f-Divergence Measures for Evaluation in Community Detection

Mariam Haroutunian[1], Karen Mkhitaryan[1], and Josiane Mothe[2]

[1] Institute for Informatics and Automation Problems of NAS of RA,
1 P. Sevak, Yerevan, Armenia
iiap.sci.am
[2] IRIT, UMR5505 CNRS & ESPE, Univ. de Toulouse,
118 Route de Narbonne, TOULOUSE CEDEX 9, Toulouse, France
info@irit.fr

Abstract. Community detection is a research area from network science dealing with the investigation of complex networks such as biological, social and computer networks, aiming to identify subgroups (communities) of entities (nodes) that are more closely related to each other than with remaining entities in the network [1]. Various community detection algorithms are used in the literature however the evaluation of their derived community structure is a challenging task due to varying results on different networks. In searching good community detection algorithms diverse comparison measures are used actively [2]. Information theoretic measures form a fundamental class in this discipline and have recently received increasing interest [3]. In this paper we first mention the usual evaluation measures used for community detection evaluation We then review the properties of f-divergence measures and propose the ones that can serve community detection evaluation. Preliminary experiments show the advantage of these measures in the case of large number of communities.

Keywords: Community Detection · Information Theory · f-divergence · distance measures.

1 Introduction

The goal of community detection is to partition the given network into communities to understand its topological structure [1].

When a community detection algorithm is applied and the studied network is partitioned into communities, the output is a n dimensional random vector $X = (x_1, x_2, ..., x_n)$, where n is the number of nodes in the network and each $x_i \in \{1, ..., k\}$ element represents the community assignment of node i, where k is the number of communities $(i \in \{1, ..., n\})$.

In order to quantitatively assess the goodness of the derived network partition, it can either be compared with other partitions of a network or with pre-known ground truth [2, 4].In the literature of the domain it is mostly accomplished by using evaluation measures imported from clustering problems and information theory [2, 4, 5].

From the list of available evaluation measures, the application of information theoretic measures in community detection is more prospective because of their strong mathematical foundation and ability to detect non-linear similarities [3, 7].

Let $X = (x_1, x_2, ..., x_n)$ and $Y = (y_1, y_2, ..., y_n)$ be two different partitions of the network. We assume that community assignments x_i and y_i are values of random variables X and Y respectively with joint probability distribution $p(x, y) = P(X = x, Y = y)$ and marginal distributions $p(x) = P(X = x)$ and $p(y) = P(Y = y)$. Thus calculating the similarity of two network partitions can be viewed as comparing two random variables which is typical an encoding/decoding problem in information theory. Mutual information (MI) is a popular measure in information theory, measuring the mutual dependence of two random variables X and Y. It measures how much information about one random variable is obtained through the other random variable [6]. However MI is not a normalized measure making it unsuitable to quantitatively evaluate and compare different partitions. Several normalized variants of MI called normalized mutual information (NMI) were introduced by Yao [8], Kvalseth [9] and Strehl et al. [10]. Later Meila [11] introduced variation of information (VI) which unlike MI is a metric measure. Finally normalized variation of information (NVI) and normalized information distance (NID) were proposed by Kraskov et al. [12].

In [3], the authors performed an organized study of information theoretic measures for clustering comparison; it has been mathematically proved that NVI and NID satisfy both the normalization and the metric properties. Moreover, it was showed that NID is preferable since it better uses the $[0, 1]$ range. Despite the fact that NMI, NVI, NID have many advantages, some experiments challenge their effectiveness [3, 7].

According to Amelio and Pizzuti [13] normalized mutual information needs adjustment as it has unfair behavior especially when the number of communities in the network is large. The authors suggested to adjust the NMI by introducing a scaling factor which also compares the number of communities detected by an algorithm and the actual number of communities in the ground truth.

Another modification was suggested by Zhang [14] who claims that NMI is affected by systematic errors as a result of finite network size which may result in wrong conclusions when evaluating community detection algorithms. Relative normalized mutual information (rNMI) introduced by Zhang takes into account the statistical significance of NMI by comparing it with the expected NMI of random partitions.

An important class of information theoretic measures are so called f - divergences. These are measures of discrimination between two probability distributions. Their properties, connection inequalities and applications in information theory, machine learning, statistics and other applied branches were studies in many publications, see for example [15–18].

However, they have never been considered as community detection evaluation measures despite their properties that make them good candidate for this task. In this paper we investigate the properties of some f-divergences from community detection evaluation point of view. We think that some of them could serve as a good alternative to existing information theoretic measures in community detection evaluation framework.

The paper is organized as follows: we demonstrate some popular information theoretic measures in Section 2, show some f-divergence measures and discuss their useful properties for considering them in community detection evaluation in Section 3 and conclude in Section 4.

2 Information Theoretic Measures

Various measures are used in community detection problems to evaluate network's partition into communities, which are imported from other disciplines such as cluster analysis and information theory [2, 4, 5].

In recent years information theoretic tools were applied in various fields such as coding theory, statistics, machine learning, genomics, neuroscience etc. [6]. The same tools are also useful when in community detection since they provide a bunch of measures to compare network partitions. One of the basic measures is the mutual information between two random variables, which tells how much knowing one of clusterings reduces the uncertainty about the other. Mutual information of two discrete random variables is defined as [6]:

$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} = H(X) - H(X|Y),$$

where $H(X)$ is the entropy of X and $H(X|Y)$ the conditional entropy of X given Y.

$$H(X) = -\sum_{x} p(x) \log p(x)$$

$$H(X|Y) = -\sum_{x,y} p(x,y) \log p(x|y)$$

Considering random variables as random community distributions of a network, it is used to compare network partitions.

To evaluate and compare community structures, it is highly desired that the measures satisfy two main properties: normalization property and metric property.

– **Normalization property**

A measure is normalized if the range of values it takes fall into a fixed interval. Normalized measures are easy to interpret and in community detection problems it is of paramount importance as it is necessary to quantitatively assess the similarity of a given partition with other partitions or with ground truth. In community detection evaluation most of the measures fall into intervals $[0, 1]$ or $[-1, 1]$.

– **Metric property**

A measure d is a metric if it satisfies the following properties:

– Non-negativity:	$d(x, y) \geq 0$,
– Identity:	$d(x, y) = 0 \Leftrightarrow x = y$,
– Symmetry:	$d(x, y) = d(y, x)$,
– Triangle inequality:	$d(x, y) + d(y, z) \geq d(x, z)$.

The metric property conforms to the intuition of distance [11] and it is important in the case of complex space of clusterings as many theoretical results already exist for metric spaces.

Based on the properties of MI, that is non-negativity and upper boundedness:

$$\begin{aligned}0 \leq I(X \cap Y) \leq \min\{H(X), H(Y)\} \leq \sqrt{H(X)H(Y)} \leq \frac{1}{2}(H(X) + H(Y)) \leq \\ \leq \max\{H(X), H(Y)\} \leq H(X, Y)\end{aligned}$$

several normalized variants of NMI can be considered as similarity measures [3, 8–10]:

$$\mathrm{NMI}_{\mathrm{joint}} = \frac{I(X;Y)}{H(X,Y)}, \qquad \mathrm{NMI}_{\mathrm{max}} = \frac{I(X;Y)}{\max(H(X), H(Y))},$$

$$\mathrm{NMI}_{\mathrm{sum}} = \frac{2I(X;Y)}{H(X) + H(Y)}, \qquad \mathrm{NMI}_{\mathrm{sqrt}} = \frac{I(X;Y)}{\sqrt{H(X)H(Y)}},$$

$$\mathrm{NMI}_{\mathrm{min}} = \frac{I(X;Y)}{\min\{H(X), H(Y)\}}.$$

Based on the five upper bounds for $I(X;Y)$ also five distance measures can be defined as follows.

$$\begin{aligned}
D_{\mathrm{joint}} &= H(X,Y) - I(X;Y), \\
D_{\mathrm{max}} &= \max\{H(X), H(Y)\} - I(X;Y), \\
D_{\mathrm{sum}} &= H(X) + H(Y) - 2I(X;Y), \\
D_{\mathrm{sqrt}} &= \sqrt{H(X)H(Y)} - I(X;Y), \\
D_{\mathrm{min}} &= \min\{H(X)H(Y)\} - I(X;Y).
\end{aligned}$$

$D_{\text{joint}} = 2D_{\text{sum}}$ is known as variation of information (VI) introduced by Meila [11], it satisfies the properties of metrics but not the one of normalization. In [3] it was proved that $D_{\max}$ is a metric, while $D_{\min}$ and D_{sqrt} are not metrics.

Later Kraskov et al. [12] introduced normalized variant of variation of information called normalized variation of information (NVI) and normalized information distance (NID) which are both normalized and metric measures.

$$\text{NVI} = \frac{H(X,Y) - I(X;Y)}{H(X,Y)} = 1 - \frac{I(X;Y)}{H(X,Y)}$$
$$\text{NID} = \frac{\max(H(X),H(Y)) - I(X;Y)}{\max(H(X),H(Y))} = 1 - \frac{I(X;Y)}{\max\{H(X),H(Y)\}}$$

Vinh et al. [3] proved that NVI and NID are metrics.

3 f-divergences and some useful properties

Definition:
Let $f : (0,\infty) \to R$ be a convex function with $f(1) = 0$ and let P and Q be two probability distributions. The f-divergence from P to Q is defined by

$$D_f(P \parallel Q) \triangleq \sum_x q(x) f(\frac{p(x)}{q(x)}).$$

Among others, f-divergences include well known notions from information theory listed below.

Kullback-Leibler divergence which is also known as relative entropy

$$D(P \parallel Q) = \sum_x p(x)\log(\frac{p(x)}{q(x)}),$$

is a f-divergence with $f(t) = t\log(t)$. Also $D(Q \parallel P)$ can be obtained from $f(t) = -t\log(t)$.

Total variational distance

$$V(P,Q) = \sum_x |p(x) - q(x)| = \sum_x q(x)|\frac{p(x)}{q(x)} - 1|,$$

is coming from the same f-divergence formula when $f(t) = |t-1|$.

Hellinger distance defined by

$$H(P,Q) = \sum_x (\sqrt{p(x)} - \sqrt{q(x)})^2,$$

is a f-divergence with $f(t) = (\sqrt{t} - 1)^2$. The Hellinger distance is closely related to the total variational distance, but it has several advantages such as being well suited for the study of product measures.

Jeffrey divergence is the symmetrisized Kullback-Leibler divergence

$$J(P \parallel Q) = D(P \parallel Q) + D(Q \parallel P) = \sum_x (p(x) - q(x)) \log(\frac{p(x)}{q(x)}),$$

that is obtained from $D_f(P \parallel Q)$ with $f(t) = \frac{1}{2}(t-1)\log(t)$.

Capacitory discrimination (also known as Jensen-Shannon divergence) is given by

$$C(P,Q) = D(P \parallel \frac{P+Q}{2}) + D(Q \parallel \frac{P+Q}{2}) = 2H(\frac{P+Q}{2}) - H(P) - H(Q)$$

which comes from $D_f(P,Q)$ with $f(t) = t\log(t) - (t+1)\log(t+1) + 2\log(2)$.

χ^2 ***divergence*** is a f-divergence measure,

$$\chi^2(P,Q) = \sum_x \frac{(p(x) - q(x))^2}{q(x)} = \sum_x q(x)(\frac{p(x)}{q(x)} - 1)^2,$$

where $f(t) = (t-1)^2$.

Bhattacharyya distance given by

$$d(P,Q) = \sqrt{1 - \sum_x \sqrt{p(x)q(x)}},$$

can be obtained from $D_f(P,Q)$, when $f(t) = 1 - \sqrt{t}$.

We considered the properties of above mentioned measure to decide how they can be applied for community detection evaluation. In fact, to compare two algorithms with network partitions X and Y we must evaluate the discrimination from P_{XY} to $P_X P_Y$.

First note that there is a well known property

$$D(P_{XY} \parallel P_X P_Y) = I(X;Y)$$

and hence **Kullback-Leibler divergence** being very useful in information theory is not interesting for our task.

It is obvious that the **total variational distance** $V(P,Q)$ takes values from interval $[0,2]$

$$0 \leq V(P,Q) \leq 2$$

and hence is normalized. It is proved that $V(P,Q)$ satisfies all metric properties. Consider,

$$V(P_{XY}, P_X P_Y) = \sum_{x,y} |p(x,y) - p(x)p(y)|,$$

it equals 0 when X and Y are independent, which means that as small is the variational distance as far are the two clusterings.

For **Hellinger distance** $H(P, Q)$ the following property

$$0 \leq H(P, Q) \leq V(P, Q)$$

shows that it is normalized too. It is also proved that $\sqrt{H(P, Q)}$ is a true metric.

We are interested in

$$\sqrt{H(P_{XY}, P_x P_Y)} = \sqrt{\sum_{x,y} (\sqrt{p(x, y)} - \sqrt{p(x)p(y)})^2},$$

which as in the previous case tends to zero when X and Y are independent. **Capacitory discrimination** $C(P, Q)$ satisfies the following inequality

$$0 \leq C(P, Q) \leq V(P, Q),$$

thus taking values in $[0, 2]$. It is proved that $\sqrt{C(P, Q)}$ satisfies the metric properties [19]. We shall consider the following measure

$$\sqrt{C(P_{XY}, P_X P_Y)} = \sqrt{D(P_{XY} \parallel \frac{P_{XY} + P_X P_Y}{2}) + D(P_X P_Y \parallel \frac{P_{XY} + P_X P_Y}{2})},$$

that can be used to compare clusterings as in the two previous cases.

For **Bhattacharyya distance** the following inequality is known

$$0 \leq d \leq 1,$$

and it is proved to be a metric. In this case

$$d(P_{XY}, P_X P_Y) = \sqrt{1 - \sum_{x,y} \sqrt{p(x, y)p(x)p(y)}},$$

being equal to 0 also when X and Y clusterings are independent.

Thus Total variational distance, Bhattacharyya distance, Hellinger distance and Capacitory discrimination are good candidates for community detection evaluation as they satisfy both normalization and metric properties.

4 Conclusion and future work

Researching information-theoretic measures and their properties we suggest Total variational distance, Bhattacharyya distance, Hellinger distance and Capacitory discrimination as promising candidates for evaluation tasks in community detection. In future we plan to investigate, evaluate and compare them on both real world and synthetic networks which may highlight the strong connections of f-divergences and community detection.

References

1. S. Fortunato: Community detection in graphs. Physics Reports **486**, 75–174 (2010)
2. J. Mothe, K. Mkhitaryan and M. Haroutunian: Community detection: Comparison of state of the art algorithms. Proc. of Intern. Conf. Computer science and information technologies, 252–256, Reprint in IEEE Revised selected papers, pp. 125-129 (2017)
3. N. X. Vinh, J. Apps, J. Bailey: Information Theoretic Measures for Clusterings Comparison: Variants, Properties, Normalization and Correction for Chance. The Journal of Machine Learning Research **11**, 2837–2854 (2010)
4. J. Malek, C. Hocine, C. Chantal, H. Atef: Community detection algorithm evaluation with ground-truth data. Physica A: Statistical Mechanics and its Applications **492**, 651–706 (2018)
5. Z. Yang, R. Algesheimer, C. J. Tessone: A Comparative Analysis of Community Detection Algorithms on Artificial Networks. Scientific Reports **6**(30750) (2016)
6. T. M. Cover and J. A. Thomas: Elements of Information Theory. Wiley Series in Telecommunications and Signal Processing (2006)
7. S. Wagner and D. Wagner: Comparing Clusterings- An Overview (2007)
8. Y. Yao: Information-theoretic measures for knowledge discovery and data mining. Springer, Entropy Measures, Maximum Entropy Principle and Emerging Applications, 115–136 (2003)
9. T. O. Kvalseth: Entropy and correlation: Some comments. Systems, Man and Cybernetics, IEEE Transactions **17**(3), 517-519 (1987)
10. A. Strehl, J. Ghosh: Cluster ensembles - a knowledge reuse framework for combining multiple partitions. The Journal of Machine Learning Research **3**, 583-617 (2002)
11. M. Meila: Comparing clusterings—an information based distance. Journal of Multivariate Analysis **98**, 873–895 (2007)
12. A. Kraskov, H. Stgbauer, R. G. Andrzejak and P. Grassberger: Hierarchical clustering using mutual information. EPL (Europhysics Letters)**70**, 2837–2854 (2005)
13. A. Amelio, C. Pizzuti: Is Normalized Mutual Information a Fair Measure for Comparing Community Detection Methods?. IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (2015)
14. P. Zhang: Evaluating accuracy of community detection using the relative normalized mutual information. Journal of Statistical Mechanics: Theory and Experiment **2015** (2015)
15. I. Sason, S. Verdú: f-divergence Inequalities. IEEE Transactions on Information Theory **62**(11), 5973–6006 (2016)
16. I. Csiszár, P. C. Shields: Information Theory and Statistics: A Tutorial. Foundations and Trends in Communications and Information Theory **1**(4), 417–528 (2004)
17. I. Sason: Tight Bounds for Symmetric Divergence Measures and a Refined Bound for Lossless Source Coding. IEEE Trans. on IT **61**(2), 701–707 (2015)
18. F. Topsøe: Some inequalities for information divergence and related measures of discrimination. IEEE Trans. on IT **46**(4), 1602-1609 (2000)
19. J. Lin: Divergence measures based on Shannon entropy. IEEE Trans. on IT **37**(1), 145–151 (1991)