데이터마이닝에서 교사학습에 의한 속성 가중치 최적화

강영구*, 차진호*, 김진원
서세미디어*, 숭실대학교 컴퓨터학과
m kang@ccmedia.co.kr, jh cha@info.ssu.co.kr, mkim@computing.ssu.ac.kr

Supervised Feature Weight Optimization for Data Mining

Myung-Ku Kang*, Jin-Ho Cha, Myungwon Kim
CCmedia Inc*, School of Computing, Soongsil Univ.

요 약

최근 과학화된 분야인 데이터 마이닝이 빠르게 성장하고 있으며, 이와 같이 데이터를 수집하고 사용하기 편리하게 처리하는 DBMS(Database Management System), 데이터 웨어하우징(Data Warehousing), 데이터마이닝(Data Mining) 기법들이 많이 연구되고 있다.

다음이, 데이터마이닝에 있어서 최근에는 기업들의 중요한 경영 속에서 살아남기 위해 영업사원의 고객관리, 서비스, 고객분석 등에 걸쳐 간단히 보존하는 수많은 데이터를 고객목표와 정보의 필요에 따라 신속하고 정확하게 마케팅을 실시하는 경영기반의 고객관리(CRM: Customer Relationship Management) 기능을 하는 도구가 되고 있다.

그러나 실제 웹데이터는 시간이 지날수록 데이터의 양이 늘어나게 되어 생기는 속성과 관계를 갖게 되어 영업사원의 기법을 이용하는데 많은 시간과 비용이 소요된다. 또한 어느 속성이 중요한지 알 수 없는 중요한 속성이 중요하지 않은 속성에 의해 제대로 분석되지 않을 수 있다.

따라서, 최근 마이닝과의 속도도 향상시키고 효율을 높이기 위해 중요한 속성을 선택하는 부분숙성선택(Feature Subset Selection)과 속성의 중요도에 따라 가중치를 부여하는 속성가중치(Feature Weighting)에 관한 연구가 관심의 대상이 되고 있다.

속성의 선택은 클러스터 정보를 얻고 있는 데이터에 적용하여 좋은 분류 결과를 얻을 수 있는 속성들로 선택하는 방법으로 교차학습이 분류기별로 사용되는 반면, 속성가중치는 클러스터 정보가 없는 데이터에 주어진 정보에 의하여 속성가중치의 관계를 잘 표현하고 있는가에 따라 가중치를 부여하는 방법으로 비교학습으로 클러스터링을 한 후에 적용할 수 있다. 그러나 이 경우 클러스터 정보가 없는 상태에서 데이터의 관계만으로 가중치를 주기 때문에 사용하는 그 결과가 맞았는지에 대한 여부를 판단할 수 있는 문제점을 보이고 있다.


예를 들자, 다음 <표 1>과 같은 전자상거래에서 판매한 사례를 살펴보자. 속성으로는 '나이', '성별', '구입금액', '방문파수' 등 4개이며 42개의 데이터로 구성되어 있다. (성별의 1은 남성, 0은 여성) 품목의 구입여부가 중요할 수 있는 품목도 포함한 34회이다.)

<table>
<thead>
<tr>
<th>번호</th>
<th>나이</th>
<th>성별</th>
<th>구입금액</th>
<th>방문파수</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>23</td>
<td>1</td>
<td>22</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>47</td>
<td>0</td>
<td>33</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>30</td>
<td>1</td>
<td>24</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>36</td>
<td>1</td>
<td>25</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>54</td>
<td>1</td>
<td>26</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>44</td>
<td>1</td>
<td>27</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>35</td>
<td>1</td>
<td>28</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>33</td>
<td>0</td>
<td>29</td>
<td>1</td>
</tr>
<tr>
<td>9</td>
<td>29</td>
<td>0</td>
<td>30</td>
<td>1</td>
</tr>
<tr>
<td>10</td>
<td>23</td>
<td>1</td>
<td>31</td>
<td>0</td>
</tr>
<tr>
<td>11</td>
<td>37</td>
<td>1</td>
<td>32</td>
<td>1</td>
</tr>
<tr>
<td>12</td>
<td>23</td>
<td>1</td>
<td>33</td>
<td>1</td>
</tr>
<tr>
<td>13</td>
<td>54</td>
<td>1</td>
<td>34</td>
<td>1</td>
</tr>
<tr>
<td>14</td>
<td>36</td>
<td>1</td>
<td>35</td>
<td>1</td>
</tr>
<tr>
<td>15</td>
<td>37</td>
<td>1</td>
<td>36</td>
<td>1</td>
</tr>
<tr>
<td>16</td>
<td>33</td>
<td>0</td>
<td>37</td>
<td>1</td>
</tr>
<tr>
<td>17</td>
<td>40</td>
<td>1</td>
<td>38</td>
<td>0</td>
</tr>
<tr>
<td>18</td>
<td>37</td>
<td>1</td>
<td>39</td>
<td>1</td>
</tr>
<tr>
<td>19</td>
<td>29</td>
<td>1</td>
<td>40</td>
<td>1</td>
</tr>
<tr>
<td>20</td>
<td>48</td>
<td>1</td>
<td>41</td>
<td>1</td>
</tr>
<tr>
<td>21</td>
<td>32</td>
<td>0</td>
<td>42</td>
<td>0</td>
</tr>
</tbody>
</table>

위의 데이터에서 '구입금액'과 '방문파수'만을 산술로 표시하면 아래의 그래프 같은다. 가중치를 전혀 고려하지 않고 클러스터링을 할
이와 같이 차원 $d$가 커짐에 따라 기하급수적으로 늘어나는 공간은 모두 탐색하기 어려우며 최근에는 최적해 문제에 많이 사용되는 유전자 알고리즘들을 이용하여 부문특성을 선택하는 방법들이 많이 제시되고 있다.[1]

2. 속성 가중치

유전자의 중요한 기법 중 하나인 클러스터링은 클러스터의 정보가 없는 데이터를 분석하는 기법으로, 클러스터 정보를 이용하여 선택한 기존의 부문 속성선택법을 제한할 수 있다. 따라서 클러스터링과 같은 비교학습 알고리즘을 적용하기 위해 클러스터의 정보가 없이 데이터의 분포와 관계를 나타내는 평균값을 이용하여 평균값만의 최소 또는 최대를 도출하여 가중치를 부여하는 속성가중기법이 제시되고 있다.[2][3][4]

속성가중치는 비교학습에 의한 속성선택으로 고속적으로 충족하는 값을 빠르게 탐색하려는 필요성을 강화하고, 데이터의 관계를 이용한 평균값으로 대체하는 가중치를 찾아주는 방법으로 분리할 수 있다.

2.2 클러스터링 방법

클랜, 파이지, 디미가 유전자 알고리즘이 같이 인공지능 기법을 이용한 학습, 훈련 알고리즘은 알고리즘에 따라 속성기반의 방법으로 이와 같은 인공지능기법을 이용한 연구가 진행되고 있다.[2][3]

3. 속성가중치 알고리즘

본 알고리즘의 순서는 다음과 같다. 전체적인 속성기반 유전자 알고리즘의 순서도를 따른다.

3.1 사용자 정보 입력

표 1과 같이 클러스터 정보가 없는 데이터를 분석하기 위해서 사용자와 빠른 계명 데이터를 수집하여 사용자가 원하는 대로 학습할 수 있는 각각의 표본 데이터에 대응한 클러스터 정보를 수집한다. 본 논문에서는 <그림 4>와 같이 입력이 33%로 표본 데이터의 입력을 부여하였다. 본 논문에서는 <그림 5>와 같이 각 주요성가중치를 산출한 결과를 표시한 것이다.

3.2 계산 과정

한 친구의 가중치를 표현하기 위한 각각의 엄체는 성과자择을 적응하기 쉽게 하기 위해 0.1로 이루어진 m개의 2차 곡선을 표시하였다. 사용자인 경우 결과의 입력은 m×l로 벡터의 수치는 0.7의 가중치로 표시할 수 있다. m의 값이 0, 1에 가까운 값이 등분수로 나누어진 경우가 설정한다. 예를 들어, m=4일 경우 m=25개로 나누어진 경우가 1개. m=8일 경우 m=25개로 나누어진 경우가 3개가 된다.

3.3. 학습도 힘 [5][6]
본 논문에서 적합도 함수가 의미하는 것은 주어진 가중치(개체)이 아닌 클러스터의 결과가 사용자가 준 정의대로 할인되는지를 나타내는 것이다. 이를 위해 본 논문에서는 클러스터링 알고리즘으로 K-means 알고리즘을 사용하였으며, 또한 두 개의 자료의 독립성 검정을 위해 x^2 함수를 변형한 Cramer의 V2를 적합도 함수로 사용하였다.

조회 한 결과 이상의 주요한 결과가 완찰되어 범주별의 상호 독립성(혹은 상호연관성) 여부를 결정하는 방식으로 유의미한 x^2^ 독립성 검정이 있다.

기본적으로 x^2 독립성 검정은 두 범주가 관찰된 빈도수의 그 값이 대응되는 기대값을 비교하는 것 기본적으로 파란. Cramer's V2는 범주가 갈수록 두 범주의 서로 독립성을 말할 수 있다. 그러나, x^2의 값은 데이타의 수에 비례하여 증가하므로 x^2 값 자체만으로는 두 범주가 얼마나 연관성이 있는지를 알 수 없다. Cramer의 V2는 같은 두 개의 범주가 상호 중복되어 있는 100로 가까운 값을, 아무런 연관성이 없이 독립적이며 0에 가까운 값을 갖는다. 따라서 범주 A를 (표준화된) 클러스터 정보로 범주 B를 클러스터링 결과로 보았을 때, 클러스터링 결과(A)는 클러스터 정보(B)에 의한 결과는 두 범주의 간의 관계를 명확히 알 수 있다. 100 가까운 값을 갖게되면, 클러스터링 결과가 사용자가 준 정의대로 되어 있지 않은 경우 두 범주는 서로 독립적이라고 볼 수 있다. 0에 가까운 값을 갖게 된다.

3.4 주제생성

다음 세대의 게임개발을 촉진하기 위한 방향으로 플레임(roulette wheel) 방법을 엘리트(elitist) 보존 방법을 사용한다.

4. 성능 및 분석

4.1 베타바이 샘플링(클러스터 정보가 없는 경우)

4.1.1 아이러니 데이터 아이러니 데이터(iris Data)


5. 결론 및 향후 연구 방향

본 논문에서는 다양한 데이터를 효율적으로 처리하기 위해 사용된 아이러니가 의무적 가중치 방법인 사용자를 취함으로써 클러스터 정보가 없는 데이터에 대해서도 사용자에게 적합한 결과를 제공할 수 있는 알고리즘을 개발하였다. 이는 데이터 분석과 분석을 위한 새로운 접근 방식을 제공하는 것으로, 향후 연구의 방향을 제시하는 것이 필요하다. 또한, 이러한 접근 방식이 다양한 분야에서 활용될 수 있도록 해야 할 것이다.