비형식의 굴집 유효화 지수의 분석과 새로운 지수 개발

김민호, R.S. Ramakrishna
광주과학기술원 정보통신공학과
{mhkim0, rsr}@gist.ac.kr

Analysis and New Indices of Cluster Validity Indices in Ratio Type

Minho Kim, R.S. Ramakrishna
Dept. Information and Communications, Gwangju Institute of Science and Technology (GIST)

요 약

군집 유효화 평가는 굴집화 알고리즘을 진정한 의미의 비감독 학습이 가능하도록 만드는 의미에서 그 중요성이 대단히 크다. 본 논문에서는 이 굴집 유효화 평가에 일반적으로 이용되는 굴집 유효화 지수들의 설계 원리를 분석하고 기존 지수들의 부정적부를 분석한다. 우리는 제 1) 부에서 합 형식의 지수들을 다루었으나, 본 논문에서는 비형식의 지수를 다룬다. 합형식의 CVI에서처럼 지역 패턴의 문제를 해결하였으며, 또한 부적절한 비형식의 지수들의 성능을 항상시킬 수 있는 새로운 기법을 제시한다. 새로운 지수들의 성능은 실험 학습을 통해 제시한다.

1. 서론

많은 굴집화 알고리즘의 결과 품질은 굴집화의 데이터 집합의 특성과 입력 변수에 대한 결정적인 영향을 받는다. 이것은 입력의 정보화 알고리즘에 대한 특징 데이터 집합의 특성을 제대로 고려하지 않은 부적합한 입력 변수를 사용하면 실제 데이터 집합과 다른 굴집 결과를 낼 수 있음을 의미한다. 즉, 어떠한 입력 변수가 실제 데이터 집합에 가장 적합한 굴집 결과를 낼든지 찾아내기 위해서는 굴집 결과 자체를 실제 데이터 집합에 비추어 최적성을 평가하는 기술이 요구되며, 굴집 유효화 지수 (Cluster Validity Index, CVI)가 바로 이러한 굴집 결과의 평가에 널리 사용되고 있다 [1] [2] [3] [4] [5].

제 1) 부에서는 ‘합형식의 CVI’에 대한 기본 설계 원리와 기존 CVI들의 부정적성을 분석하였다. 이 때, 굴집 내부 거리 (Distance Within a Cluster, $dW$) 계산에 있어 평균화로 인한 변형 효과의 문제점은 제기하였으며 대안을 제시하였다. 하지만, 합형식의 CVI 들은 $dW$ 와 $dB$ (군집 사이 거리, Distance Between Clusters)의 결합합 때 발생하는 가중치화 문제점은 가지고 있다.


2. 기존의 CVI

비형식을 취하는 CVI는 그 이름에서도 알 수 있듯이 $dW$와 $dB$의 비를 통해서 얻어진다. 대표적인 CVI로는XB [6]를 들 수 있다. XB의 정의는 Eq. (1)과 같다. XB는

$$XB(nc) = \frac{\sum_{i,j=1}^{N} u^2_w(x_i, c_i) + \sum_{i=1}^{N} u_i^2 d(c_i, c_j)^2}{N \cdot \min_{i,j} d(c_i, c_j)^2}$$

3. 분석과 새로운 CVI들

비형식을 취하는 CVI는 다음과 같은 가정에서 설계되었다:

1) $dW$는 $nc = nc_{optimal}$에서 $nc = nc_{optimal} - 1$로 감소할 때 값의 급격한 증가가 발생한다.
2) $dB$는 $nc = nc_{optimal} + 1$에서 $nc = nc_{optimal}$로 감소할 때
한국컴퓨터종합학술대회 2005 논문집 Vol. 32, No. 1(B)

4. 실험 결과


실험에 사용된 데이터 집합은 총 4개이며, Fig. 2에서는 Dataset 4를 보여주고 있다(17개 군집). Dataset 1~3은 ‘제 (i) 부: 협향식의 CVI’ 논문에 활용할 수 있다.

Fig. 2. 합성 데이터 집합 (Dataset 4)

다음 Dataset 3에 대한 XB와 XB+의 비교 실험을 살펴보자 (Fig. 3). Fig. 3 (과 Fig. 4)의 그레프들은 세 그래프 (dW, dβ, XB)를 그림 하나에 모두 집중시키기 위해 음직을 조절하였는데, 이것은 원본 그래프와 비교하여 그래프의 원활한 해석을 위한 것이며 실제 정량화 기준을 방해하지는 않는다. 이번 실험에서 XB+ 역시 dW의 평균화의 문제를 효과적으로 처리하고 있음을 보여주고 있다. dW와 dβ 그래프 상으로는 앞에서 언급한 이유로 모양이 바뀌지 않았고, 실험에서 nc = 13 (nCoptimal)에서 nc = 12로 변화할 때 XB의 경우 12.8%의 증가가 발생했으며 반면 XB+의 경우 34.2%의 증가가 발생했다. 이로도 제한성을 가진 경우에는 XB와 XB+에 대해 각각 11.4%와 25.5%의 감소가 있었다. 즉, 제한된 방법에 의해 nc = 13에서 dW 이 다른 곳보다 상대적으로 작은 값을 가지므로 XB+ dW < dβ 가 되어,XB가 최소값을 가능하도록 하고 있음을 알 수 있다.

다음 실험은 Dataset 4에 대해 XB와 XB+의 비교한 것이다. Fig. 4에 이를 보이고 있다. Fig. 4 (a)에서 보여 nc = 13 (nCoptimal)에서 Fig. 1에서와 같은 dW와 dβ의 상대적인 급격한 변화가 발생했음에도 불구하고 XB+가 최소값을 가지는데 실패했다. 그 이유는 이러한 평가표 nc < nCoptimal에서 여러 번 발생하였으며 그들 중 하나 (nc = 4)가 nc = 13 일 때 보다 더 작은 값을 가지고 있기 때문이다. 다시 말해 실제세의 데이터 집합은 Fig. 1의 이상적인 경우처럼 단 한 번만 평가표를 보유하는 것이 아닌을 알 수 있다. 다른 데이터집합에서도 다른 문제를 확인할 수 있다. XB+는 maxDiff를 이용하여 XB의 보정한 것인데 실제 월안과 동일하게 nc < nCoptimal에서 큰 값을 가질수록 dW (nc) > dβ (nc)로 만들어 이 문제점을 해결할 것을 볼 수 있다.
표 1. 각 CVI에 의해 제안된 구조수와 성공여부

<table>
<thead>
<tr>
<th></th>
<th>( \nu_{p1} )</th>
<th>( \nu_{p1}^* )</th>
<th>SD</th>
<th>SD*</th>
<th>XB</th>
<th>XB*</th>
<th>XB**</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dataset</td>
<td>15</td>
<td>18</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>1</td>
<td>(X)</td>
<td>(X)</td>
<td>(O)</td>
<td>(O)</td>
<td>(O)</td>
<td>(O)</td>
<td>(O)</td>
</tr>
<tr>
<td>Dataset</td>
<td>13</td>
<td>13</td>
<td>7</td>
<td>13</td>
<td>13</td>
<td>13</td>
<td>13</td>
</tr>
<tr>
<td>2</td>
<td>(O)</td>
<td>(O)</td>
<td>(X)</td>
<td>(O)</td>
<td>(O)</td>
<td>(O)</td>
<td>(O)</td>
</tr>
<tr>
<td>Dataset</td>
<td>11</td>
<td>13</td>
<td>6</td>
<td>11</td>
<td>13</td>
<td>13</td>
<td>13</td>
</tr>
<tr>
<td>3</td>
<td>(X)</td>
<td>(O)</td>
<td>(X)</td>
<td>(X)</td>
<td>(O)</td>
<td>(O)</td>
<td>(O)</td>
</tr>
<tr>
<td>Dataset</td>
<td>15</td>
<td>16</td>
<td>7</td>
<td>4</td>
<td>4</td>
<td>17</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>(X)</td>
<td>(X)</td>
<td>(X)</td>
<td>(X)</td>
<td>(X)</td>
<td>(X)</td>
<td>(O)</td>
</tr>
</tbody>
</table>

5. 결론

본 논문에서는 ‘제 (1) 부: 합형식의 CVI’에 이어 비형식의 CVI에 대한 기본 설계 원리 분석하고 분석을 통해 발전한 이들의 성과를 보다 뛰어난 결과를 보여 주는 다양한 CVI들로 제시하였다. 대표적인 비형식의 CVI는 XB에 대해서는 기본 합형식의 CVI로서의 \textit{Low-Pass Filtering}으로 인한 문제점을 확인하였다. 이 문제점을 \( \nu_{p1}^* \), SD*와 유사한 방법으로 XBB*를 정의함으로써 해결하였다. 또한, 복잡한 데이터 집합에 대해서도 비형식 CVI들의 설계 원리 효율적으로 보완할 수 있는 \textit{maxDiff}를 제안하였으며, 이것을 XB*에 적용하여 XB**를 제안하였다. 실험 결과에서 합형식의 CVI들보다 비형식의 CVI들이 더 좋은 성능을 보여 주었으며, 그 중에서도 XB**가 가장 좋은 결과를 보여주었다. ‘제 (1) 부: 합형식’의 논문과 더불어 본 논문에서 제시한 설계 원리 및 보완 기법은 이들 논문에서 제시한 CVI들 뿐만 아니라 다른 CVI에 대해서도 적용이 가능할 것으로 기대되며, 향후 이에 대한 연구를 보장하고자 한다.

[참고문헌]


