이즈를 선별하고 연관성 있는 데이터를 분류할 수 있는 기술이 필요하다. 본 논문은 사용자로 부터 수집된 데이터를 k-NN(Nearest Neighbor) 기법을 이용하여 노이즈를 선별한다. 노이즈가 제거된 데이터는 의사결정나무(Decision Tree)방법을 이용하여 학습하였고, 노이즈가 분류되기 전과 비교 분석 하였다. 실험에서는 홈 인테리어 학습 컨텐츠인 DOLLS-HI를 이용하여 수집된 학습자의 데이터를 이용하였고, 생성된 학습자 모델링의 신뢰도가 높아지는 것을 확인하였다.

Key Words : ITS, Noise, k-NN Method

1. 서 론

주어진 데이터를 분석하여 의미 있는 정보를 추출하기 위하여, 전처리 과정은 매우 중요하게 여겨진다. 특히 노이즈 데이터의 분류/제거 작업은 분석결과의 신뢰성과 연관되는 결정적 요인이기 때문에 전체 분석 작업 중에 많은 시간을 소요하며, 우선시 되는 작업 중에 하나이다. 노이즈를 선별하기 위한 방법에는 노이즈가 들어간 데이터를 이용한 학습알고리즘들의 성능비교에 대한 연구[1]와 클러스터링에서 노이즈를 제거하기 위한 방법[2]이 소개 되었으며, 잡음패턴에서 음질 개선을 위해 노이즈 선별 방법[3]이 소개되었으나, 주로 노이즈 선별을 위한 연구는 음원, 전파 등과 같은 시스템에서 얻은 데이터를 기반으로 주로 실시되었다. 반면, 다양성과 애매함을 내포한 학습자(인간)의 데이터를 분류하기 위한 연구는 찾아보기 힘들다. 본 논문은 k-NN 기법을 이용하여, 학습자의 학습 과정에서 수집된 데이터에 섞여 있는 노이즈를 선별 할 수 있는 방법을 소개한다.

본 논문의 구성은 다음과 같다. 2장에서는 노이즈 선별에 사용될 학습자 데이터를 수집하기 위한 환경을 설명하고, 3장에서는 제안하는 방법인 k-NN 기법을 이용한 노이즈 선별 방법을 소개한다. 4장에서는 실험을 통하여 유효성을 확인하고, 끝으로 5장에서는 결론과 향후 연구로 맺는다.

2. 학습자의 데이터 분석을 통한 학습자 모델링에 관련한 연구

2.1 학습자의 성향 파악을 위한 연구

학습자는 학습 과정에서 정보를 받아들이고 이해하는 방식에서 다양한 모습을 보여주고 있다. 예를 들면 문자로 설명된 내용 보다는 그림으로 설명된 학습 컨텐츠를 더 선호하거나, 학습 과정에 있어서 순서대로 학습하는 것보다 순서에 상관없이 자신이 원하는 정보를 자유롭게 찾아보면서 학습하는 것을 더 선호하는 학습자가 있을 것이다. Felder & Silverman[5]은 앞의 예에서와 같이 학습 정보를 이해하는 차

원에서 Global 과 Sequential, 정보를 습득하는 차원에서 Visual과 Auditory, 정보를 인지하는 차원에서 Sensing과 Intuitive, 그리고 정보를 활용하는 차원에서 Active와 Reflective로 네 가지 영역에서 학습 성향을 분류하였다. 표1은 학습 성향에 대한 설명이다.

표 1. 학습 성향에 따른 설명

학습 성향		설명
정보이해	Global	Global 학습자는 부분적으로 보고 이해하지 못하며 전체 학습의 큰 그림이나 개요 등을 통해 더 잘 이해하는 경향이 있다.
	Sequential	Sequential 학습자는 세부내용을 참을성 있게 학습하며, 표준적으로 정해진 방법을 통해 더 잘 이해하는 경향이 있다.
정보습득	Visual	Visual 학습자는 그림, 차트, 영화, 데모 등으로 본 것을 잘 기억하며, 글이나 말로 하는 설명보다는 실습 동영상을 통해 더 잘 이해하는 경향이 있다.
	Auditory	Auditory 학습자는 학습을 통해 들은 것, 말한 것을 좀 더 잘 기억하고 토론을 통해 많은 것을 얻는 경향이 있다.
정보인지	Sensing	Sensing 학습자는 학습의 세부내용을 주의 깊게 공부하며 구체적인 정보로 구성된 학습 자료를 활용할 때 효과적인 경향이 있다.
	Intuitive	Intuitive 학습자는 어떤 상징화된 것을 다루는데 익숙하며 추상화된 개념을 해석하는 것을 잘하는 경향이 있다.
정보활용	Active	Active 학습자는 토론하고 설명하고 테스트하는 등의 실험적인 성향을 가지고 있다.
	Reflective	Reflective 학습자는 습득한 지식과 정보를 시험해보고, 처리하는 성향을 가지며, 혼자 또는 다른 사람과 짝을 지어 공부할 때 효과적이다.

2.2 학습 스타일에 따른 인터페이스 구성

Global & Sequential : 정보를 이해할 때 학습자의 선호에 의해 교육 컨텐츠를 선택하는지, 교육 전문가가 의도 하는 학습 순서에 따라 학습하는지를 알아보기 위하여 개요보기 버튼, 목차를 통한 이동 버튼과 항목 이동 화살표 등으로 인터페이스를 설계하였다.

Visual & Auditory : 정보를 습득할 때 그림 위주의 설명을 선호하는지 텍스트 위주의 설명을 선호하는지를 알아보기 위하여 그림위주의 학습 설명 버튼과 텍스트 위주의 학습 설명 버튼을 인터페이스에 설계하였다.

Sensing & Intuitive : 정보를 인지할 때 주위 깊게 문제를 풀어나가는지 직관적으로 문제를 풀어나가는지를 알아보기 위하여 퀴즈 풀기와 부가 학습 및 학습에 소요되는 시간을 측정할 수 있도록 설계하였다.

Active & Reflective : 정보를 활용 할 때 적극적인지 수동적인지를 알아보기 위하여 선생님께 질문하기, 학습 게시판에 의견 달기, 의견 보기 등의 기능을 설계하였다.

그림 1은 학습 스타일 관점에 따른 홈 인테리어 학습 컨텐츠인 DOLLS-HI(Diagnosis of Learner's Learning Styles - Housing Interior)[4]에 인터페이스의 일부분이다.

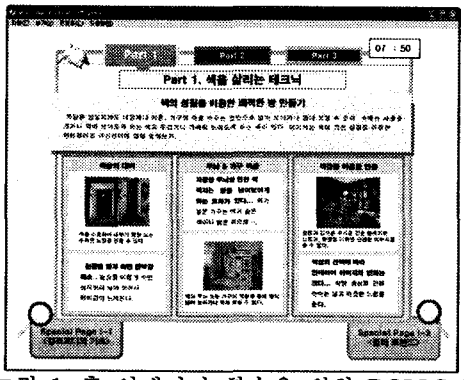

그림 1. 홈 인테리어 학습을 위한 DOLLS-HI

홈 인테리어 학습을 위한 DOLLS-HI은 학습과정에서 수집된 버튼 클릭 정보, 시간 정보, 퀴즈 풀이과정에서의 정답률 등을 수집하고 분석하여 학습자의 성향을 알아내는데 이용된다.

3. k-NN 기법을 이용한 학습자 데이터에서 노이즈의 분류

본 논문에서 제안하는 노이즈 선별 방법은 k-NN 기법에 기반을 두고 있다. 주어진 데이터를 공간상에 사상시키고, 한 개체와 다른 개체들 사이의 분포 정도를 계산한다. 주변 분포 정도를 측정하여 노이즈 인지 여부를 판단하는 것이다.

3.1 k-NN 기법을 이용한 노이즈 선별 방법

k-NN 기법을 이용한 노이즈 선별 방법은 각 인스턴스 간에 거리를 기반으로 계산한다. 예를 들어 그림 2와 같이 10개의 인스턴스가 있다고 가정하자 각 인스턴스가 가지고 있는 속성은 x와 y 두개이며 정수형의 값을 갖는다. 각 속성의 종속변수인 Class는 Yes와 NO 두 가지 값을 가질 수 있다. 그림 2는 10개의 인스턴스를 좌표상에 표현한 것이다.

그림에서 보면 알 수 있듯이 10번째 인스턴스의 Class를 알 수 없다. 그렇다면 여기서 10번째 Class를 무엇이라 할 수 있는가? 아마도 No라고 대답하는 사람이 대다수 일 것이다. 그 이유는 현재 알지 못하는 10번째 인스턴스의 주변 분포가 Yes보다는 No가 더 많기 때문이다. 만약 주어진 데이터에 10번째 인스턴스의 Class가 Yes라고 주어졌다고 가정하자.

	x	y	class
1	5	5	Yes
2	6	17	No
3	8	9	No
4	9	10	No
5	9	14	Yes
6	11	9	No
7	12	8	No
8	13	13	Yes
9	15	3	Yes
10	9	8	?

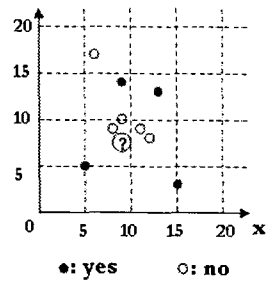

그림 2. 수집데이터의 공간상의 표현

주변 분포를 볼 때 No가 더 많으므로, 10번째 데이터를 노이즈로 결정할 수 있을 것이다. 이처럼 주어진 전체 데이터에서 한 인스턴스의 주변 분포를 측정하여 자기 자신의 노이즈 여부를 판단하는 것이 k-NN 기법을 이용한 노이즈 선별 방법이다.

3.2 k-NN 기법 노이즈 선별을 위한 고려사항

분표 측정을 위한 범위 설정 - 한 인스턴스에서 주변의 분포를 측정하기 위해서는 범위 설정이 필요하다. 여기서 범위는 노이즈 여부를 판단하기 위한 인스턴스로부터 일정 거리를 기준으로 할 것인지(그림 3 우), 근접한 개수를 기준으로 할 것인지(그림 3 좌)에 대한 여부이다. 두 가지 방법은 실제 적용을 통하여 비교 사용하여야 한다.

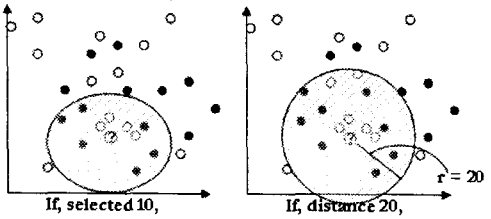

그림 3. (좌) 기준 인스턴스에서 가장 근접한 10개를 선정하여 판단 (우) 기준 인스턴스로부터 거리 20 만큼의 범위를 선정하여 판단

분포내 인스턴스와의 거리 비율 - 범위가 설정되고 나면 범위내의 인스턴스 개수를 파악하여 노이즈 여부를 판단하게 된다. 하지만 단순히 개수만을 비교한다면 근접거리에 존재하는 인스턴스의 의미가 무시될 수 있다.

예를 들어, 그림 4는 근접한 10개의 인스턴스를 선택하였다. 개수만을 비교한다면 "Black"가 되어야 하지만, 원안의 거리비율에 따른 분포를 본다면 "White"에 가깝다고 할 수 있다. 선정된 범위 안에서 비교가 되는 인스턴스의 거리에 대한 가중치가 부여 된다면 그림 4의 (좌) 결과에서 그림 4 (우)결과로 바뀔 것이다.

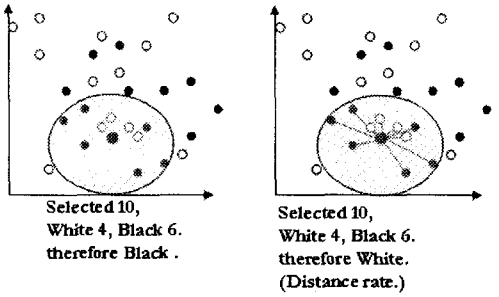

그림 4. (좌) 범위 내에서 단순히 개수만 비교하여 판단 (우) 범위 내에서 거리 비율을 계산하여 판단

전체 데이터의 분포에 따른 비율 - 제안하는 방법은 노이즈 판단을 위한 인스턴스 주변의 분포를 측정하여 결정한다. 다시 말하면 분포의 측정이란 내 주변에 나와 비슷한 인스턴스가 얼마나 존재하는가 하는 여부이다. 주변의 범위를 설정하고 범위내 인스턴스들의 거리 비율에 따라 개수를 파악한다고 하더라도 초기에 주어진 데이터의 비율이 편중 되어 있다면, 그 부분 또한 연산에 고려되어야 할 것이다. 예를 들어 그림5 에서 보면 노이즈 판단을 하고자 하는 데이터의 주변 10개 분포를 비교하면 "White"가 6개이고 "Black"이 4개 이므로 "White"라고 판단해야 한다. 하지만, 초기에 주어진 전체 데이터의 개수가 "White"에 편중되어 있으므로 전체 데이터의 비율을 고려하여 "Black"으로 판단된다.

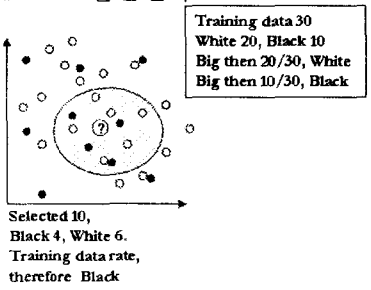

그림 5. 전체 데이터의 비율을 고려한 노이즈 선별

3.3 노이즈 선별을 이용한 학습자 진단 과정

학습 성향을 분류하기 위한 전체 프로세스는 Felder & Silverman이 제시한 학습 스타일을 기반으로 학습 성향을 미리 알아보기 위한 ILS (Index of Learning Styles) Questionnaire 프로세스와 학습 성향 데이터를 수집하기 위한 홈 인테리어 학습용 교육 컨텐츠 인터페이스 학습 프로세스로 구성되어 있다. 우선 학습자는 ILS 온라인 설문에 참여하여, 학습 성향별로 선호도를 알아본다. 이 후 학습자는 학습자 성향 수집을 위한 홈 인테리어 학습용 컨텐츠를 학습하며, 이 인터페이스에서 제공하는 홈 인테리어 학습, 퀴즈 풀기, 인테리어 배치 경험하기 등을 수행한다. 이 때 학습자의 데이터(시간, 이동, 학습을 위한 버

튼의 클릭 등)는 XML 형태의 파일로 기록된다. 이후 전체 피 실험자의 XML 데이터가 수집이 되면, 전처리 과정을 거쳐 제안하는 방법인 k-NN 기법을 이용하여 노이즈를 제거한다. 노이즈가 제거된 데이터는 의사결정 트리 방법을 이용하여 학습시키고, 생성된 결과는 학습자 진단을 위한 기반 정보로 사용한다. 그림 6은 위에서 설명한 작업흐름을 표현한 것이다.

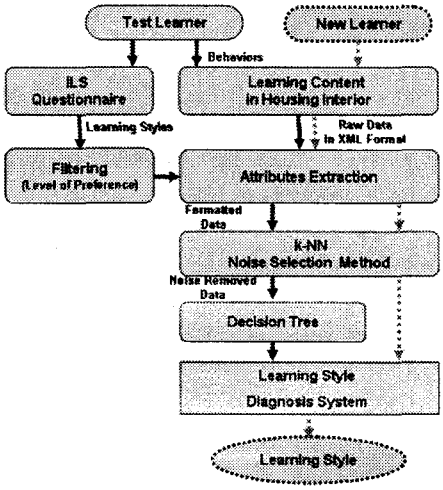

그림 6. 학습자 진단을 위한 작업 흐름도

4. 실험

실험을 위하여 DOLLS-HI 학습 컨텐츠를 이용하여 성균관대학교 신입생중 483명을 대상으로 15분동안 학습하게 하였다. 학습을 통하여 얻은 데이터는 속성 추출과정을 거쳐 각 학습 성향별로 의사결정 나무를 이용하여 학습에 사용하였다. 표 2는 483명의 데이터중에서 교차검증방법을 이용하여, 300명의 데이터는 학습 데이터로 사용하고 183명의 데이터는 테스트 데이터로 사용하여 얻은 에러율이다.

표 2. 483명 학습자 데이터의 분석

	Active& Reflective	Global & Sequential	Sensing& Intuitive	Visual & Auditory
1	0.5196	0.5027	0.3715	0.2404
2	0.4371	0.4590	0.3715	0.2131
3	0.5136	0.5136	0.3660	0.1912
4	0.5191	0.4863	0.3551	0.2513
5	0.5136	0.4808	0.3606	0.2005
평균	0.5005	0.4885	0.3650	0.2193

다음으로, 제안하는 방법인 k-NN 기법을 이용하여 노이즈를 제거한 뒤 의사결정 나무(C4.5) 방법을 이용하여 학습시키고 에러율을 확인하였다. 표 3는 483명의 데이터중에서 노이즈를 제거하고 남은 데이터를 학습 성향별로 표현한 것이다.

표 3. k-NN 기법을 이용하여 노이즈를 제거하고 남은 데이터

Active& Reflective	Global & Sequential	Sensing& Intuitive	Visual & Auditory
230	276	263	291

표 4은 노이즈를 걸러낸 남은 데이터를 이용하여 60%는 학습 데이터로 이용하고, 나머지 40% 는 테스트 데이터로 이용하였다.

표 4. 노이즈를 제거한 후의 분석 결과

	Active& Reflective	Global & Sequential	Sensing& Intuitive	Visual & Auditory
Learning	38(60%)	165(60%)	157(60%)	174(60%)
Test	91(40%)	111(40%)	106(40%)	117(40%)
1	0.4835	0.3243	0.3018	0.0854
2	0.3736	0.2882	0.3301	0.1367
3	0.4395	0.2072	0.3867	0.1111
4	0.4285	0.3513	0.3584	0.094
5	0.4835	0.2702	0.2924	0.1025
평균	0.4417	0.288	0.333	0.105

첫 번째 실험은 수집된 데이터에서 노이즈를 제거하지 않고 학습시켜 에러율을 확인하고, 두 번째 실험은 노이즈를 제거한 뒤 학습에 사용하여 에러율을 확인하였다. 표의 평균 에러율에서도 알 수 있듯이 노이즈를 제거한 결과가 그러하지 않은 결과 보다 좋게 나온 것을 확인 할 수 있다.

5. 결론 및 향후 연구

본 논문은 k-NN 기법을 이용하여 노이즈가 포함된 학습자의 데이터를 분석하는 방법을 소개하였다. 실험을 통하여 에러율이(Active & Reflective:0.05, Global & Sequential:0.20, Sensing & Intuitive:0.032, Visual & Auditory:0.11) 향상된 것을 확인할 수 있다. 또한 학습자 데이터뿐만 아니라, 노이즈가 포함된 다른 데이터에서도 사용가능하다. 향후 연구로는 노이즈에 대한 정의와 검증, 그리고 노이즈 데이터를 이용한 전처리 방법이 연구되어야 할 것이다.

참 고 문 헌

[1] 김현숙, "정규화되고 노이즈가 들어간 데이터를 이용한 신뢰네트웍 학습알고리즘들의 성능 비교 분석", 석사학위논문, 포항공과대학, 1995.

[2] 이성렬, "점진적 클러스터링에서 노이즈 제거", 석사학위논문, 숭실대학교, 2000.

[3] 서정국, 차형태, "잡음 패턴의 지능적 추정을 통한 음질 개선 알고리즘", 퍼지및지능시스템학회 논문지 2005, Vol. 15, No. 2, pp. 230-235, 2005.

[4] Cha, H. J., Kim, Y. S., Park, S. H., Yoon, T. B., Jung, Y. M., and Lee, J. H., "Learning Styles Diagnosis based on User Interface Behaviors for the Customization of Learning Interfaces," Proc. 8th Int'l. Conf. on Intelligent Tutoring Systems (ITS), Jhongli, June, 2006.

[5] Felder, R., Silverman, L.,"Learning and Teaching Styles in Engineering Education," Engineering Education, Vol. 78, No. 7, pp. 674-681, 1988.