INFORMATION-BOTTLENECK BASED ON THE JENSEN-SHANNON DIVERGENCE WITH APPLICATIONS TO PAIRWISE CLUSTERING

Jacob Goldberger *Yaniv Opochinsky*

Faculty of Engineering, Bar-Ilan University, Ramat-Gan, Israel

ABSTRACT

The information-bottleneck (IB) principle is defined in terms of mutual information. This study defines mutual information between two random variables using the Jensen-Shannon (JS) divergence instead of the standard definition which is based on the Kullback-Leibler (KL) divergence. We reformulate the information-bottleneck principle using the proposed mutual information and apply it to the problem of pairwise clustering. We show that applying IB to clustering tasks using JS divergences instead of KL yields improved results. This indicates that JS-based mutual information has an expressive power at least as the standard KL-based mutual information.

Index Terms— Jensen-Shannon (JS) divergence, pairwise clustering, information bottleneck

1. INTRODUCTION

The information bottleneck (IB) method is a technique in information theory introduced by Tishby, Pereira, and Bialek [1]. It is designed for finding the best tradeoff between accuracy and complexity (compression) when summarizing (e.g. clustering) a random variable x, given a joint probability distribution $p(x, y)$ between x and an observed relevant variable y. The IB principle is mostly applied to form clustering algorithms (see e.g. [2] [3] [4]). Recently IB has been also suggested as a theoretical foundation for deep learning [5]. The IB principle is defined in terms of the mutual information between the feature information y and compressed version of the r.v. x. The Mutual information between two random variables is defined as Kullback-Leibler (KL) divergence between their joint distribution and the product of marginal distributions.

Measuring the difference between two distributions – their divergence – is a key element in many data analysis tasks. The most popular is the Kullback-Leibler (KL) divergence which measures the expected number of extra bits required to code samples from one distribution with a code optimized for another. An alternative measure is the Jensen-Shannon (JS) divergence [6]. As a pure distance measure, JS seems superior in that it is symmetric, it is bounded from above and its square root is even a metric [7]. The KL divergence, however, is by far the most frequently used in the formulation of machine learning algorithms since it is related to maximum-likelihood estimation.

The JS divergence between distributions p and q can be seen as measuring the performance of the optimal binary classifier that needs to decide whether a given data point was sampled from p or from q. This interpretation recently resulted in connections between JS and the analysis of several deep learning algorithms. Generative adversarial networks (GANs) [8] are a class of methods for learning generative models based on game theory. Goodfellow et al. [8] showed that training a GAN is equivalent to minimizing the Jensen-Shannon divergence between the generator and the data distributions. A popular word embedding algorithm is *word2vec* [9] [10]. Melamud and Goldberger [11] showed that the global optimum of the *word2vec* objective function is the JS divergence between word and context joint distribution and the product of their marginal distributions. They also showed that *word2vec*'s algorithm finds the optimal low-dimensional approximation of this JS measure. These and other examples motivate re-considering JS as the preferred measure for a range of machine learning algorithms.

In this study we define mutual information between random variables by using JS instead of KL. We dub this mutual information the Jensen-Shannon Mutual Information (JSMI). The IB principle was originally defined based on KL-based mutual information. We reformulate the information bottleneck principle using the proposed, JS-based mutual information, JSMI.

We utilize the proposed IB based on JSMI measure for the task of pairwise clustering. In a pairwise clustering setup features representation is not available and we are only given pairwise similarity information between data points. Previous works used either Normalized-Cut [12] or IB based on mutual information [4] as a criterion that is optimized to find the best clustering. We show that applying IB to pairwise clustering tasks using JSMI instead of KL based mutual-information yields improved results.

2. MUTUAL INFORMATION BASED ON JENSEN-SHANNON

In this section, we define a dependency measure between two random variables, which is based on the Jensen-Shannon

divergence. There are several standard methods of measuring the distance between two discrete probability distributions, defined on a given finite set $\mathcal{A}$. The KL divergence of a distribution p from a distribution q is defined as follows: $\mathrm{KL}(p||q) = \sum_{i\in\mathcal{A}} p_i \log \frac{p_i}{q_i}$. The mutual information between two jointly distributed random variables X and Y is defined as the KL divergence of the joint distribution $p(x,y)$ from the product $p(x)p(y)$ of the marginal distributions of X and Y, i.e. $I(X;Y) = \mathrm{KL}(p(x,y)||p(x)p(y))$.

The Jensen-Shannon (JS) divergence [6] between distributions p and q is:

$$\begin{aligned}\mathrm{JS}_\alpha(p,q) &= \alpha\mathrm{KL}(p||r) + (1-\alpha)\mathrm{KL}(q||r) \qquad (1)\\ &= H(r) - \alpha H(p) - (1-\alpha)H(q)\end{aligned}$$

such that $0 < \alpha < 1$, $r = \alpha p + (1-\alpha)q$ and H is the entropy function (i.e. $H(p) = -\sum_i p_i \log p_i$). Unlike KL divergence, JS divergence is bounded from above and $0 \le \mathrm{JS}_\alpha(p,q) \le 1$.

We next propose a new measure for mutual information using the JS-divergence between $p(x,y)$ and $p(x)p(y)$ instead of the KL-divergence. We define the Jensen-Shannon Mutual information (JSMI) as follows:

$$\mathrm{J}_\alpha(X;Y) = \mathrm{JS}_\alpha(p(x,y), p(x)p(y)). \qquad (2)$$

It can be easily verified that X and Y are independent if and only if $\mathrm{J}_\alpha(X;Y) = 0$.

We next derive an alternative definition of the JSMI dependency measure. Assume we choose between the two distributions, $p(x,y)$ and the product of marginal distributions $p(x)p(y)$, according to a binary random variable B, such that $p(B=1) = \alpha$. We first sample a binary value for B and next, sample a r.v. W as follows:

$$p(W=(x,y)|B) = \begin{cases} p(x)p(y) & \text{if} \quad B=0 \\ p(x,y) & \text{if} \quad B=1. \end{cases} \qquad (3)$$

The divergence measure $\mathrm{J}_\alpha(X;Y)$ can alternatively be defined in terms of the mutual information between W and B. The mutual information between W and B is:

$$\begin{aligned}I(W;B) &= H(W) - \sum_{i=0,1} p(B=i)H(W|B=i)\\ &= H(\alpha p(x,y) + (1-\alpha)p(x)p(y))\\ &-\alpha H(p(x,y)) - (1-\alpha)H(p(x)p(y)).\end{aligned}$$

Eq. (1) thus implies that:

$$\mathrm{J}_\alpha(X;Y) = I(W;B). \qquad (4)$$

We note that the JSMI satisfies the fundamental data processing inequality. If $X \to Y \to Z$ is a Markov chain, then

$$\mathrm{J}_\alpha(X;Z) \le \mathrm{J}_\alpha(X;Y).$$

To validate it we can use the fact that Jensen-Shannon is an f-divergence and that the data processing inequality for KL divergence extends to all f-divergences [13].

3. A JSMI VERSION OF THE IB PRINCIPLE

In this section we define a variant of the Information Bottleneck (IB) principle [1] using the proposed JS-based mutual information. Assume we are given a joint distribution $p(x,y)$ of objects X and features Y and we want to cluster the object-set in such a way that the clusters are maximally correlated with the features. The IB principle [1] states that among all the possible clusterings of the object set into a fixed number of clusters, the desired clustering is the one that minimizes the loss of mutual information between the objects and the features. According to the (hard version of the) IB principle we seek a clustering C of object space X such that the information loss $I(X;Y) - I(C;Y)$ is minimized. Note that C is a deterministic grouping of the objects and therefore $C \to X \to Y$ is a Markov chain. We use the notation $x \in c$ to denote that x is in cluster c. The clustering cost function we thus want to minimize is:

$$\begin{aligned}S_{mi}(C) &= I(X;Y) - I(C;Y) \qquad (5)\\ &= \sum_c \sum_{x\in c} p(x)\mathrm{KL}(p(y|x)||p(y|c)).\end{aligned}$$

We next define a JSMI version of the IB principle where we use JSMI instead of the mutual information as a measure of the correlation between two random variables. The modified cost function we aim to minimize is:

$$S_{jsmi}(C) = \mathrm{J}_\alpha(X;Y) - \mathrm{J}_\alpha(C;Y). \qquad (6)$$

The data processing lemma [13] guarantees that the clustering operation indeed causes a JS information loss, i.e. $S_{jsmi}(C) \ge 0$. Hence, defining the IB principle using JS instead of KL makes sense.

There is no closed-form solution for the minimal information-loss criterion. Several standard optimization algorithms can be utilized to find the best clustering. In this study we apply a greedy sequential algorithm (see e.g. [2]).

4. PAIRWISE CLUSTERING BASED ON THE JSMI CRITERION

In this section the JS-based information bottleneck principle is applied to pairwise clsutering. We first represent the problem as graph clustering and then translate it into the problem of clustering the states of a Markov chain. Finally, we apply the JSMI version of the IB principle to define a clustering cost function. The optimal clustering is the one that minimizes the proposed cost function.

Given a set of data points $x_1, ..., x_n$ and some symmetric notion of similarity $w_{ij} \ge 0$ between all pairs of data points x_i and x_j, the goal of clustering is to divide the data points into several groups such that points in the same group are similar and points in different groups are dissimilar to each other. In many cases feature representation is not available

and we are only given pairwise similarity information between the data points. For example, in social networks, only binary neighborhood relations are given. In these cases feature based clustering algorithms (such as k-means) cannot be applied in a straightforward way. Instead, we are looking for a partition of the data based only on the similarity measure between the points.

We can represent the data as a similarity graph $G = (V, E)$. Each vertex i in this graph represents a data point x_i. Two vertices are connected if the similarity w_{ij} between the corresponding data points x_i and x_j is positive and the edge is weighted by w_{ij}. The problem of clustering can now be reformulated using the similarity graph: we want to find a partition of the graph in which existing edges between different groups have low weights and edges within a group have high weights.

Denote the similarity matrix by $W = (w_{ij})$. The degree of a vertex $i \in V$ is defined as $d_i = \sum_{j=1}^{n} w_{ij}$. The degree matrix D is defined as the diagonal matrix with the degrees $d_1, ..., d_n$ on the diagonal. The matrix $P = D^{-1}W$ is a stochastic matrix (non-negative entries, row sums are all 1) and therefore it defines a stationary Markov chain that corresponds to a random walk on the graph nodes. Let $X = \{X_t\}$ be the n-valued stationary Markov chain defined by:

$$P_{ij} = (D^{-1}W)_{ij} = p(X_2 = j | X_1 = i) = \frac{w_{ij}}{\sum_k w_{ik}} \quad (7)$$

The transition probability P_{ij} of jumping in one step from i to j is proportional to the edge weight w_{ij}. Let $\pi = (\pi_1, ..., \pi_n)$, where $\pi_i = d_i/(\sum_j d_j)$. It can be easily verified that $P^\top \pi = \pi$. Hence, if the graph is connected and non-bipartite, then π is the unique stationary distribution of the Markov chain defined by P [14]. Therefore, the joint stationary probability of X_1 and X_2 is:

$$p(X_1 = i, X_2 = j) = \frac{w_{ij}}{\sum_{kl} w_{kl}}. \quad (8)$$

Given the random walk model (7) we thus translated the graph clustering problem, into the problem of clustering the states of a Markov chain.

Let $\{A_1, ..., A_m\}$ be a partition of the states of a Markov chain $\{1, ..., n\}$ into m clusters and let C denote the subset membership function, i.e. $C(i) = j$ if $i \in A_j$. For each t we define a random variable $C_t = C(X_t)$ indicating the cluster membership of the state visited by the random walk at time t. The joint distribution of the random variables (C_1, C_2) defined on the clusters is:

$$p(C_1 = i, C_2 = j) = p(X_1 \in A_i, X_2 \in A_j). \quad (9)$$

Note that the joint clustering forms a Markov chain:

$$C_1 \to X_1 \to X_2 \to C_2.$$

The original walk over the points also determines the walk over the clusters. The goal of clustering is to choose the clustering such that the loss in mutual information due to clustering is minimized. A good Markov-state clustering should preserve maximum information on the visited points. Using the mutual information criterion, the best clustering of the given n points into m clusters is the one that minimizes the information loss of the mutual information $I(X_1; X_2) - I(C_1; C_2)$ over all the partitions of the state-space into m subsets [4]. The clustering score we thus want to minimize is:

$$\text{score}_{mi}(C) = I(X_1; X_2) - I(C_1; C_2) \quad (10)$$

where $C_1 = C(X_1)$, $C_2 = C(X_2)$. The optimal state-clustering is the one that minimizes the information-loss function $\text{score}_{mi}(C)$.

Using the JSMI version of the IB principle, the modified cost function we aim to minimize is:

$$\text{score}_{jsmi}(C) = \text{J}_\alpha(X_1; X_2) - \text{J}_\alpha(C_1; C_2). \quad (11)$$

The data processing inequality guaranties that the score is non-negative. As described above, the minimization can be done by a greedy strategy. The proposed algorithm is summarized in Table 1. In the next section we empirically compare the mutual-information (MI) and JSMI clustering criteria and show that using JSMI for information-bottleneck clustering indeed yields improved results.

Table 1. The JSMI Pairwise Clustering Algorithm.

Input: A similarity graph defined by the $n \times n$ weight matrix W.
Output: A partition of the graph vertices into m clusters.

Algorithm:

1. Convert the graph into a Markov chain:
$$\widetilde{w}_{ij} \triangleq p(X_1 = i, X_2 = j) = \frac{w_{ij}}{\sum_{kl} w_{kl}}$$
2. Choose a random partition C of the Markov states.
3. Loop until there is no change:
   - for $i = 1, ..., n$
     Move state i into the cluster that minimizes the information loss:
     $$\text{score}_{jsmi}(C) = \text{J}_\alpha(X_1; X_2) - \text{J}_\alpha(C_1; C_2).$$

5. EXPERIMENTAL RESULTS

In this section we demonstrate our proposed clustering objective function on the following commonly used real-world datasets. **USPS-245:** 1650 instances of handwritten digits

Table 2. Clustering results on standard datasets.

	NCut		MI		JSMI	
	NMI	RI	NMI	RI	NMI	RI
USPS-245	.56	.77	.76	.87	**.81**	**.91**
Iris	65	.78	.71	.83	**.78**	**.88**
Wine	**.86**	**.94**	.79	.89	.85	.93
Face	**.52**	.63	.43	.63	.49	63

2,4 and 5 from the USPS dataset [15]. **Iris** contains flower petal and sepal measurements from three species of irises, 150 instances [16]. **Wine** are the results of a chemical analysis of wines. The analysis determined the quantities of 13 constituents found in each of three types of wines. 178 instances. **Olivetti Faces (OlFace5)** 10 images of 5 different people, 64×64 size [17].

To construct the pairwise similarity matrix we used the k-nearest neighbor graph, based on the Euclidean distance, with $k = 10$. We set $w_{ij} = 1$ if node i is a k-nearest neighbor of node j or j is a k-nearest neighbor of i. Otherwise, we set $w_{ij} = 0$. We used a simple sequential clustering algorithm to find the best clustering of the data. It starts with a random clustering of the object-set. We then go over the data points in a circular manner and check for each point whether its removal from one cluster to another can reduce the information loss. This loop is iterated until no single-point transition offers an improvement. Since there is no guarantee that the algorithm will find the global optimum, we ran the algorithm on several initial random partitions and chose the best local optimum. We use the same random initialization for all methods.

We implemented three optimization criteria MI (10) [4], JSMI (11) and also the non information-theoretical measure Normalized Cut (NCut) [18] [12]. Ncut is is defined as follows:

$$\text{score}_{ncut}(C) = \sum_{i=1}^{m} p(X_2 \notin A_i | X_1 \in A_i). \qquad (12)$$

In all the implementations of the JSMI we set the parameter α to be 1/2. We didn't observe a significant change when trying other values of α.

To evaluate the performance of the clustering methods we measured the clustering results against the true labels using two standard external validation indices: normalized mutual information (NMI) [19], and the Rand index (RI) [20]. We refer the reader to [21] for details regarding these measures. The results are shown in Table 2. As can be seen, using JSMI criterion for IB clustering (10) we get better results than using the standard MI criterion (11) to find the optimal clustering. Also in most cases the NCut was much worse or similar to the JSMI based clustering algorithm.

We next evaluated the proposed clustering approach using a controlled experiment on simulated datasets. In this experiment we directly compare the MI criterion (10) for pairwise clustering to the proposed JSMI criterion (11). We simulated

Table 3. NMI clustering results on three concentric circles as a function of the noise SD.

σ	NCut	MI	JSMI
0.1	0.902	0.982	**0.993**
0.2	0.715	0.754	**0.765**
0.3	0.678	0.746	**0.750**

a standard clustering problem in which the data are formed of concentric circles, where each circle represents a different cluster. In this problem the clusters are non-convex and it is considered to be impossible for the k-means algorithm to solve. We created a dataset composed of three concentric circles (with radii 1, 2 and 3) with 50 points equally spaced on each circle. An isotropic Gaussian noise with variance σ^2 was added to each point. The pairwise similarity matrix was constructed using the same procedure as for the dataset described below. We applied two variants of the sequential greedy clustering algorithm. In the first variant we aimed to minimize the MI score and in the second we aimed to minimize the JSMI score. We used the same random clustering initialization for the two algorithms. Therefore, the only difference between the two methods was the pairwise clustering criterion used for optimization. Table 3 shows the results (as a function of σ the noise standard deviation) averaged over 100 experiments. The clustering quality was measured by the NMI index. Similar results were obtained by measuring performance by the Rand Index. We also show the clustering result based on the NCut criterion. The results show that by using the JSMI criterion rather than the MI led to better result. This indicates that JSMI has better expressive power as the standard mutual information.

6. CONCLUSIONS

To conclude, in this study we defined a mutual information notion based on the Jensen-Shannon divergence dubbed JSMI and we developed a corresponding Information Bottleneck principle. We then used it as a pairwise clustering optimization criterion and obtained better performance compared to the standard clustering algorithm based on the MI criterion. Recent deep-learning analysis showed the potential usefulness of the JSMI measure. We hope that this study will encourage future machine learning applications of it for clustering and other tasks.

7. REFERENCES

[1] N. Tishby, F. C. Pereira, and W. Bialek, "The information bottleneck method," in *Allerton Conf. on Communication, Control, and Computing*, 1999.

[2] N. Slonim, N. Friedman, and N. Tishby, "Unsupervised document classification using sequential informa-

tion maximization," in *Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, 2002.

[3] L. Faivishevsky and J. Goldberger, "A nonparametric information theoretic clustering algorithm," in *Int. Conf. on Machine Learning*, 2010.

[4] A. Alush, A. Friedman, and J. Goldberger, "Pairwise clustering based on the mutual-information criterion," *Neurocomputing*, vol. 182, pp. 284–293, 2016.

[5] R. Shwartz-Ziv and N. Tishby, "Opening the black box of deep neural networks via information," *CoRR*, vol. abs/1703.00810, 2017.

[6] J. Lin, "Divergence measures based on the shannon entropy," *IEEE Transactions on Information Theory*, vol. 37, no. 1, pp. 145–151, 1991.

[7] D. M. Endres and J. E. Schindelin, "A new metric for probability distributions," *IEEE Trans. Inf. Theory*, vol. 49, pp. 1858–60, 2003.

[8] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems*, 2014.

[9] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in Neural Information Processing Systems*, 2013.

[10] O. Levy and Y. Goldberg, "Neural word embedding as implicit matrix factorization," in *Advances in Neural Information Processing Systems*, 2014.

[11] O. Melamud and J. Goldberger, "Information-theory interpretation of the skip-gram negative-sampling objective function," in *Proceedings of ACL*, 2017.

[12] M. Meila and J. Shi, "A random walks view of spectral segmentation," *AISTATS*, 2001.

[13] M. Pardo and I. Vajda, "About distances of discrete distributions satisfying the data-processing theorem of information theory," *IEEE Transactions on Information Theory*, vol. 43, no. 4, pp. 1288–1293, 1997.

[14] U. von Luxburg, "A tutorial on spectral clustering," *Statistics and Computing*, pp. 395–416, 2007.

[15] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, Springer, 2001.

[16] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Annual Eugenics*, vol. 7 (2), pp. 179–188, 1936.

[17] F. Samaria and A. Harter, "Parameterisation of a stochastic model for human face identification," in *IEEE Workshop on Applications of Computer Vision*, 1994.

[18] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Trans. pattern Anal. Machine Intell*, vol. 22, (8), pp. 888–905, 2000.

[19] D. Cai, X. He, and J. Han., "Locally consistent concept factorization for document clustering," *IEEE Transactions on Knowledge and Data Engineering*, vol. 23, no. 6, pp. 902–913, 2011.

[20] K. Y. Yeung and W. L W. L. Ruzzo, "Details of the adjusted rand index and clustering algorithms, supplement to the paper an empirical study on principal component analysis for clustering gene expression data," *Bioinformatics*, vol. 19, no. 9, pp. 763–774, 2001.

[21] C. Manning, P. Raghavan, and H. Schutze, "Introduction to information retrieval," *Cambridge University Press*, 2008.