k-NN 기법을 이용한 학습자의 학습 행위 데이터의 이상치 분석

Outlier Analysis of Learner's Learning Behaviors Data using k-NN Method

윤태복, Taebok Yoon*, 정영모, Youngmo Jung*, 이지형, Jeehyong Lee**,
차현진, Hyunjin Cha*, 박선희, Seonhee Park*, 김용세, Yongse Kim*

*성균관대학교 창의적설계추론 지적교육시스템 연구단,
**성균관대학교 정보통신공학부

요약 지능형 학습 시스템은 학습자의 학습 과정에서 수집된 데이터를 분석하여 학습자에게 맞는 전략을 세우고 적합한 서비스를 제공하는 시스템이다. 학습자에게 적합한 서비스를 위해서는 학습자 모델링 작업이 우선시 되며, 이 모델 생성을 위해서 학습자의 학습 과정에서 발생한 데이터를 수집하고 분석하게 된다. 하지만, 수집된 데이터가 학습자의 일관되지 못한 행위나 비예측 학습 성향을 포함하고 있다면, 생성된 모델을 신뢰하기 어렵다. 본 논문에서는 학습자에게서 수집된 데이터를 거리기반 이상치 선별 방법인 *k*-NN 을 이용하여 이상치를 선별한다. 실험에서는 홈 인테리어 컨텐츠 기반에 학습자의 학습 행위에 대한 학습 성향을 진단하기 위한 DOLLS-HI 를 이용하여, 수집된 학습자의 데이터에서 이상치를 분류하고 학습 성향 진단을 위한 모델을 생성하였다. 생성된 모델은 이상치 분류전과 비교하여 신뢰가 향상된 것을 확인하였다.

핵심어: Outlier Analysis, Learner Modeling, Educational Data Mining

1. 서론

지능형 학습 시스템 (ITS : Intelligent Tutoring System)은 학습자에게 지능적인 학습 환경을 제공해 주기 위한 시스템을 말하며, 학습자의 수준, 능력, 성향 등에 따라 그에 맞는 적합한 서비스를 제공해 준다[1]. 학습자에게 적합한 학습 환경을 제공해 주기 위해서는 학습 과정에서 수집된 데이터를 분석하여 모델을 생성한다. 수집된 학습자의 데이터를 분석하여 의미 있는 정보를 추출하기 위하여, 전처리 과정은 매우 중요하게 여겨진다. 특히 학습자의 일관되지 않은 정보는 생성된 모델의 신뢰성과 성능에 큰 영향을 미치기 때문에 비정상적인 데이터를 선별하기 위한 방법이 요구된다. 본 논문은 학습자의 학습 과정에서 수집된 학습 행위 데이터를 이용하여 학습자의 성향을 진단하는 DOLLS-HI(Diagnosis of Learner's Learning Styles Housing Interior)[2][3][4]를 소개한다. 그리고 이 시스템의 진단 모델을 만들기 위해 수집된 데이터를 분석하는 방법에 의사 결정 트리를 이용하여 진단 모델을 만들었다. 모델 생성에 사용되는 데이터는 거리 기반의 이상치 분류 방법인 *k*-NN(Nearest Neighbor)을 이용하여 이상치를 제거한 경우와 이상치를 제거하지 않은 경우를 비교 실험하여 이상치 제거후의 모델이 더 유효함을 확인 하였다. 본 논문의 구성은 다음과 같다. 2장에서는 이상치의 정의와 학습 환경에서 발생할 수 있는 이상치에 대해 이야기 하고, 3장에서는 거리 기반 이상치 분류 방법인 *k*-NN에 대한 이야기 한다. 4장에서는 DOLLS-HI에 대해 설명하고, 5장에서는 이상치 선별을 위한 실험에 대해 이야기 한다. 끝으로 6장에서는 결론과 향후 연구로 맺는다.

2. 학습자 데이터의 이상치

2.1 이상치 데이터

흔히, 데이터의 모델이나 일반적인 행동에 대응하지 못하는 데이터 객체들이 존재한다. 남아 있는 데이터 집합에 불일치 되거나, 심하게 다른 데이터 객체들을 이상치라고 부른다[5]. 이상치들은 측정이나 인위적인 실행 오류에 의해 발생할 수 있는데, 예를 들어 지능형 문제 시스템이 있다고 가정하자. 문제 풀이 과정으로부터 학습자의 학습 정보를 수집하고, 문제의 오답율 및 문제 풀이에 소요된 시간 등을 분석하여 학습자 모델을 만들게 된다. 그런데, A 라는 학생이 학습을 빨리 끝내고 싶은 마음에 성의 없게 문제를 풀었다고 가정하자. 이 때 A학생의 데이터는 모델 생성에 신뢰도를 떨어지게 하는 요인으로 작용하게 되어 서비스에 질을 저하시킨다. 이때 학생 A의 데이터를 이상치라고 할 수 있다.

2.2 학습 환경에서의 이상치 데이터

앞서 설명한 바와 같이 학습 환경에서 이상치 분석은 매우 중요한 작업이며, 이상치의 원인을 파악하는 작업 역시 중요하게 여겨진다. 교육 환경에서 수집된 학습자 데이터의 이상치가 발생 될 수 있는 요인은 크게 두 가지로 나뉠수 있으며, 아래 와 같다.

첫 번째, 학습자의 학습과정에서 발생한 심리적 변이로 인한 일관되지 못한 데이터 – 내적 요소

학습자의 심리적 변화에는 의도한 경우와 의도하지 않은 경우로 나눌 수 있다. 학습자가 의도한 경우는 현재 학습 상황의 재미를 잃거나 더 흥미를 끄는 다른 무엇가로 인하여 의미 없는 학습 데이터가 수집되는 경우이다. 의도하지 않은 경우는 학습 과정에서 학습자의 수준이 향상하거나 학습 스타일에 대한 가치관이 변화하여 기존에 보이던 학습 성향과 다른 모습을 보이는 경우이다. 이 경우 의도한 학습자의 이상치 데이터 보다 변화 기복이 작은 특징이 있다. 전자의 경우는 이상치 이면서 데이터의 이상치 성향이 강하므로 제거하여 학습자 모델의 신뢰도를 높이는데 사용할 수 있으며, 후자의 경우 동적인 학습자 모델 생성에 사용 가능하다.

두 번째, 학습자의 학습에 영향을 줄 수 있는 주변 환경 요인의 변화로 인한 데이터 – 외적 요소

학습 환경은 학습자의 학습 성향에 주요한 영향을 미친다. 동일한 컨텐츠를 제공하고 아주 짧은 시간과 아주 긴 시간 두 가지 경우의 학습 제한 시간이 주어진다고 가정하자. 이 때 학습자는 본래 자기가 가지고 있는 학습 성향을 뒤로하고 시간이란 요소에 억매여 학습 데이터를 만들어낸다. 이 데이터는 학습자 모델을 위한 데이터로 사용하기에 부족한 부분이 있다고 하겠다. 마찬가지고 제공되는 학습 컨텐츠의 수준이나 학습에 사용되는 장비, 주변 소음, 밝기 등외에 학습 상황에 발생 할 수 있는 모든 외적 요인을 들 수 있다. 이런 경우의 이상치 데이터는 학습자의 학습 활동에 미치는 환경적 요인을 찾는 다거나 학습 컨텐츠의 수준 결정을 위한 기초 자료로 사용 할 수 있다.

위에서는 학습 환경에서 학습자의 내/외적의 요인으로 인하여 발생 할 수 있는 이상치에 대하여 기술하였다. 추출된 이상치는 아래와 같은 용도로 사용 할 수 있다.

- 이상치 제거를 통한 학습자 모델링의 신뢰도 향상
- 학습자의 학습 활동에 미치는 환경적 요인 분석
- 동적인 학습자 모델링 생성의 기초가 되는 자료
- 예측되지 않은 새로운 학습자의 학습 모델링 생성

3. *k*-NN 방법을 이용한 이상치 분석

본 논문에서 이용한 이상치 선별 방법은 *k*-NN 기법에 기반을 두고 있다. 주어진 데이터를 공간상에 사상시키고, 한 개체와 다른 개체들 사이의 분포 정도를 계산한다. 주변 분포 정도를 측정하여 이상치 여부를 판단하는 것이다[6].

3.1 *k*-NN 기법을 이용한 이상치 선별 방법

k-NN 기법을 이용한 이상치 선별 방법은 각 개체간에 거리를 기반으로 계산한다. 예를 들어 그림 1와 같이 10개의 개체가 있다고 가정하자 각 개체가 가지고 있는 속성은 x와 y 두 개이며 정수형의 값을 갖는다. 각 속성의 종속변수인 class는 Yes와 NO 두 가지 값을 가질 수 있다. 그림 2는 10개의 개체를 좌표상에 표현한 것이다.

그림에서 보면 알 수 있듯이 10번째 개체의 class를 알 수 없다. 그렇다면 여기서 10번째 class를 무엇이라 할 수 있는가? 아마도 “No”라고 대답하는 사람이 대다수 일 것이다. 그 이유는 현재 알지 못하는 10번째 개체의 주변 분포가 Yes보다는 No가 더 많기 때문이다. 만약 주어진 데이터에 10번째 개체의 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	?

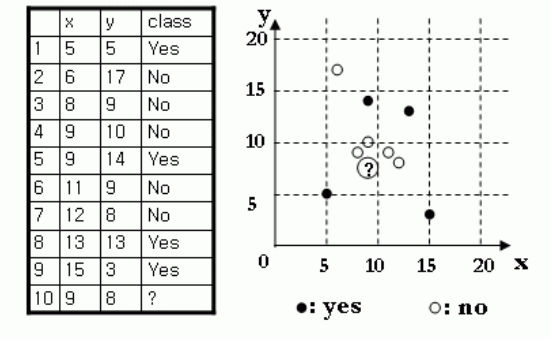

그림 1. *k*-NN 방법의 예

주변 분포를 볼 때 No가 더 많으므로, 10번째 데이터를 이상치로 결정할 수 있을 것이다. 이처럼 주어진 전체 데이터에서 한 개체의 주변 분포를 측정하여 이상치 여부를 판단하는 것이 *k*-NN 기법을 이용한 이상치 선별 방법이다.

3.2 *k*-NN 기법의 이상치 선별을 위한 고려사항

분포 측정을 위한 범위 설정: 한 개체 주변의 분포를 측정하기 위해서는 범위 설정이 필요하다. 여기서 범위는 이상치 여부를 판단하기 위한 개체로부터 일정 거리를 기준으로 할 것인지(그림 2 (b)), 근접한 개수를 기준으로 할 것인지(그림 2 (a))에 대한 여부이다. 두 가지 방법은 실제 적용을 통하여 비교 사용하여야 한다.

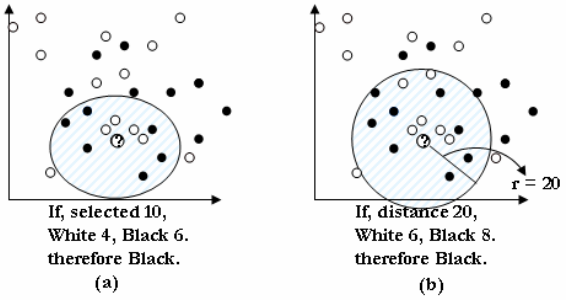

그림 2. 분포 측정을 위한 범위 설정 예:
(a) 개수를 이용한 방법 (b) 거리를 이용한 방법

분포내 개체와의 거리 비율: 범위가 설정되고 나면 범위내의 개체 개수를 파악하여 이상치 여부를 판단하게 된다. 하지만 단순히 개수만을 비교한다면 근접거리에 존재하는 개체의 의미가 무시될 수 있다.

예를 들어, 그림 3은 근접한 10개의 개체를 선택하였다. 그림 3의 (a)의 경우처럼 개수만을 비교한다면 "White"는 4개이고 "black"는 6개 이므로 "?"는 "Black"가 되어야 한다. 하지만, 원안의 개체간에 거리비율에 따른 분포를 고려한다면 그림 3의 (b)와 같이 "?"는 "White"에 가깝다고 할 수 있다.

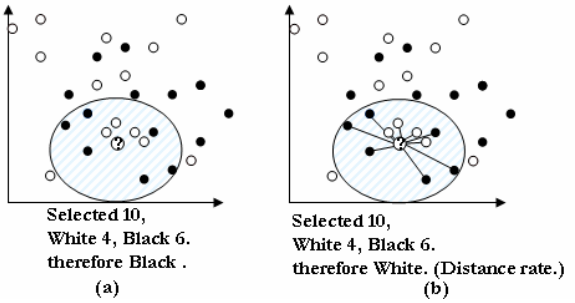

그림 3. 분포내 객체와 거리 비율 예:
(a) 단순 개수 비교 (b) 개체의 거리에 따른 가중치 부여

전체 데이터의 분포에 따른 비율: 제안하는 방법은 이상치 판단을 위한 개체 주변의 분포를 측정하여 결정한다. 다시 말하면 분포의 측정이란 내 주변에 나와 비슷한 개체가 얼마나 존재하는가 하는 여부이다. 주변의 범위를 설정하고 범위내에 개체들의 거리 비율에 따라 개수를 파악한다고 하더라도 초기에 주어진 데이터의 비율이 편중 되어 있다면, 그것 또한 연산에 고려되어야 할 것이다. 예를 들어 그림 4에서 보면 이상치 판단을 하고자 하는 데이터의("?") 주변 10개 분포를 비교하면 "White"가 6개이고 "Black"이 4개 이므로 "White"라고 판단해야 한다. 하지만, 초기에 주어진 전체 데이터가 30개 이며 이중에서 "White"는 20개, "Black"는 10개로 "White"에 편중되어 있으므로 전체 데이터의 비율을 고려하여 "Black"으로 판단된다.

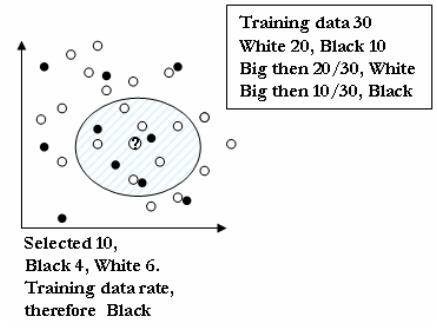

그림 4. 전체 데이터의 비율을 고려한 이상치 판단

4. 학습 행위 데이터를 이용한 학습자 성향 진단

4.1 학습자 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[7]은 앞의 예에서와 같이 학습 정보를 이해하는 차원에서 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. 학습자의 학습 성향 및 설명

학습성향		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 세부 내용을 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.

4.2 학습자 성향 파악을 위한 연구

Global & Sequential : 정보를 이해할 때 학습자의 선호에 의해 교육 컨텐츠를 선택하는지, 교육 전문가가 의도 하는 학습 순서에 따라 학습하는지를 알아보기 위하여 정해진 순서에 따라 학습 할 수 있는 방법과 학습자가 학습 순서를 정하여 학습 할 수 있는 방법이 가능하도록 인터페이스를 설계하였다(그림 5).

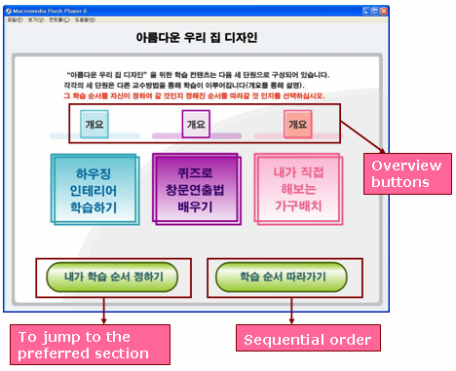

그림 5. Global & Sequential를 위한 인터페이스

Visual & Auditory : 정보를 습득할 때 그림위주의 설명을 선호하는지 텍스트 위주의 설명을 선호하는지를 알아보기 위하여 그림위주의 학습 설명 버튼과 텍스트 위주의 학습 설명 버튼을 인터페이스에 설계하였다(그림 6).

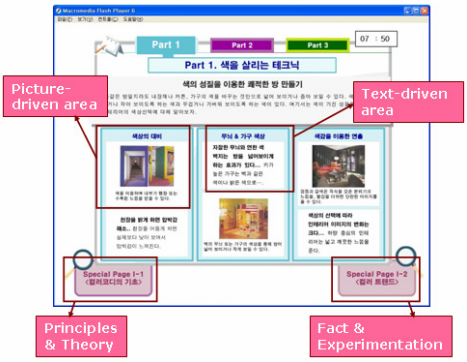

그림 6. Visual & Auditory를 위한 인터페이스

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(그림 7).

그림 7. Sensing & Intuitive를 위한 인터페이스

Active & Reflective : 정보를 활용 할 때 적극적인지 수동적인지를 알아보기 위하여 학습 게시판에 의견 달기, 의견 보기 등의 기능을 설계하였다(그림 8).

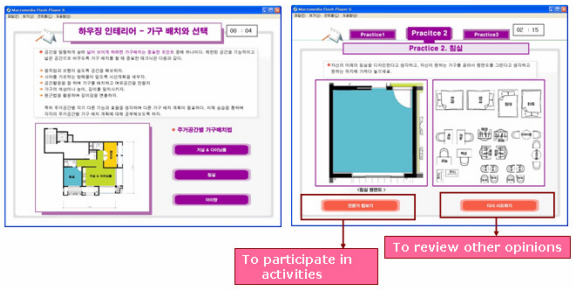

그림 8. Active & Reflective를 위한 인터페이스

그림 1,2,3,4는 학습 성향 관점에 따른 홈 인테리어 학습 컨텐츠인 DOLLS-HI(Diagnosis of Learner's Learning Styles Housing Interior)[4]에 인터페이스의 일부분이다.

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

4.2 이상치 선별을 이용한 학습자 진단 과정

학습 성향을 분류하기 위한 전체 프로세스는 Felder & Silverman이 제시한 학습 성향을 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[8]. 우선 학습자는 ILS 온라인 설문에 참여하여, 학습 성향별(Global & Sequential, Visual & Auditory, Sensing & Intuitive, Active & Reflective)로 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 버튼의 클릭 등)는 XML 파일로 기록된다. 이후 전체 피 실험자의 XML 데이터가 수집이 되면, 전처리 과정을 거쳐 제안하는 방법인 k-NN 기법을 이용하여 이상치를 제거한다. 이상치가 제거된 데이터는 의사결정 트리 방법을 이용하여 학습시키고, 생성된 결과는 학습자 진단을 위한 기반 정보로 사용한다. 그림 9는 위에서 설명한 작업흐름을 표현한 것이다[9].

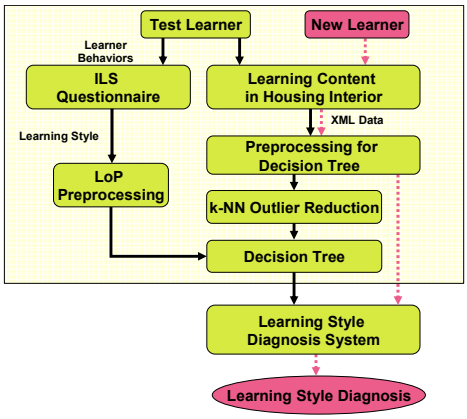

그림 9. 학습 성향 진단을 위한 전체 흐름도

5. 실험

실험을 위하여 DOLLS-HI 학습 컨텐츠를 이용하여 성균관대학교 신입생중 483명을 대상으로 실험을 실시하였다. 먼저 실험 대상학생들에게 ILS 온라인 설문지를 통하여 학습 성향(Global & Sequential, Visual & Auditory, Sensing & Intuitive, Active & Reflective)을 파악하였다. 표 2는 실험에 참가한 학생들의 학습 성향별 분포이다.

표 2. ILS 온라인 설문지를 통한 학습자 학습 성향 분포(개수)

	Global	Sequential	Visual	Auditory
Num.	264	219	416	67
	Sensing	Intuitive	Active	Reflective
Num.	357	126	262	221

그리고, 학생들에게 DOLLS-HI를 이용하여 15분 동안 하게 하였다. 학습 후에 얻은 학생들의 데이터는 전처리 과정을 거쳐 의사 결정 트리를 이용하여 분석하였다. 표 3은 의사 결정 학습을 위해 정의된 속성이다. 표에서는 Visual & Auditory 성향에 대해서 표현한 것이다. 예를 들어 "VA1_MainImgClick"의 경우 학습 화면에서 이미지 컨텐츠를 클릭한 수를 나타내는 속성이며, "VA9_Relevant-ButImgClick"는 학습과 연관된 이미지 컨텐츠를 보기 위하여 버튼을 클릭한 수이다. 다른 속성들도 Visual & Auditory와 같이 학습자의 학습 의도를 파악하기 위한 요소를 의사 결정 트리 방법을 위해 사용하였다.

표 3. 의사 결정 트리를 위한 진단 속성

Style	Attribute List	
Visual & Auditory	VA1_MainImgClick	VA10_RelevantButImgTime
	VA2_MainImgTime	VA11_RelevantButTxtClick
	VA3_MainTxtClick	VA12_RelevantButTxtTime
	VA4_MainTxtTime	VA13_DetailImgTime
	VA5_OptionButImgClick	VA14_DetailTxtTime
	VA6_OptionButImgTime	VA15_ChosenImgClick
	VA7_OptionButTxtClick	VA16_ChosenTxtClick
	VA8_OptionButTxtTime	VA17_TotalImgTime
	VA9_RelevantButImgClick	VA18_TotalTxtTime

표 4는 483명의 데이터 중에서 교차검증방법을 이용하여, 300명의 데이터는 학습 데이터로 사용하고 183명의 데이터는 테스트 데이터로 사용하여 얻은 에러율이다.

표 4. 의사결정나무 방법에서 얻은 에러율(%)

	Global & Sequential	Visual & Auditory	Sensing & Intuitive	Active & Reflective
1	45.9	24.59	45.35	50.27
2	50.81	24.04	44.26	48.63
3	41.53	22.4	49.72	48.08
4	44.8	18.57	42.62	49.18
5	49.18	20.76	54.09	45.9
6	48.63	22.4	47.54	44.26
7	43.16	31.69	44.26	54.64
8	43.71	27.86	42.62	49.72
9	45.35	26.22	43.71	51.91
10	47.54	22.95	44.26	57.37
avg.	46.061	24.148	45.843	49.996

표 4와 같이 각 스타일에 따른 평균 에러율은 매우 높은 편이며 학습자의 성향을 진단하기 위한 모델로 사용하기엔 부족하다.

다음은 수집된 데이터에서, 앞서 설명한 *k*-NN 방법을 이용하여 이상치를 선별하고 남은 데이터를 학습에 사용하였다. 한 개체를 기준으로 주변 개체를 선택하는 범위는 10, 20 그리고 30개, 세 번 실시하였다. 즉, 한 개체를 기준으로 분석에 사용되는 주변개체의 수를 다르게 한 것이다. 또한 주어진 데이터의 비율(예 Global : Sequential = 262:221)을 고려하여 이상치 유무를 판단하였다. 표 5,6은 처음에 주어진 데이터와 이상치 선별 후 남은 데이터의 개수를 각각 표현한 것이다.

표 5. 학습 성향별 이상치 선별후 남은 데이터(개수)
(Global & Sequential, Visual & Auditory)

	Global	Sequential	Visual	Auditory
Init.	264	219	416	67
10	161	116	266	27
20	149	109	240	34
30	133	127	300	28

표 6. 학습 성향별 이상치 선별후 남은 데이터(개수)
(Sensing & Intuitive, Active & Reflective)

	Sensing	Intuitive	Active	Reflective
Init.	357	126	262	221
10	163	63	132	118
20	183	52	143	102
30	128	77	110	117

처음 데이터의 개수에서 각각의 이상치를 제외하고 남은 데이터는 의사 결정 트리 방법을 이용하여 학습하고 테스트하여 에러율을 확인하였다 (표 7, 8). 학습에 사용된 개수는 전체 데이터의 60%, 테스트 데이터는 40%를 이용하였다. Global & Sequential에서 주변 분포 10개 선정후 이상치를 제거한 경우 남은 데이터가 각각 161, 116이다(표 5). 이때 학습 데이터는 166개(60%)이며, 테스트에 사용된 데이터는 111개(40%)로 랜덤하게 선정된다.

표 7. 이상치 분석 방법 적용후 에러율(%)
(Global & Sequential, Visual & Auditory)

	Global & Sequential			Visual & Auditory		
	10	20	30	10	20	30
1	22.58	9.708	9.615	11.22	10	5.343
2	17.2	13.59	8.653	9.183	14.54	6.87
3	15.05	12.62	7.692	10.2	10.9	5.343
4	13.97	10.67	13.46	9.183	8.181	6.87
5	18.27	15.53	10.57	16.32	10.9	3.816
6	15.05	9.708	5.769	11.22	9.09	8.396
7	17.2	10.67	8.653	8.163	13.63	6.106
8	12.9	9.708	11.53	11.22	10.9	5.343
9	18.27	7.766	10.57	15.3	8.181	6.106
10	13.97	3.883	5.769	10.2	11.81	7.633
Avg.	16.446	10.385	9.228	11.22	10.81	6.182

표 8. 이상치 분석 방법 적용후 에러율(%)
(Sensing & Intuitive, Active & Reflective)

	Sensing & Intuitive			Active & Reflective		
	10	20	30	10	20	30
1	31.57	23.65	34.14	40.00	31.63	36.26
2	34.21	22.58	24.39	39.00	30.61	35.16
3	27.63	27.95	28.04	40.00	30.61	29.67
4	32.89	25.8	15.85	42.00	29.59	31.86
5	39.87	31.18	23.17	39.00	30.61	31.86
6	27.63	24.73	30.48	42.00	24.48	29.67
7	35.52	30.1	31.70	40.00	33.67	23.07
8	25	22.58	19.51	36.00	30.61	28.57
9	31.57	23.65	24.39	34.00	31.63	34.06
10	30.26	26.88	25.60	39.00	32.65	35.16
Avg.	31.61	25.91	25.73	39.10	30.61	31.53

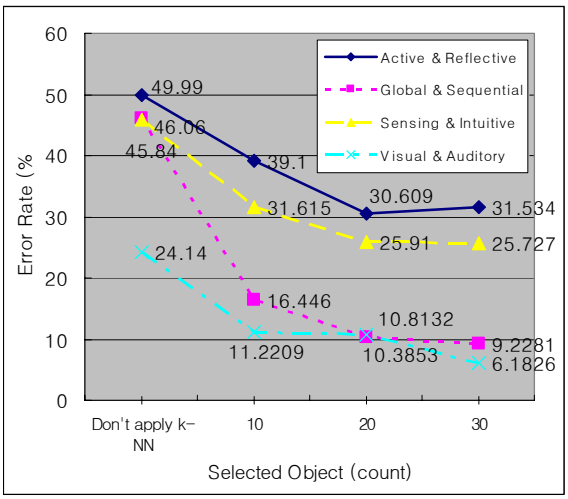

그림 10. 개체 수에 따른 에러율 비교

6. 결론 및 향후 연구

본 논문은 학습자의 학습과정에서 얻은 학습 행위 데이터를 분석하여 학습자의 학습 성향을 진단하는 방법을 소개하였다. 홈 인테리어 컨텐츠 기반한 DOLLS-HI를 이용하여 학습 데이터를 수집하고 의사 결정 트리를 이용하여 진단 모델을 만들었다. 또한 학습자의 데이터를 분석하는 과정에서 k-NN 방법을 이용하여 수집된 데이터와 상이한 분포를 보이는 데이터는 이상치라고 판단하고 선별하였다. 이상치를 선별하는 방법에는 3가지를 실험을 하였으며 이상치 선별 전과 비교하였다. 그림 10은 이상치 선별을 적용하지 않은 경우와 이상치 선별에 사용된 인수가 각각 10, 20, 30인 경우에 대한 의사결정 트리 에러율을 그래프로 표현한 것이다. 그림에서 보면 알 수 있듯이 이상치를 선별하지 않은 경우 보다 낮은 에러율을 보이고 있다. 만약 이상치를 제거하지 않은 결과를 진단 모델로 사용한다면, 새로운 학습자에 대한 진단 결과를 의심하지 않을 수 없을 것이며, 데이터 분석에 이상치 선별은 생성된 진단 모델의 신뢰성을 높이는데 중요한 역할을 한다.

향후 연구로는 이상치로 선별된 데이터를 활용하는 방법이 필요할 것이다. 새로운 학습자의 데이터가 수집되었을 때, 진단 모델을 통하여 진단하기 앞서 이상치 인지 여부를 먼저 판단한다면 결과의 신뢰성이 향상될 것이다. 또한 이상치 데이터들의 분석을 통하여 사전에 예측하지 못했던 새로운 학습 스타일에 대한 모델도 만들 수 있을 것이다.

참고문헌

[1] Ueno, M., Nagaoka, K. (2002). Learning Log Database and Data Mining system for e-Learning -On-Line Statistical Outliers Detection of irregular learning processes, IEEE Int'l. Conf. on Advanced Learning Technologies (ICALT).

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

[3] Cha, H. J., Kim, Y. S., Lee, J. H., and Yoon, T. B. (2006). An Adaptive Learning System with Learning Style Diagnosis based on Interface Behaviors, Workshop Proc. of Int'l Conf. on E-learning and Games (Edutainment).

[4] Cha, H. J., Kim, Y. S., Park, S. H., Cho, Y. J., and Pashkin, M. (2005). Adaptive Learning Interface Customization Based on Learning Styles and Behaviors, Proc. Int'l. Conf. on Computers in Education (ICCE).

[5] Han, J., Kamber, M. (2001). Data Mining: Concepts and Techniques. Academic Press.

[6] Knorr, E., Ng. R. (1997). A unified notion of outliers: Properties and computation. Int'l. Conf. Knowledge Discovery and Data Mining (KDD), pp219-222.

[7] Felder, R., Silverman, L. (1988). Learning and Teaching Styles in Engineering Education, Engineering Education, 78(7), pp674-681.

[8] Constantine, S., Charalampos, K., Adamantios, K. (1997). Decision Making in Intelligent User Interface, Intelligent User Interface1997 Conf., pp195-202.

[9] Margaret, H. D. (2003). Data Mining Introductory and Advanced Topics, Pearson Education Inc., New Jersey.