http://dx.doi.org/10.11627/jkise.2012.35.4.244 pISSN 2005-0461
eISSN 2287-7975

F_n-Measure : An External Cluster Evaluation Measure

Kyeongtaek Kim†

Department of Industrial and Management Engineering, Hannam University

클러스터 평가 외부기준 척도 F_n-Measure

김 경 택†

한남대학교 산업경영공학과

F-Measure is one of the external measures for evaluating the validity of clustering results. Though it has clear advantages over other widely used external measures such as Purity and Entropy, F-Measure has inherently been less sensitive than other validity measures. This insensitivity owes to the definition of F-Measure that counts only most influential portions. In this research, we present F_n-Measure, an external cluster evaluation measure based on F-Measure. F_n-Measure is so sensitive that it can detect their difference in the cases that F-Measure cannot detect the difference in clustering results. We compare F_n-Measure to F-Measure for a few clustering results and show which measure draws better result based upon homogeneity and completeness

Keywords : External Clustering Measure, F-Measure, F_n-Measure

1. 서 론

객체(object)를 오버랩이 허용되지 않는 여러 개의 클러스터(cluster)로 나누는 클러스터링(clustering)에서는 같은 클러스터내의 객체들과의 유사성이 다른 클러스터에 있는 객체들과의 유사성보다 크도록 클러스터를 형성한다. 클러스터링을 지도(supervision)의 유무 및 정도에 따라 분류하면 자율(unsupervised)클러스터링, 반지도(semi-supervised) 클러스터링, 지도(supervised) 클러스터링으로 분류된다. 클러스터링은 전통적으로 자율학습(unsupervised learning) 도구의 하나로 여겨져 왔다. 자율 클러스터링은 특정 목적함수를 최적화하기 위하여 학습 프레임웍을 사용한다. 목적함수의 예로는, 각 객체로부터 클러스터 중심까지의 거리의 합의 최소화를 들 수 있다. 지도 클러스터링에서는 각 클러스터마다 어느 한 클래스(class)에 속한 객체가 높은 확률로 존재하도록 클러스터링 하는 것을 목적으로 한다는 점에서 전통적인 자율 클러스터링과 다르다[1]. 반지도 클러스터링에서는 일부의 미리 분류된 객체들이 존재하며, 클러스터링 과정에서 서로 다른 클래스에 속한 객체들은 서로 다른 클러스터에 속해야 하며, 이들 정보를 이용하여, 분류되지 않은 객체들을 클러스터링 한다.

객체가 주어지면, 각 클러스터링 알고리즘은 실제로 그 구조가 존재하는지 여부에 관계없이 일련의 클러스터를 결과로서 산출한다. 서로 다른 알고리즘은 일반적으로 서로 다른 클러스터 집합을 산출한다. 게다가, 동일한 알고리즘을 적용하는 경우에도, 파라메터 값의 변경이나 객체의 입력 순서가 최종 결과에 영향을 미칠 수 있다. 따라서, 사용된 알고리즘으로부터 얻은 결과의 적정성을 제시하기 위하여 결과에 대한 효과적인 평가 기준이나

Received 15 November 2012; Finally Revised 11 December 2012; Accepted 13 December 2012
† Corresponding Author : kkim610@gmail.com

평가 척도가 필요하다. 이러한 평가는 객관적이어야 하며, 특정 알고리즘에 편향되어서는 않아야 한다. 평가척도 값을 제시함으로써, 몇 개의 클러스터로 클러스터링하는 것이 타당한지, 또는 왜 특정 알고리즘을 선택하였는지 등에 관하여 설명할 수 있다[2].

대다수의 클러스터링 알고리즘의 성능은 객체의 특성과 입력 파라메터에 의해 크게 좌우된다. 객체는 주어지므로, 객체의 특성은 주어진 조건이 되어 변경 불가능한 사항이 된다. 부적절한 입력 파라메터는 객체들의 특성에서 벗어난 클러스터를 형성시킨다. 주어진 객체들에 가장 적합한 클러스터를 만드는 입력 파라메터를 결정하기 위하여, 클러스터를 평가하는 신뢰성 있는 지표가 필요한데, 클러스터 타당성 인덱스(cluster validity index)가 이 역할을 하며, 이를 사용하여 클러스터의 타당성을 평가한다.

클러스터 타당성은 크게 두 가지 접근방법으로 분류한다[3]. 첫 번째 방법은 내부기준(internal criteria)에 근거한 접근법이다. 이 방법에서는 근접성 행렬처럼 객체의 벡터들에 포함된 정보를 이용하여 클러스터링 알고리즘의 결과를 평가한다. 두 번째는 외부기준(external criteria)에 근거한 접근법이다. 이 방법은 객체에 미리 명시된 클래스(또는 카테고리)에 근거하여 클러스터링 알고리즘의 결과를 평가하는 것을 의미한다. 이 클래스는 클러스터링 알고리듬의 입력정보로 사용되지 않는다.

외부기준에 의한 타당성 평가 방법으로 Purity와 Entropy에 근거한 척도가 사용되어 왔다[4]. 그러나, 이 방법들은 주어진 클래스의 모든 멤버들이 하나의 클러스터에 포함되었는지 여부를 측정할 수 없다는 단점이 지적되었다[5]. 그래서, 이들의 단점을 극복한 F-Measure는 널리 쓰였다. 그러나, 최근에는 F-Measure의 단점을 지적하는 논문들이 늘고 있는 추세이다[5, 6]. 본 논문에서는, 두 클러스터링 결과에 대한 F-Measure가 같을 경우, 이들 결과의 우열을 보다 세밀히 나타낼 수 있는 척도를 제시한다.

2. 외부기준의 방법들

2.1 평가 척도를 평가하는 기준

내부 기준의 평가에서는 객체들의 원래 클래스(또는 카테고리)를 알 수 없다. 내부 기준에 근거한 타당성 인덱스는 다음과 같은 두 평가 기준의 결합에 의해 정의된다[7].

(1) 밀집성(compactness) : 이것은 같은 클러스터를 형성하는 데이터들의 상호 근접도(closeness)을 나타낸다. 전형적인 예가 분산(variance)이다. 분산은 멤버들이 얼마나 다른가를 나타내지만, 분산이 작다는 것은 상호 근접성의 지표이다.

(2) 분리성(separability) : 이것은 클러스터들이 얼마나 뚜렷이 구별되는 지를 클러스터들 사이의 거리로써 나타낸다.

한편, 외부기준에 의한 타당성 평가에서는, 동일한 클래스의 데이터들이 모두 한 클러스터로 클러스터링 된다면 가장 완벽한 해(solution)가 된다. 따라서, 완벽한 해를 이미 알고 있으므로, 클러스터링 알고리듬의 결과와 비교하여 얼마나 차이가 나는지를 나타내는 타당성 인덱스를 사용하여야 한다. 외부기준에 의한 클러스터링 결과의 타당성을 평가하는 기준으로 다음 두 기준이 제시되고 있다[5, 6, 8].

(1) 동질성(homogeneity) : 각 클러스터의 모든 객체들이 동일한 클래스로부터 온 객체들일 때, 클러스터링 결과는 동질성을 만족시킨다.

(2) 완전성(completeness) : 각 클래스의 모든 객체들이 동일한 클러스터의 멤버가 될 때, 클러스터링 결과는 완전성을 만족시킨다.

2.2 외부 기준의 방법들

Melia[6]는 외부기준의 평가척도를 counting pairs 방법과 set matching 방법으로 분류하였다. Reichart와 Rappoport[8]는 mapping 기준 방법, counting pairs 방법, Information theoretic 방법으로 분류 하였다. Amigo et al.[9]은 counting pairs 방법, set matching 방법, 엔트로피에 의한 방법으로 분류하였다. Wu et al.[10]은 contingency table 방법, 엔트로피와 purity 방법, mutual information과 variation of information 방법, micro-average precision 방법으로 분류하였다.

이러한 분류에서 공통적으로 언급되는 외부기준 척도로 Purity와 Entropy가 있다. Purity와 Entropy는 동질성을 측정하는 좋은 척도이다[5, 8, 10]. 동질성이 증가하면 Purity는 증가하고, Entropy는 감소한다. 그러나, Purity와 Entropy는 완전성을 측정하지 못한다는 단점을 가지고 있다[5, 8]. 널리 쓰이는 또 다른 외부 기준 척도로 F-Measure[11]가 있다. F-Measure는 동질성과 완전성을 모두 측정한다는 점에서 Purity나 Entropy에 비하여 비교우위를 가지고 있다[5].

3. 새로운 클러스터링 척도

C를 평가할 클러스터들의 집합이라 하고, L을 클래스들(또는 카테고리들)의 집합이라 하자. 각 클러스터에 대한 정밀도(precision)는 다음과 정의된다.

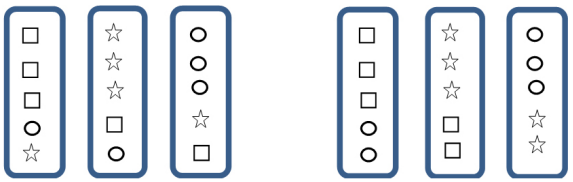
$$\text{Precision}(C_i,L_j)=\frac{|C_i\cap L_j|}{|C_i|}$$

정밀도는 클러스터의 객체들 중에 동일한 클래스에서 온 객체의 비율을 나타내며, 모든 객체들이 동일한 클래스로부터 온 객체들일 때 1값을 갖게 된다. Precision은 외부기준의 평가척도의 타당성을 판단하는 기준 중 동질성을 표현하는 지표이다.

정밀도를 이용하는 외부기준의 평가척도로 Purity가 있으며 그 정의는 다음과 같다.

$$\text{Purity}=\sum_i\frac{|C_i|}{n}max\ \text{Precision}(C_i,L_j)$$

한편, 각 클래스에 대한 Recall은 다음과 같이 정의된다.



F-Measure = 0.6

<Figure 1> Clustering A and Clustering B[5]

$$\text{Recall}(C_i,L_j)=\frac{|C_i\cap L_j|}{|L_j|}$$

Recall은 같은 클래스에 속한 객체들이 같이 그룹핑 되는 비율을 나타내며, 같은 클래스에 속한 객체 모두가 동일한 클러스터로 그룹핑 되는 경우 1값을 갖게 된다. Recall은 외부기준의 평가척도의 타당성을 판단하는 기준 중 완전성을 표현하는 지표이다.

Recall을 이용하는 외부기준의 평가척도로 역순수도(inverse purity)가 있으며 그 정의는 다음과 같다.

$$\text{Inverse Purity}=\sum_j\frac{|L_j|}{n}max\ \text{Recall}(C_i,L_j)$$

Larson과 Aone[11]은 Precision과 Recall의 조화평균의 최대값을 이용하여 F-Measure를 다음과 같이 정의하였다.

$$\text{F}(C_i,L_j)=\frac{2\times\text{Recall}(C_i,L_j)\times\text{Precision}(C_i,L_j)}{\text{Recall}(C_i,L_j)+\text{Precision}(C_i,L_j)}$$

$$\text{F}-\text{Measure}=\sum_j\frac{|L_j|}{n}max_i\ \{F(C_i,L_j)\}$$

앞에서도 언급 하였듯이 Precision은 동질성을 표현하고, Recall은 완전성을 표현하므로, 이들의 조화평균을 사용하는 F-Measure는 동질성과 완전성을 모두 측정하는 척도가 된다. 따라서, 동질성만을 표현하는 척도인 Purity나 Entropy보다 우수한 척도라 할 수 있다.

그러나, F-Measure도 그 정의로 인한 약점을 가지고 있다. F-Measure는 각 클래스 L_j에 대하여 $\max_i\{F(C_i,L_j)\}$ 값을 갖는 i에 대한 $\text{F}(C_i,L_j)$값들만이 기여를 하므로, $\max_i\{F(C_i,L_j)\}$ 값을 갖지 않는 $\text{F}(C_i,L_j)$의 변화에 둔감할 수 밖에 없다. 이에 따라, <Figure 1>에서 보듯이 객관적으로 두 클러스터링 결과의 우열이 분명한데도, $\max_i\{F(C_i,L_j)\}$값은 변화가 없으므로, F 값의 변화가 없는 경우가 발생한다. 이러한 경우, 각 클래스 L_j에 대하여 $\max_i\{F(C_i,L_j)\}$ 값을 갖는 i에 대한 $\text{F}(C_i,L_j)$값 뿐만 아니라, 그 다음 최대치를 갖는 $\text{F}(C_i,L_j)$까지 계산한다면, 차이에 보다 더 민감한 척도가 될 것이다.

따라서, 본 논문에서는 F-Measure의 정의를 확장하여 다음과 같은 새로운 클러스터링 타당성 척도를 제시한다.

클래스 L_j에 대하여 $\max_i\{F(C_i,L_j)\}$값을 갖는 i를 i^*라 하자. 새로운 척도 $\triangle\text{F}_2$-Measure 및 F_2-Measure를 다음과 같이 정의한다.

$$\triangle\text{F}_2-\text{Measure}=\sum_j\frac{|L_j|}{n}max_{i\neq i^*}\{F(C_i,L_j)\}$$

$$\text{F}_2-\text{Measure}=F+\alpha\triangle F_2$$

$$=\sum_j\frac{|L_j|}{n}max_i\{F(C_i,L_j)\}+\alpha\sum_j\frac{|L_j|}{n}max_{i\neq i^*}\{F(C_i,L_j)\}$$

α값은 0과 1사이의 값을 취한다. α 값은 각 클래스 L_j에 대하여 $\max_i\{F(C_i,L_j)\}$값의 반영비율이 1일때 $\max_{i\neq i^*}\{F(C_i,L_j)\}$의 반영 비율을 의미한다. $\alpha=0$이면, F_2-Measure는 F-Measure와 동일하게 된다. $\alpha=1$이면, F_2-Measure는 F-Measure와 $\triangle F_2$-Measure의 합과 동일하게 된다.

F_2-Measure를 정의하는 방법과 유사한 방법으로 F_3-Measure, F_4-Measure, …, F_n-Measure를 정의할 수 있다.

4. 사례 연구

앞장에서 제시한 새로운 인덱스가 유용하게 사용되는

경우를 살펴보자. <Figure 1>에 있는 Clustering A와 Clustering B의 동질성과 완전성을 비교해보면 <Table 1>과 같다. 따라서 Clustering B가 Clustering A보다 동질성과 완전성 양면에서 더 우수하다.

Clustering A의 Purity와 F-Measure값은 다음과 같다.

$$\text{Purity}=\frac{5}{15}\times\frac{3}{5}+\frac{5}{15}\times\frac{3}{5}+\frac{5}{15}\times\frac{3}{5}=0.6$$

$$\text{F}-\text{Measure}=\frac{5}{15}\cdot\max\left(\frac{3}{5},\frac{1}{5},\frac{1}{5}\right)+\frac{5}{15}\cdot\max\left(\frac{3}{5},\frac{1}{5},\frac{1}{5}\right)$$
$$+\frac{5}{15}\cdot\max\left(\frac{3}{5},\frac{1}{5},\frac{1}{5}\right)=0.6$$

<Table 1> Homogeneity and Completeness

	Clustering A	Clustering B
Homogeneity	Each class consists of objects classified as 3 clusters.	Each class consists of objects classified as 2 clusters.
Completeness	Each cluster consists of objects from 3 classes.	Each cluster consists of objects from 2 classes.

동일한 방법으로 계산하면, Clustering B의 Purity는 0.6, F-Measure 값은 0.6이 된다. Purity 값이 동일한 이유는 각 클러스터마다 가장 큰 비중을 차지하는 객체들이 그 클러스터에서의 비율이 Clustering A와 Clustering B에서 모두 동일하기 때문이다. 이와 유사하게, F-Measure 값이 동일한 이유는 각 클래스에 대한 $\max_i\{F(C_i,L_j)\}$ 값이 Clustering A에서와 Clusterting B에서 모두 동일하기 때문이다.

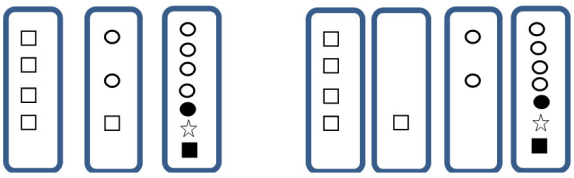
한편, 두 클러스터링에 대한 $\triangle F_2$-Measure를 구하면 다음과 같다.

Clustering A : $\triangle F_2$-Measure = 0.21
Clustering B : $\triangle F_2$-Measure = 0.42

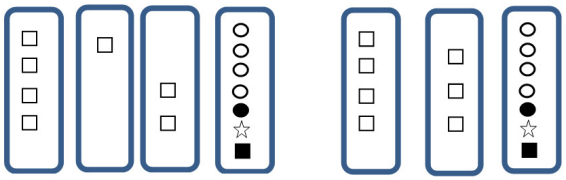
이를 바탕으로, 두 클러스터링에 대한 F_2-Measure를 구하면 다음과 같다.

($\alpha=1$일 때)
Clustering A : F_2-Measure = 0.81
Clustering B : F_2-Measure = 1.02

따라서, Clustering B가 Clustering A보다 낫다는 결론에 도달하며, 이는 <Table 1>에서 제시한 동질성 및 완전성 양면에서 Clustering B가 Clustering A보다 낫다는 사실과 부합한다. $0<\alpha\leq 1$이면,



<Figure 2> Clustering C and Clustering D[9]



<Figure 3> Clustering E and Clustering F[9]

Clustering B의 F_2-Measure 값이 Clustering A의 F_2-Measure값보다 항상 크다.

<Figure 2>에서 Clustering D가 Clustering C보다 동질성과 완전성 양면에서 더 우수하다. 그러나, 이 경우에도 F-Measure 값은 동일하여, 이러한 사실을 제대로 반영하고 있지 않다. 만일 대한 F_2-Measure를 사용한다면, $\alpha>0$이면, Clustering D의 F_2-Measure 값이 Clustering C의 F_2-Measure 값보다 항상 크다. 따라서, Clustering D가 Clustering C보다 낫다는 결론에 도달한다.

<Figure 3>의 Clustering E와 Clustering F의 경우에, Purity나 F-Measure 값을 사용하면 두 클러스터링의 우열을 구별할 수 없다. 그러나, F_2-Measure를 사용한다면, Clustering F의 F_2-Measure 값이 Clustering E의 F_2-Measure 값보다 크게 되어, Clustering F가 Clustering E보다 낫다는 결론에 도달한다.

5. 결 론

본 연구에서는 클러스터링을 평가할 때 사용되는외부 기준 인덱스중의 하나인 F-Measure가 갖는 약점을 보완하기 위한 새로운 인덱스 F_n-Measure를 제안하였다. F-Measure로는 우열을 가릴 수 없는 경우에 제안된 인덱스 F_n-Measure에 속하는 $\triangle F_2$-Measure를 사용하여 우열을 가릴 수 있음을 그림과 함께 제시하였다.

F_n-Measure를 적용하여 클러스링의 우열을 가린 결과가 F_m-Measure에서는 동일하게 유지 되지 않을 수 있다 (단 $n\neq m$). 그러나, 이것은 관점의 차이로, 클러스터링의 우열을 가릴 때에, 동질성과 완전성에 몇 번째로 큰 영향을 끼치는 객체들까지 살펴볼 것이냐에 따른 것이

다. 그러나, F_n-Measure를 적용해서 클러스터링의 우열을 구분할 수 없을 때 F_{n+1}-Measure를 사용한다면, 항상 결론의 일관성을 유지할 수 있다.

Acknowledgement

This study has been partially supported by a 2012 Research Fund of Hannam University, Korea.

References

[1] Xu, L., Mo, H., and Wang, K., Immune Algorithm for Supervised Clustering. *Proceedings of the 5th IEEE International Conference on Cognitive Informatics*, 2006, p 953-958.

[2] Xu, R. and Wunsch, D. II, Survey of Clustering Algorithms. *IEEE Transactions on neural Networks*, 2005, Vol. 16, No. 3, p 645-678.

[3] Halkidi, M., Batistakis, Y., and Vazirgiannis, M., *Cluster Validity Methods : Part I, SIGMOD Record*, 2002, Vol. 31, No. 2, p 40-45.

[4] Zhao, Y. and Karypis, G., Criterion functions for document clustering : Experiments and Analysis. *Technical Report TR 01-40, Dept. of Computer Science*, U. of Minnesota, 2001.

[5] Rosenberg, A. and Hirschberg, J., V-Measure : A Conditional Entropy-based External Cluster Evaluation Measure, Proceedings of the 2007 Joint Conf. on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, June 2007, p 410-420.

[6] Melia, M., Comparing Clustering-an Information Based Distance. *J. of Multivariate Analysis*, 2007, Vol. 98, p 873-895.

[7] Berry, M. and Linoff, G., Data Mining Techniques for Marketing, Sales and Customer Support, John Wiley and Sons, 1996.

[8] Reichart, R. and Rappoport, A., The NVI Clustering Evaluation Measure. *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, 2009, p 165-173.

[9] Amigo, E., Gonzalo, J., and Artiles, J., A Comparison of Extrinsic Clustering Evaluation Metrics based on Formal Constraints. *Information Retrieval*, Aug 2009, Vol. 12, No. 4, p 461-486.

[10] Wu, J., Chen, J., Xiong, H., and Xie, M., External Validation measures for K-means Clustering : A Data Distribution perspective. *Expert Systems with Applications*, 2009, Vol. 36, p 6050-6061.

[11] Larson, B. and Aone, C., Fast and Effective Text Mining Using Linear Time Document Clustering. *Proceedings of the Conference on Knowledge Discovery and Data Mining*, 1999, p 16-22.