INTERNAL PURITY: A DIFFERENTIAL ENTROPY BASED INTERNAL VALIDATION INDEX FOR CLUSTERING VALIDATION

Anonymous authors
Paper under double-blind review

ABSTRACT

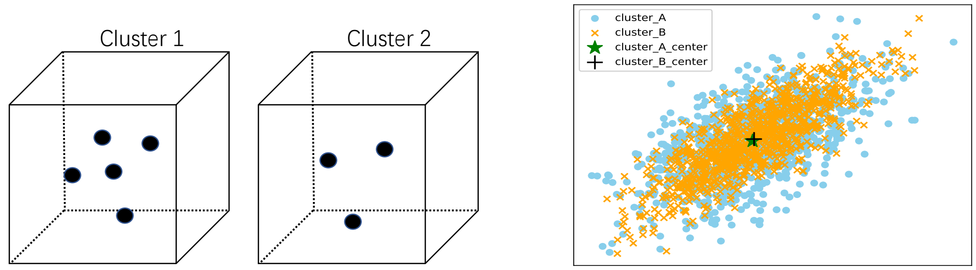
In an effective process of cluster analysis, it is indispensable to validate the goodness of different partitions after clustering. Existing internal validation indices are implemented based on distance and variance, which cannnot catpure the real "density" of the cluster. Moreover the time complexity for distance based indices is usually too high to be applied for large datasets. Therefore, we propose a novel internal validation index based on the differential entropy, named *internal purity* (IP). The proposed IP index can effectively measure the purity of a cluster without using the external cluster information, and successfully overcome the drawbacks of existing internal indices. Based on six powerful deep pre-trained representation models, we use four basic clustering algorithms to compare our index with thirteen other well-known internal indices on five text and five image datasets. The results show that, for 90 test cases in total, our IP index can return the optimal clustering results in 51 cases while the second best index can merely report the optimal partition in 30 cases, which demonstrates the significant superiority of our IP index when validating the goodness of the clustering results. Moreover, theoretical analysis for the effectiveness and efficiency of the proposed index are also provided.

1 INTRODUCTION

The goal of clustering is to divide a set of samples into different clusters such that similar samples are grouped in the same clusters. As one of the most fundamental tasks in machine learning, clustering has been extensively studied in many fields, such as text mining (Guan et al., 2012a), image analysis (Zhou et al., 2011) and pattern recognition (Guan et al., 2012b). With the advance of deep learning, it has been proved that running any classical clustering algorithm (e.g., K-means) over the learned representation can yield better results (Xie et al., 2016; Huang et al., 2020; Dang et al., 2021). The main reason behind this is that the deep neural networks can effectively extract highly non-linear features that are helpful for clustering.

However, besides the data representation, the outcome of clustering is still affected by other factors (Xu & Wunsch, 2005; Yang et al., 2017). For example, different clustering algorithms usually lead to different clustering results in a specific dataset. Even for the same algorithm, the selection of different parameters may affect the final clustering results (Halkidi et al., 2000). Thus, within an effective process of cluster analysis, it is inevitable to validate the goodness of different partitions after clustering and select the best one for application. Here the best one refers to not only the proper parameters but also the best partition that fits the underlying data structure. In fact, many clustering validation measures have been proposed over past years and they can be categorized to *external* validation and *internal* validation (Wu et al., 2009; Liu et al., 2013). Specifically, external validation indices assume the "true" cluster information is known in advance, and they use the supervised information to quantify how good is the obtained partition with respect to prior ground truth clusters. However, such prior knowledge is rarely available in many real scenarios. Then, internal validation indices become the only option for evaluating the clustering result.

Internal clustering validation usually measures the clustering result based on following two criteria (Liu et al., 2013; Fraley & Raftery, 1998): (1) *compactness*, which measures how closely related are



(a) Based on distance (b) Based on variance

Figure 1: Drawback illustration for computing compactness based on distance and variance

the samples within the same cluster, and (2) *separation*, which measures how clusters are separated from each other. In general, distance and variance are two main strategies to implement compactness and separation. However, validation indices based on these implementations suffers following drawbacks that limit their performance.

First, for distance based index, given two clusters, the same distance computation result cannot guarantee the same compactness. We use the example shown in Fig.1(a) for explanation. Particularly, two clusters are grouped into two cubes respectively and each cube represents the volume of the corresponding vector space. Suppose the volume of two cluster vector spaces are the same and the average pairwise distance of each cluster is also the same, measures like Silhouette index (Rousseeuw, 1987) will consider they have the same compactness. However, from the density perspective, the left cluster should be tighter than the right one. Though indices like S_Dbw (Halkidi & Vazirgiannis, 2001), DBCV (Moulavi et al., 2014) and DCVI (Xie et al., 2020) also propose the density related concept, the density is still calculated based on the distance instead of the volume of vector space. Then, many existing indices require computing pairwise distances for compactness or separation, which would be prohibitively time-consuming for a large high-dimensional data set (Cheng et al., 2018).

Second, variance based compactness computation usually view lower variance as the indicator of higher tightness. However, this is misleading in some cases. As shown in Fig. 1(b), we can easily observe that cluster B is more compact than cluster A. But the variances for both clusters are the same, which makes measures like standard deviation index (SD) (Halkidi et al., 2000) fail to provide a reliable validation. In fact, covariance here is more suitable to measure the compactness, i.e., the covariances among pairs of variables in cluster A is smaller than that of cluster B.

Therefore, we dedicate this paper to a novel internal validation index, named *internal purity* (IP). Here the purity refers to how "pure" the semantic of a set of samples are. For example, a cluster of similar texts describe a specific event. Hence, from the perspective of compactness, we want a cluster to be as pure as possible, while for the separation perspective, a clustering partition of lower purity is favored. To evaluate the purity, we apply the idea of information entropy (Shannon, 1948) and more specifically, the differential entropy (Cover & Thomas, 1991) is used for evaluation due to that the embedding variables are usually continuous in deep clustering. Furthermore, using differential entropy can help us to overcome the aforementioned drawbacks of existing measures. First of all, unlike the average pairwise distance, the nature of the entropy lends itself to measuring the "information density" (Zu et al., 2020) of a vector space (Cheng et al., 1999). Then, theoretical studies have shown that computation of differential entropy actually considers the variance of each variable and the covariance between variables (Johnson et al., 2014), which makes it more effectively measure the compactness of a cluster. Last but not least, since the differential entropy computation requires merely one iteration over the whole cluster, it needs less computation time compared with computing average pairwise distance.

In fact, from the perspective of information theory, there is another information criterion (Bishop & Nasrabadi, 2006) that can be used for evaluating clustering partitions (Akogul & Erisoglu, 2016), i.e., estimating the quantity of information loss based on model performance and complexity, where model performance can be evaluated by using a probalilistic framework, such as log-likelihood and model complexity by the number of parameters in the model (Akogul & Erisoglu, 2016). Akaike information criterion (AIC) (Akaike, 1974) and the Bayesian information criterion (BIC) (Schwarz, 1978) are two indices of this criterion. As will be shown later in the paper, our IP index can actually be regarded as a form of such criterion through mathematical derivation. But different from AIC and BIC, our IP index also takes into account the traditional internal clustering criteria, i.e. *compactness* and *separation*, whereas AIC and BIC can not. Moreover, some existing works have shown that AIC and BIC indices have poor performance and are prone to overestimate the number of clusters in dataset (Windham & Cutler, 1992; Hu & Xu, 2004). Hence, our IP index is more competitive for evaluating the clustering results based on its capability to capture different criteria.

Note that, although the term ”purity” has been used in external validation (Wu et al., 2009) and the entropy has also been used as the measurement, the definition as well as the computation of our internal purity is quite different since no supervised information is known in advance in our validation setting. To summarize, our contributions are threefold:

1. To tackle the effectiveness and efficiency problems of existing measures, we propose to use differential entropy to measure the purity of a cluster as well as a partition.
2. Following the traditional perspective of compactness and separation, we design a new internal validation index based on the nature of the proposed purity measure. Moreover, theoretical analysis for the effectiveness and efficiency of the proposed internal purity is also provided.
3. Based on six powerful deep pre-trained models, we use four basic clustering algorithms to compare our index with thirteen other well-known internal indices on five text and five image datasets. The results show that for 90 test cases, our IP index can return the optimal clustering results in 51 cases while the second best index can merely report the optimal partition in 30 cases.

The remainder of this paper is organized as follows. Section 2 briefly reviews the existing internal indices. Section 3 introduces preliminaries. Section 4 presents our internal purity index. Section 5 reports the results of experimental evaluation. Finally, Section 6 concludes the paper.

2 RELATED WORK

Table 1: The description of well known internal clustering validation indices.

Measure	Notation	Description
Root-mean-square standard deviation	RMSSTD	RMSSTD of a cluster is the square root of the variance of all the attributes.
Calinski-Harabasz index	CH	CH is the ratio of the sum of between-clusters scatter and of within-cluster scatter for all clusters.
I index	I	I measures compactness by calculating the distance from the samples in a cluster to their cluster center and the separation based on the maximum distance between cluster centers.
Dunn index	Dunn	Dunn uses the farthest distance between samples in cluster as the compactness and uses the distance between the nearest samples in different clusters as the separation.
Silhouette index	S	S measures compactness based on the pairwise distance in a cluster and separation based on the average distance between a sample and all other samples in the next nearest cluster.
Davies-Bouldin index	DB	DB measures compactness by calculating the average distance between samples in a cluster to their cluster center and the separation based on the distance between cluster centers.
Xie-Beni index	XB	XB defines compactness as average center-based distance and separation as the minimal squared distances between cluster centers .
Standard deviation index	SD	SD measures compactness based on variances of samples in a cluster, and separation based on distances between cluster centers.
S_Dbw validation index	S_Dbw	S_Dbw measures compactness based on variances of samples in a cluster, and separation based on average density among clusters.
Clustering validation index based on nearest neighbors	CVNN	CVNN measures compactness based on pairwise distance in a cluster and separation based on the idea of k-nearest neighbor (kNN).
Density-Based Clustering Validation	DBCV	DBCV measures compactness based on maximum edge weight of the minimum spanning tree and separation based on minimum reachability distance between nodes in different clusters.
Density core based clustering validation index	DCVI	DCVI measures compactness based on maximum weight value of the minimum spanning tree and separation based on the minimum value of minimum distances between samples in different clusters.

The internal clustering validation indices measure the quality of clustering by the internal information of the dataset and no other external information is used. We group the internal indices to two categories based on how the criteria of compactness and separation is calculated, i.e., distance-based and variance-based. We briefly summarize several well known indices in Table 1.

Distance-based. Indices in this category either use pairwise distance or cluster center based distance to measure the compactness and the separation. DB (Davies & Bouldin, 1979), I (Maulik &

Bandyopadhyay, 2002) and XB (Xie & Beni, 1991) measure compactness based on average center-based distance. Dunn (Dunn, 1973) measures compactness based on maximum pairwise distance. DBCV and DCVI select the most representative pairwise distance for measuring the compactness. However, using center based or a single pairwise distance in the cluster to represent the compactness for the entire cluster cannot provide stable performance (Liu et al., 2013). Our IP index uses all the samples in the cluster which is more stable. Although S and CVNN can use all samples in the cluster by computing average pairwise distance, they have high time complexity. Moreover, as mentioned in Section 1, these average pairwise distance-based indices cannot correctly reflect the density of a vector space.

Variance-based. Indices in this category assume lower variance indicates better compactness. RMSSTD (Halkidi et al., 2001), CH (Caliński & Harabasz, 1974), SD and S_Dbw measure compactness based on variances of samples in a cluster. As a representative, CH further measures the separation by computing the between-cluster variance based on cluster centroids. However, these variance-based indices are not good measures since two clusters with same variance may have distinct density.

3 PRELIMINARIES

In this section, we first present the concept of entropy and differential entropy. Then we introduce the computation of differential entropy for a multivariate normal distribution. Note that, though the normal distribution assumption made here seems restrictive, it still works for datasets that deviate from the assumption as long as we use their deep representations. This is because the distribution of deep representation has been proved to be close to the normal distribution (Lee et al., 2018; Daneshmand et al., 2021).

3.1 ENTROPY AND DIFFERENTIAL ENTROPY

In information theory, the entropy of a random variable is the average level of "surprise" or "uncertainty" inherent in the variable's possible outcomes (Shannon, 1948). In other words, it can be used to measure the uncertainty of data. Entropy includes two classes: entropy for discrete random variables and entropy for continuous random variables. i.e., differential entropy.

Given a discrete random variable V, with possible outcomes $v_1, ..., v_n$, we assume the probability of V being v_i is P_i. The entropy of V is formally defined as (Shannon, 1948):

$$Entropy(V) = -\sum_{i=1}^{n} P_i \log P_i \tag{1}$$

Given a continuous random variable V, with a probability density function $p(v)$. The differential entropy is defined as (Cover & Thomas, 1991):

$$DiffEntropy(V) = -\int_{-\infty}^{+\infty} p(v) \log p(v) dv \tag{2}$$

3.2 DIFFERENTIAL ENTROPY OF THE MULTIVARIATE NORMAL DISTRIBUTION

The multivariate normal distribution of a d-dimensional random vector $H = \left(H^1, \ldots, H^d\right)^{\mathrm{T}}$ can be written in the following form (Goodfellow et al., 2016):

$$H \sim \mathcal{N}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{3}$$

with d-dimensional mean vector

$$\boldsymbol{\mu} = \mathrm{E}[H] = \left(\mathrm{E}\left[H^1\right], \mathrm{E}\left[H^2\right], \ldots, \mathrm{E}\left[H^d\right]\right)^{\mathrm{T}} \tag{4}$$

and $d \times d$ covariance matrix

$$\boldsymbol{\Sigma}_{i,j} = \mathrm{E}\left[\left(H^i - \mu^i\right)\left(H^j - \mu^j\right)\right] = \mathrm{Cov}\left[H^i, H^j\right] \tag{5}$$

such that $1 \leq i, j \leq d$, and $\boldsymbol{\Sigma}$ is a positive definite matrix. The differential entropy of multivariate normal distribution is given by (Ahmed & Gokhale, 1989):

$$DiffEntropy(H) = \frac{rank(\boldsymbol{\Sigma})}{2} + \frac{rank(\boldsymbol{\Sigma})}{2} \ln(2\pi) + \frac{1}{2} \ln |\boldsymbol{\Sigma}| \tag{6}$$

where, $rank(\mathbf{\Sigma})$ is the rank of $\mathbf{\Sigma}$ and $|\mathbf{\Sigma}|$ is the determinant of covariance matrix. If the covariance matrix $\mathbf{\Sigma}$ is not full rank, the multivariate normal distribution is degenerate (Rao et al., 1973). Then, the $rank(\mathbf{\Sigma}) < d$ and the determinant of covariance matrix is degenerated as the pseudo-determinant.

4 INTERNAL PURITY

In this section, we first present the implementation details for our proposed internal purity (IP) index, and then provide theoretical analysis for its effectiveness and efficiency.

4.1 IP IMPLEMENTATION

Following the traditional criteria of internal clustering validation, IP index consists of following two main components, i.e., compactness purity (CP) and seperation purity (SP). The former measures the average differential entropy of clusters to judge the compactness of the clustering result, while the latter evaluates the separation between clusters based on the differential entropy of the space formed by the center of each cluster.

Compactness. Let H be the feature space, where $H = \{h_1, ..., h_N\}^{\mathrm{T}}$. Supposing that N samples are clustered into k clusters, i.e. $H_1, ..., H_k$, the compactness CP for k clusters is defined as average differential entropy of k clusters:

$$CP = \frac{1}{k}\sum_{i=1}^{k} DiffEntropy(H_i) \tag{7}$$

Given a specific dataset, we can get a feature representation of each sample x_i through the pre-trained deep model $f(\cdot)$, i.e. $h_i = f(x_i)$. Then, for k-th cluster we can get $\{h_1^k, ..., h_m^k\}^{\mathrm{T}} \in H_k$, where m is the number of data samples in the k-th cluster. So H_k is a $m \times d$ matrix composed of feature representations corresponding to samples in the H_k. H_k can be viewed as a set of points in d-dimensional space. Moreover, we usually assume feature matrix H_k follows multivariate normal distribution when we have no prior knowledge of the distribution of these points (Goodfellow et al., 2016) and existing works have demonstrated that the distribution of the hidden representation is close to the normal distribution (Lee et al., 2018; Daneshmand et al., 2021). Hence we can obtain the differential entropy of a cluster H_k by Eq. (6).

Separation. Let μ_k be the centroid of the k-th cluster H_k, i.e., $\mu_k = \frac{1}{m}\sum_{i=1}^{m} h_i$, $h_i \in H_k$, where m is the number of data samples in the k-th cluster.. The separation SP for k clusters is defined as differential entropy of the feature matrix H_μ formed by k cluster centers

$$SP = DiffEntropy(H_\mu) \tag{8}$$

where $H_\mu = \{\mu_1, ..., \mu_k\}^{\mathrm{T}}$. Here we also assume that the centers of clusters follow a multivariate normal distribution. Hence, we can obtain the differential entropy of this distribution by Eq. (6).

IP Index. Based on CP and SP, the IP index for a clustering result of k clusters is defined as

$$IP = CP - SP \tag{9}$$

As shown above, our IP index takes the form of subtracting the intercluster separation from the intracluster compactness. A lower value of IP indicates a better clustering result.

4.2 THEORETICAL ANALYSIS

Effectiveness. According to Eq. (6), we can see that the differential entropy of multivariate normal distribution is proportional to the determinant of the covariance matrix. The determinant of the covariance matrix is usually called generalized variance (Johnson et al., 2014). For a fixed set of data, generalized variance is proportional to the square of the volume generated by the d deviation vectors (Johnson et al., 2014). i.e.,

$$\text{Generalized variance } = |\mathbf{\Sigma}| = (N-1)^{-d}(\text{ volume })^2 \tag{10}$$

Based on Eq.(10), we know the reason why our internal purity can avoid the drawbacks of distance or variance based indices. Specifically, the form $\frac{volume}{N}$ embedded in the Eq. (10) indicates the real density for a given vector space. Then, according to Eq. (9), we can further obtain following formula (detailed derivation is listed in Appendix A):

$$IP = \underbrace{\frac{1}{2}\left(\frac{1}{k}\sum_{i=1}^{k}\ln|\mathbf{\Sigma}_i| - \ln|\mathbf{\Sigma}_\mu|\right)}_{part\_1} + \underbrace{\frac{1+\ln(2\pi)}{2}\left[\frac{1}{k}\sum_{i=1}^{k}rank(\mathbf{\Sigma}_i) - rank(\mathbf{\Sigma}_\mu)\right]}_{part\_2} \tag{11}$$

This formula is similar to the information criterion (Bishop & Nasrabadi, 2006) mentioned in the introduction. According to this criterion, the $part\_1$ of Eq. (11) can be considered as the evaluation of the model performance and the $part\_2$ as the measure for the model complexity. This formula also indicates that our proposed IP index can also capture what traditional information criterion based indices want to measure.

Efficiency. The time complexity of IP computation is decided by the complexity of both compactness and separation. For compactness, the time complexity of calcualting the covariance matrix is $O(d^2\,|C_k|)$ and the determinant of covariance matrix complexity is $O(d^{2.376})$ (Aho & Hopcroft, 1974). Hence, the complexity of calcualting the compactness of a cluster is $O(d^2\,|C_k| + d^{2.376})$, then the complexity for k clusters is $O(Nd^2 + kd^{2.376})$. For seperation, we have to calculate k cluster centroids, and the covariance matrix formed by k cluster centroids and the corresponding determinant. So the complexity of seperation is $O(Nd + kd^2 + d^{2.376})$. Usually, $d \ll N$ and $k \ll N$, then, the complexity of IP index is $O(Nd^2)$, which makes it affordable for large-scale and high-dimensional datasets.

5 Experiments

In this section, we compare the performance of IP with other 13 well-known internal indices, namely, SD, Dunn, I, XB, S, CH, DB, AIC, BIC, S_Dbw, CVNN, DBCV and DCVI. Note that, we didn't consider RMSSTD since it need subjective determination for the shift point of its curve (Halkidi et al., 2001; Vendramin et al., 2010). Similar to existing works (Halkidi et al., 2000; Liu et al., 2013; Moulavi et al., 2014), we use the task of determining optimal cluster number for evaluation purpose. The general evaluation procedure is as follows: (1) we first use the existing pre-trained deep model to transfer the dataset to a representation matrix; (2) a set of clustering algorithms are then applied to the representation matrix and different clustering partitions can be obtained with different parameters; (3) finally, we compute the internal index for each partition and get the best partition as well as its corresponding optimal cluster number.

5.1 Settings

Datasets. We use five text datasets and five image datasets. The statistics of these datasets are shown in Table 2 and 3 respectively. For text datasets usages, we cluster the train sets of SearchSnippets (Phan et al., 2008), Biomedical (Xu et al., 2017), StackOverflow (Xu et al., 2017) and WebofScience-11967 (Kowsari et al., 2017), and 10,000 randomly selected texts from each class on Yahoo!Answers (Zhang et al., 2015). For image datasets usages, we cluster the test sets of CIFAR-10 (Krizhevsky et al., 2009), MINST (LeCun et al., 1998) and FashionMNIST (Xiao et al., 2017), and 10,000 randomly selected imagess in the test set of CINIC-10 (Darlow et al., 2018). For ImageNet-10 (Chang et al., 2017), the train set is directly used. Moreover, we also use five real-world datasets from UCI Machine Learning Repository (Frank, 2010) for evaluating how our IP index performs over dataset without deep representation, and details can be seen in Appendix E.

Evaluation Metrics. Since the information of true clusters are known for above datasets, i.e., classes are given, we can use external indices to evaluate the best clustering result each internal index selected and then further judge which internal index is better. Three external indices are used in our experiments, i.e., Accuracy (ACC) (Wu & Schölkopf, 2006), Adjusted Rand Index (ARI) (Hubert & Arabie, 1985) and Normalized Mutual Information (NMI) (Chen et al., 2010). Larger ACC, ARI and NMI indicate better clustering result. Details of the implementation of the three external indices can be seen in Appendix B.

Pre-trained Representation Models. To extract feature representations, we use following six pre-trained models. For text datasets, Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019), Sentence-BERT (SBERT) (Reimers & Gurevych, 2019) and Simple contrastive sentence embedding framework (SimCSE) (Gao et al., 2021) are used. We use the average of all output embeddings as the feature representation of each text sentence on the three models. For image datasets, we use Vision Transformer (ViT) (Dosovitskiy et al., 2021), Swin Transformer (Swin) (Liu et al., 2021) and Bidirectional Encoder Representations from Image Transformers (BEiT) (Bao et al., 2021). The generation of feature representations in different image models follows the suggestion by their original work. Note that since the unsupervised nature of clustering, we don't further fine-tune above pre-trained models with the experimental datasets.

Experimental Setup. Above models are implemented based on Huggingface's transformers (Wolf et al., 2020) and Sentence Transformers (Reimers & Gurevych, 2019). Detailed model configurations are listed in Appendix C. Our internal index is implemented based on scikit-learn (Pedregosa et al., 2011) and SciPy (Virtanen et al., 2020). Then based on above setup, we use four different clustering algorithms for generating partitions, including K-Means, GMM, agglomerative hierarchical clustering (AHC) (Ward Jr, 1963) and density-based spatial clustering of applications with noise (DBSCAN) (Ester et al., 1996). In the experiments, the numbers of clusters or components are in the search range from 2 to $\lfloor\sqrt{N}\rfloor$ (Pal & Bezdek, 1995) (Bezdek & Pal, 1998) for K-Means, GMM and AHC. Two parameters are needed for DBSCAN, $min\_samples$ and eps. For $min\_samples$, we choose $min\_samples \in [5, 100]$ and step size is 5. For eps, we obtain the minimum and maximum values of pairwise distances for each dataset and employ 50 different values of eps equally distributed within this range. Morever, since the AIC and BIC are implemented based on the current clustering model, within-cluster-sum-of squares is applied in K-Means , AHC and DBSCAN (Manning, 2008), maximum likelihood function applied in GMM. The experimental results are averaged over five random runs for each validation index.

Table 2: Statistics of text datasets.

Dataset	Split	Samples	Classes
SearchSnippets	Train	12340	8
Biomedical	Train	20000	20
StackOverflow	Train	20000	20
WebofScience-11967	Train	11967	7
Yahoo!Answers	Train	1,400,000	10

Table 3: Statistics of image datasets.

Dataset	Split	Samples	Classes
CIFAR-10	Train+Test	50,000+10,000	10
MNIST	Train+Test	60,000+10,000	10
FashionMNIST	Train+Test	60,000+10,000	10
ImageNet-10	Train	13000	10
CINIC-10	Train+Test	90,000+90,000	10

5.2 Effectiveness Study based on External Evaluation

The evaluation results based on external indices of ACC, ARI and NMI are shown in Table 5, 6, 7, 8, 9, and 10, where the best results are highlighted in bold and the optimal k each index select, i.e. opt_k, and the true cluster number in each dataset, i.e. $dataset\text{-}k$, are provided. Moreover, we also count the number of best results that each internal index can achieve and report them in Table 4. Obviously, compared with other 13 indices, our IP index has significant advance in achieving the best results. Specifically, for all 90 cases, our IP index can produce 51 best results, while the second top index AIC only has 30 best results. Finally, IP index outperforms other indices greatly with respect to different domains, i.e., text and image. Due to space limitations, every clustering algorithm evaluation results can be seen in the Appendix D respectively.

Table 4: Number of best clustering results based on counting over Table 5, 6, 7, 8, 9, and 10.

	SD	Dunn	I	XB	S	CH	DB	S_Dbw	CVNN	DCVI	DBCV	AIC	BIC	**IP**
Text datasets	0	0	0	0	3	0	0	0	0	0	0	17	0	**26**
Image datasets	0	0	0	0	0	7	0	0	0	0	0	13	3	**25**
All datasets	0	0	0	0	3	7	0	0	0	0	0	30	3	**51**

Text datasets. In Table 5, we can also observe that our IP index outperforms other indices except StackOverflow and Yahoo!Answers. In Table 6, the index S has the best results in three cases. However, it is distance based method which has higher time complexity than our IP index. In Table 7, although the index AIC has the results in nine cases, the optimal k value selected by it is far from the true number of classes on Biomedical and Yahoo!Answers.

Image datasets. In Table 8, CH obtains three best cases in terms of ACC, ARI and NMI on ImageNet-10, respectively. However as shown in Table 9 where Swin is the representation model,

Table 5: BERT based clustering results on five text datasets. (15 cases: 5 datasets × 3 evaluation metrics)

	SearchSnippets - 8				**Biomedical - 20**				**StackOverflow - 20**				**WebofScience - 7**				**Yahoo!Answers - 10**			
	ACC	ARI	NMI	opt_k	ACC	ARI	NMI	opt_k	ACC	ARI	NMI	opt_k	ACC	ARI	NMI	opt_k	ACC	ARI	NMI	opt_k
SD	21.56	0	0.01	2	5	0	0.01	2	5.01	0	0.01	2	17.6	0	0.02	2	10.01	0	0.02	2
Dunn	21.56	0	0.01	2	5	0	0.01	2	5.01	0	0.01	2	17.61	0	0.02	2	10.04	0	0.05	2
I	21.56	0	0.01	2	5	0	0.01	2	5.01	0	0.01	2	17.6	0	0.02	2	10.01	0	0.02	2
XB	21.56	0	0.01	2	5	0	0.01	2	5.01	0	0.01	2	17.6	0	0.02	2	10.01	0	0.02	2
S	21.56	0	0.01	2	5	0	0.01	2	5.01	0	0.01	2	17.6	0	0.02	2	10.01	0	0.02	2
CH	33.63	12.22	16.68	2	9.66	2.52	8.21	2	6.29	0.17	0.53	2	29.39	17.99	28.63	2	12.9	0.7	1.85	2
DB	21.56	0	0.01	2	5	0	0.01	2	5.01	0	0.01	2	17.6	0	0.02	2	10.01	0	0.02	2
S_Dbw	21.64	0.01	0.18	3	5	0	0.01	2	5.01	0	0.01	2	17.6	0	0.02	2	10.05	0	0.07	3
CVNN	21.67	0.01	1.01	14	5.34	0	0.66	2	5.62	0.02	0.43	2	17.85	0.02	0.46	2	10.04	0	0.05	2
DCVI	21.56	0	0.01	2	5	0	0.01	2	5.01	0	0.01	2	17.6	0	0.02	2	10.01	0	0.02	2
DBCV	23.41	0.5	1.13	3	8.28	0.74	3.87	6	5.95	0.06	0.58	3	19.43	0.14	0.8	3	10.01	0	0.02	2
AIC	27.84	18.61	37.90	25	24.50	13.88	**28.20**	40	**22.34**	**9.99**	**22.61**	44	40.88	24.69	36	3	**29.52**	**11.86**	**19.31**	19
BIC	34.65	14.06	19.73	2	9.82	2.94	9.93	2	6.32	0.15	0.46	2	29.42	18.43	29.34	2	12.23	0.24	1.01	2
IP	**54.18**	**33.58**	**42.31**	11	**31.96**	**15.86**	27.29	16	5.38	0.01	0.39	2	**47.93**	**31.90**	**40.83**	9	19.97	2.59	6.90	4

Table 6: SBERT based clustering results on five text datasets. (15 cases: 5 datasets × 3 evaluation metrics)

	SearchSnippets - 8				**Biomedical - 20**				**StackOverflow - 20**				**WebofScience - 7**				**Yahoo!Answers - 10**			
	ACC	ARI	NMI	opt_k	ACC	ARI	NMI	opt_k	ACC	ARI	NMI	opt_k	ACC	ARI	NMI	opt_k	ACC	ARI	NMI	opt_k
SD	21.89	0.06	1.26	13	6.24	0.03	3.8	38	5.11	0	0.38	7	18.92	-0.02	6.77	12	10.4	0	1.22	14
Dunn	21.58	0.01	0.07	2	5	0	0.02	2	5.01	0	0.01	2	17.62	0	0.02	2	10.01	0	0.02	2
I	21.55	0	0.01	2	5.1	0	0.19	2	5.01	0	0.01	2	17.62	0	0.02	2	10.01	0	0.02	2
XB	21.55	0	0.01	2	5.04	0	0.09	3	5.01	0	0.01	2	17.62	0	0.02	2	10.01	0	0.02	2
S	21.56	0	0.02	2	5	0	0.02	2	**69.32**	**58.87**	**68.85**	24	24.83	22.11	50.31	35	11.46	7.73	33.2	96
CH	34.87	13.81	20.16	2	9.9	4.49	14.4	2	9.87	5.42	24.15	2	32.96	24.4	40.34	2	16.73	4.27	7.83	2
DB	21.55	0	0.01	2	5.04	0	0.09	3	5.01	0	0.01	2	17.62	0	0.02	2	10.01	0	0.02	2
S_Dbw	21.6	0	0.09	2	5.46	0	1.46	20	5.28	0	1.13	17	17.62	0	0.02	2	10.04	0	0.07	3
CVNN	21.87	0.01	0.63	4	6.1	0.03	2.27	7	5.13	0	0.36	4	22.97	0.51	14.77	6	10.04	0	0.04	2
DCVI	21.55	0	0.01	2	6.24	0.03	3.8	38	5.01	0	0.01	2	17.62	0	0.02	2	10.01	0	0.02	2
DBCV	21.49	-0.03	0.74	3	12.8	0.76	13.8	6	21.17	2.26	29.04	18	18.42	0.25	0.86	3	16.78	0.34	12.52	22
AIC	23.92	18.75	47.28	39	31.33	21.43	40.99	55	35.87	35.11	61.24	62	23.08	20.95	49.89	36	25.25	16.81	**34.92**	35
BIC	35.67	15.15	22.76	2	9.90	4.88	15.77	2	9.88	5.96	24.34	2	32.96	22.36	38.61	2	17.64	5.83	10.69	2
IP	**57.30**	**43.66**	**53.67**	12	**46.85**	**27.87**	**41.49**	17	54.99	19.35	61.4	14	**55.19**	**42.03**	**53.75**	10	**44.1**	**20.28**	30.29	7

Table 7: SimCSE based clustering results on five text datasets. (15 cases: 5 datasets × 3 evaluation metrics)

	SearchSnippets - 8				**Biomedical - 20**				**StackOverflow - 20**				**WebofScience - 7**				**Yahoo!Answers - 10**			
	ACC	ARI	NMI	opt_k	ACC	ARI	NMI	opt_k	ACC	ARI	NMI	opt_k	ACC	ARI	NMI	opt_k	ACC	ARI	NMI	opt_k
SD	21.65	0.05	1.03	11	8.86	0.25	7.34	12	6.64	0.02	3.18	10	17.62	0	0.02	2	10.01	0	0.02	2
Dunn	21.56	0	0.01	2	5	0	0.02	2	5.02	0	0.03	2	17.62	0	0.02	2	10.01	0	0.02	2
I	21.56	0	0.01	2	5	0	0.02	2	5.04	0	0.08	2	17.62	0	0.02	2	10.01	0	0.02	2
XB	21.56	0	0.01	2	5	0	0.02	2	5.04	0	0.08	2	17.62	0	0.02	2	10.01	0	0.02	2
S	21.56	0	0.01	2	5.03	0	0.08	2	5.02	0	0.06	2	17.62	0	0.02	2	10.01	0	0.02	2
CH	30.44	6.29	10.4	2	9.75	3.21	10.21	2	9.21	3.1	9.67	2	31.69	21.95	35.76	2	17.7	4.22	7.92	2
DB	21.56	0	0.01	2	5.1	0	0.43	9	5.04	0	0.08	2	17.62	0	0.02	2	10.01	0	0.02	2
S_Dbw	21.6	0	0.09	2	6.1	0.03	2.85	22	5.45	0	1.12	14	17.62	0	0.02	2	10.04	0	0.07	3
CVNN	21.65	-0.01	0.52	4	9.15	0.34	7.9	5	8.86	0.21	7.41	4	18.13	0.04	1.01	2	10.03	0	0.04	2
DCVI	21.56	0	0.01	2	5	0	0.02	2	6.33	0.01	3.89	42	17.62	0	0.02	2	10.01	0	0.02	2
DBCV	21.04	0.27	13.25	114	8.91	1.93	6.03	3	9.29	1.65	5.86	3	18.9	0.27	0.83	3	10.09	0	0.2	4
AIC	24.42	16.24	37.81	3	**26.95**	**16.24**	**32.03**	46	41.91	34.70	55.67	52	**55.65**	**38.40**	**44.50**	8	**22.73**	**11.32**	**23.95**	30
BIC	31.38	9.01	14.31	2	9.79	2.99	9.59	2	7.96	1.15	3.42	2	32.04	22.47	36.68	2	18.14	5.02	9.49	2
IP	**50.33**	**30.74**	**39.79**	12	11.31	0.99	12.65	3	**68.74**	**46.44**	**59.22**	20	48.98	33.74	**44.50**	11	14.72	0.92	6.19	2

the best partitions in terms of ACC, ARI and NMI are found by our IP index again, which indicates that the representation model can influence the clustering result. Moreover, when comparing with the results of ImageNet-10 in Table 9 and 10 it is interesting to find that: (1) Different representation models may have great impacts on the clustering result, e.g., the best ARI score of Swin based clustering can be as high as 99.64 while it drops to 0.65 dramatically when the representation model is changed to BEiT; (2) Even for the same representation model, the optimal clustering results found by different internal indices could vary a lot, e.g., the ARI score of I index is 0 while our IP index and several indices can provide optimal partition ARI score of 99.64.

Table 8: ViT based clustering results on five image datasets. (15 cases: 5 datasets × 3 evaluation metrics)

	CIFAR-10 - 10				**MNIST - 10**				**FashionMNIST - 10**				**ImageNet-10 - 10**				**CINIC-10 - 10**			
	ACC	ARI	NMI	opt_k	ACC	ARI	NMI	opt_k	ACC	ARI	NMI	opt_k	ACC	ARI	NMI	opt_k	ACC	ARI	NMI	opt_k
SD	10.49	0	1.39	7	11.34	0	0.02	2	10.01	0	0.02	2	34.34	5.11	36.03	9	10.01	0	0.02	2
Dunn	10.01	0	0.02	2	11.34	0	0.02	2	10.01	0	0.02	2	10.03	0	0.06	2	10.01	0	0.02	2
I	10.01	0	0.02	2	11.34	0	0.02	2	10.01	0	0.02	2	10.04	0	0.08	2	10.01	0	0.02	2
XB	10.01	0	0.02	2	11.34	0	0.02	2	10.01	0	0.02	2	87.52	82.8	91.19	9	10.01	0	0.02	2
S	10.01	0	0.02	2	11.34	0	0.02	2	10.02	0	0.04	2	89.13	88.55	90.82	13	10.01	0	0.02	2
CH	19.55	8.52	29.94	2	15.87	1.9	3.53	2	19.81	11.31	25.14	2	**92.21**	**89.12**	**93.75**	9	19.02	7.6	26.43	2
DB	10.01	0	0.02	2	11.34	0	0.02	2	10.01	0	0.02	2	10.04	0	0.08	2	10.01	0	0.02	2
S_Dbw	11.52	0.02	4.32	16	11.34	0	0.02	2	10.01	0	0.02	2	20.41	1.38	18.56	14	10.03	0	0.08	2
CVNN	11.09	0.02	2.75	7	12.95	0.12	2.46	2	12.82	0.34	4.56	3	15.5	0.43	10.18	4	10.86	0.03	1.71	2
DCVI	10.01	0	0.02	2	11.34	0	0.02	2	10.01	0	0.02	2	10.03	0	0.06	2	10.01	0	0.02	2
DBCV	11.46	0.13	1.51	3	14.66	1.24	3.17	3	15.69	1.79	4.44	3	67.96	45.14	71.31	10	28.71	4.25	30.43	7
AIC	28.21	26.67	57.69	46	13.55	**7.57**	**22.90**	40	16.94	**11.68**	**33.89**	41	33.01	36.74	72.87	54	27.31	21.82	48.74	44
BIC	19.55	8.52	29.94	2	**16.02**	1.99	3.69	2	**19.87**	11.66	26.08	2	19.74	6.44	27.37	2	19.02	7.60	26.43	2
IP	**75.87**	**58.89**	**70.24**	10	14.52	0.43	3.28	2	12.48	0.39	4.11	2	53.08	17.88	56.16	8	**59.47**	**38.70**	**54.43**	10

Table 9: Swin based clustering results on five image datasets. (15 cases: 5 datasets × 3 evaluation metrics)

	CIFAR-10 - 10				**MNIST - 10**				**FashionMNIST - 10**				**ImageNet-10 - 10**				**CINIC-10 - 10**			
	ACC	ARI	NMI	opt_k	ACC	ARI	NMI	opt_k	ACC	ARI	NMI	opt_k	ACC	ARI	NMI	opt_k	ACC	ARI	NMI	opt_k
SD	10.01	0	0.02	2	15.31	0.78	8.24	3	10.01	0	0.02	2	56.1	21.67	59.71	10	10.01	0	0.02	2
Dunn	10.01	0	0.02	2	11.37	0	0.04	2	10.01	0	0.02	2	10.02	0	0.03	2	10.01	0	0.02	2
I	10.01	0	0.02	2	11.37	0	0.04	2	10.01	0	0.02	2	10.02	0	0.03	2	10.01	0	0.02	2
XB	10.01	0	0.02	2	11.37	0	0.04	2	10.01	0	0.02	2	10.02	0	0.03	2	10.01	0	0.02	2
S	10.01	0	0.02	2	11.37	0	0.04	2	10.04	0	0.08	2	10.02	0	0.03	2	10.04	0	0.08	2
CH	20.0	17.35	40.25	2	19.47	4.94	9.45	2	19.99	13.57	33.58	2	**99.84**	**99.64**	**99.47**	10	19.89	14.32	30.65	2
DB	10.01	0	0.02	2	11.15	-0.01	0.38	2	10.01	0	0.02	2	10.02	0	0.03	2	10.01	0	0.02	2
S_Dbw	13.62	0.13	7.21	11	13.77	0.31	6.63	4	10.52	0	1.16	7	13.92	0.2	8.88	13	10.03	0	0.08	2
CVNN	12.9	0.11	5.5	5	13.63	0.29	6.37	3	11.31	0.03	2.25	4	12.82	0.09	5.55	6	14.87	0.27	9.23	8
DCVI	10.01	0	0.02	2	11.38	0	0.07	2	10.01	0	0.02	2	10.12	0	0.27	2	10.01	0	0.02	2
DBCV	23.44	1.51	23.05	15	16.18	1.23	9.44	3	13.68	0.36	7.42	3	57.78	22.57	60.58	10	22.38	1.95	24.05	12
AIC	43.81	47.01	73.79	33	16.28	**8.50**	**18.35**	32	21.25	13.79	34.29	30	50.88	55.04	79.93	42	35.38	29.84	54.39	33
BIC	20.00	17.35	40.25	2	**19.49**	4.99	9.56	2	17.47	6.73	15.88	2	19.81	4.56	23.70	2	19.89	14.32	30.65	2
IP	**95.38**	**90.12**	**90.14**	10	16.74	1.48	8.57	2	**35.18**	**20.10**	**34.48**	7	**99.84**	**99.64**	**99.47**	10	**68.57**	**52.89**	**62.36**	10

Table 10: BEiT based clustering results on five image datasets. (15 cases: 5 datasets × 3 evaluation metrics)

	CIFAR-10 - 10				**MNIST - 10**				**FashionMNIST - 10**				**ImageNet-10 - 10**				**CINIC-10 - 10**			
	ACC	ARI	NMI	opt_k	ACC	ARI	NMI	opt_k	ACC	ARI	NMI	opt_k	ACC	ARI	NMI	opt_k	ACC	ARI	NMI	opt_k
SD	10.01	0	0.02	2	11.36	0	0.02	2	10.01	0	0.02	2	10.01	0	0.02	2	10.01	0	0.02	2
Dunn	10.01	0	0.02	2	11.38	0	0.06	2	10.01	0	0.02	2	10.01	0	0.02	2	10.03	0	0.08	2
I	10.01	0	0.02	2	11.36	0	0.02	2	10.01	0	0.02	2	10.01	0	0.02	2	10.01	0	0.02	2
XB	10.01	0	0.02	2	11.36	0	0.02	2	10.01	0	0.02	2	10.01	0	0.02	2	10.01	0	0.02	2
S	10.01	0	0.02	2	11.36	0	0.02	2	10.01	0	0.02	2	10.01	0	0.02	2	10.02	0	0.04	2
CH	17.21	4.97	9.21	2	20.38	5.66	13.48	2	19.67	11.79	24.09	2	23.77	7.23	20.93	3	17.09	3.42	6.93	2
DB	10.01	0	0.02	2	11.36	0	0.02	2	10.01	0	0.02	2	10.01	0	0.02	2	10.01	0	0.02	2
S_Dbw	10.01	0	0.02	2	10.58	-0.04	1.72	10	12.08	0.23	6.15	20	11.46	0.05	3.18	12	10.03	0	0.08	2
CVNN	10.18	0	0.36	2	20.73	2.42	14.48	3	21.9	2.85	22.8	6	14.45	0.77	5.73	3	11.32	0.07	0.99	2
DCVI	10.01	0	0.02	2	11.36	0	0.02	2	10.01	0	0.02	2	10.01	0	0.02	2	10.01	0	0.02	2
DBCV	12.28	0.23	1.62	4	12.05	0.05	1.42	3	10.18	0	0.19	3	13.15	0.39	3.34	4	10.01	0	0.03	2
AIC	13.15	5.37	**16.88**	43	17.76	**10.41**	**27.92**	37	22.93	17.30	**41.39**	36	16.72	**9.69**	**28.51**	57	11.92	4.10	**13.84**	42
BIC	17.44	5.17	9.61	2	20.41	5.71	13.57	2	19.78	12.05	24.75	2	18.29	5.62	13.19	2	17.24	3.53	7.44	2
IP	**24.71**	**8.01**	13.62	7	**23.85**	6.48	18.36	3	**37.62**	**20.75**	38.08	7	14.01	0.65	5.23	2	**21.17**	**4.69**	9.88	7

6 CONCLUSIONS

In this paper, we propose internal purity (IP), a novel internal validation index. IP index uses the differential entropy to measure the purity of a cluster and the cluster centers of a partition. Based on the theoretical analysis, the nature of our IP index can help overcome the effectiveness and efficiency drawbacks of existing internal indices. Extensive experiments over different datasets of text and image domains also show that, our IP index can significantly outperform other thirteen well known internal indices when selecting the optimal partition and cluster number with different deep representation models. Although normal distribution assumption is indeed restrictive in our IP index, it won't affect the usage of our method over the data after deep representation. Considering that deep representation is already widely used, we believe that our method is still very practical.

REFERENCES

Nabil Ali Ahmed and DV Gokhale. Entropy expressions and their estimators for multivariate distributions. *IEEE Transactions on Information Theory*, 1989.

Alfred V Aho and John E Hopcroft. *The design and analysis of computer algorithms*. Pearson Education India, 1974.

Hirotugu Akaike. A new look at the statistical model identification. *IEEE transactions on automatic control*, 1974.

Serkan Akogul and Murat Erisoglu. A comparison of information criteria in clustering based on mixture of multivariate normal distributions. *Mathematical and Computational Applications*, 2016.

Hangbo Bao, Li Dong, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv*, 2021.

James C Bezdek and Nikhil R Pal. Some new indexes of cluster validity. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 1998.

Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*. Springer, 2006.

Tadeusz Caliński and Jerzy Harabasz. A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, 1974.

Jianlong Chang, Lingfeng Wang, Gaofeng Meng, Shiming Xiang, and Chunhong Pan. Deep adaptive image clustering. In *ICCV*, 2017.

Wen-Yen Chen, Yangqiu Song, Hongjie Bai, Chih-Jen Lin, and Edward Y Chang. Parallel spectral clustering in distributed systems. *IEEE transactions on pattern analysis and machine intelligence*, 2010.

Chun-Hung Cheng, Ada Waichee Fu, and Yi Zhang. Entropy-based subspace clustering for mining numerical data. In *SIGKDD*, 1999.

Dongdong Cheng, Qingsheng Zhu, Jinlong Huang, Quanwang Wu, and Lijun Yang. A novel cluster validity index based on local cores. *IEEE transactions on neural networks and learning systems*, 2018.

Thomas M Cover and Joy A Thomas. Information theory and statistics. *Elements of information theory*, 1991.

Hadi Daneshmand, Amir Joudaki, and Francis Bach. Batch normalization orthogonalizes representations in deep random networks. 2021.

Zhiyuan Dang, Cheng Deng, Xu Yang, Kun Wei, and Heng Huang. Nearest neighbor matching for deep clustering. In *CVPR*, 2021.

Luke N Darlow, Elliot J Crowley, Antreas Antoniou, and Amos J Storkey. Cinic-10 is not imagenet or cifar-10. *arXiv*, 2018.

David L Davies and Donald W Bouldin. A cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence*, 1979.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, 2019.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.

Joseph C Dunn. A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters. 1973.

Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In Evangelos Simoudis, Jiawei Han, and Usama M. Fayyad (eds.), *KDD*, 1996.

Chris Fraley and Adrian E Raftery. How many clusters? which clustering method? answers via model-based cluster analysis. *The computer journal*, 1998.

Andrew Frank. Uci machine learning repository. *http://archive. ics. uci. edu/ml*, 2010.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. Simcse: Simple contrastive learning of sentence embeddings. In *EMNLP*, 2021.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

Naiyang Guan, Dacheng Tao, Zhigang Luo, and Bo Yuan. Nenmf: An optimal gradient method for nonnegative matrix factorization. *IEEE Transactions on Signal Processing*, 2012a.

Naiyang Guan, Dacheng Tao, Zhigang Luo, and Bo Yuan. Online nonnegative matrix factorization with robust stochastic approximation. *IEEE Transactions on Neural Networks and Learning Systems*, 2012b.

Maria Halkidi and Michalis Vazirgiannis. Clustering validity assessment: Finding the optimal partitioning of a data set. In *ICDM*, 2001.

Maria Halkidi, Michalis Vazirgiannis, and Yannis Batistakis. Quality scheme assessment in the clustering process. In *PKDD*, 2000.

Maria Halkidi, Yannis Batistakis, and Michalis Vazirgiannis. On clustering validation techniques. *Journal of intelligent information systems*, 2001.

Xuelei Hu and Lei Xu. Investigation on several model selection criteria for determining the number of cluster. *Neural Information Processing-Letters and Reviews*, 2004.

Jiabo Huang, Shaogang Gong, and Xiatian Zhu. Deep semantic clustering by partition confidence maximisation. In *CVPR*, 2020.

Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of classification*, 2(1):193–218, 1985.

Richard Arnold Johnson, Dean W Wichern, et al. *Applied multivariate statistical analysis*. Pearson London, UK:, 2014.

Kamran Kowsari, Donald E Brown, Mojtaba Heidarysafa, Kiana Jafari Meimandi, Matthew S Gerber, and Laura E Barnes. Hdltex: Hierarchical deep learning for text classification. In *ICMLA*, 2017.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 1998.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *NIPS*, 2018.

Yanchi Liu, Zhongmou Li, Hui Xiong, Xuedong Gao, Junjie Wu, and Sen Wu. Understanding and enhancement of internal clustering validation measures. *IEEE transactions on cybernetics*, 2013.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021.

Christopher D Manning. *Introduction to information retrieval*. Syngress Publishing, 2008.

Ujjwal Maulik and Sanghamitra Bandyopadhyay. Performance evaluation of some clustering algorithms and validity indices. *IEEE Transactions on pattern analysis and machine intelligence*, 2002.

Davoud Moulavi, Pablo A. Jaskowiak, Ricardo J. G. B. Campello, Arthur Zimek, and Jörg Sander. Density-based clustering validation. In *SDM*, 2014.

Nikhil R Pal and James C Bezdek. On cluster validity for the fuzzy c-means model. *IEEE Transactions on Fuzzy systems*, 1995.

Christos H Papadimitriou and Kenneth Steiglitz. *Combinatorial optimization: algorithms and complexity*. Courier Corporation, 1998.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 2011.

Xuan-Hieu Phan, Le-Minh Nguyen, and Susumu Horiguchi. Learning to classify short and sparse text & web with hidden topics from large-scale data collections. In *WWW*, 2008.

Calyampudi Radhakrishna Rao, Calyampudi Radhakrishna Rao, Mathematischer Statistiker, Calyampudi Radhakrishna Rao, and Calyampudi Radhakrishna Rao. *Linear statistical inference and its applications*. Wiley New York, 1973.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *EMNLP*, 2019.

Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 1987.

Gideon Schwarz. Estimating the dimension of a model. *The annals of statistics*, 1978.

Claude Elwood Shannon. A mathematical theory of communication. *The Bell system technical journal*, 1948.

Lucas Vendramin, Ricardo JGB Campello, and Eduardo R Hruschka. Relative clustering validity criteria: A comparative overview. *Statistical analysis and data mining: the ASA data science journal*, 2010.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, et al. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 2020.

Joe H Ward Jr. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 1963.

Michael P Windham and Adele Cutler. Information ratios for validating mixture analyses. *Journal of the American Statistical Association*, 1992.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Transformers: State-of-the-art natural language processing. In *EMNLP*, 2020.

Junjie Wu, Hui Xiong, and Jian Chen. Adapting the right measures for k-means clustering. In *KDD*, 2009.

Mingrui Wu and Bernhard Schölkopf. A local learning approach for clustering. In *NIPS*, 2006.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv*, 2017.

Jiang Xie, Zhong-Yang Xiong, Qi-Zhu Dai, Xiao-Xia Wang, and Yu-Fang Zhang. A new internal index based on density core for clustering validation. *Information Sciences*, 2020.

Junyuan Xie, Ross Girshick, and Ali Farhadi. Unsupervised deep embedding for clustering analysis. In *ICML*, 2016.

Xuanli Lisa Xie and Gerardo Beni. A validity measure for fuzzy clustering. *IEEE Transactions on pattern analysis and machine intelligence*, 1991.

Jiaming Xu, Bo Xu, Peng Wang, Suncong Zheng, Guanhua Tian, and Jun Zhao. Self-taught convolutional neural networks for short text clustering. *Neural Networks*, 2017.

Rui Xu and Donald Wunsch. Survey of clustering algorithms. *IEEE Transactions on neural networks*, 2005.

Bo Yang, Xiao Fu, Nicholas D Sidiropoulos, and Mingyi Hong. Towards k-means-friendly spaces: simultaneous deep learning and clustering. In *ICML*, 2017.

Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In *NIPS*, 2015.

Tianyi Zhou, Dacheng Tao, and Xindong Wu. Manifold elastic net: a unified framework for sparse dimension reduction. *Data Mining and Knowledge Discovery*, 2011.

Mengjie Zu, Arunkumar Bupathy, Daan Frenkel, and Srikanth Sastry. Information density, structure and entropy in equilibrium and non-equilibrium systems. *Journal of Statistical Mechanics: Theory and Experiment*, 2020.

A DERIVATION OF EQ. 11

$$
\begin{aligned}
&IP = CP - SP \\
&= \left[\frac{1}{k}\sum_{i=1}^{k} DiffEntropy(H_i)\right] - DiffEntropy(H_\mu) \qquad \text{According to Eq. 7 and 8} \\
&= \frac{1}{k}\sum_{i=1}^{k}\left[\frac{rank(\mathbf{\Sigma}_i)}{2} + \frac{rank(\mathbf{\Sigma}_i)}{2}\ln(2\pi) + \frac{1}{2}\ln|\mathbf{\Sigma}_i|\right] - \left[\frac{rank(\mathbf{\Sigma}_\mu)}{2} + \frac{rank(\mathbf{\Sigma}_\mu)}{2}\ln(2\pi) + \frac{1}{2}\ln|\mathbf{\Sigma}_\mu|\right] \\
&\qquad \text{According to Eq. 6} \\
&= \frac{1}{k}\left[\sum_{i=1}^{k}\frac{rank(\mathbf{\Sigma}_i)}{2} + \sum_{i=1}^{k}\frac{rank(\mathbf{\Sigma}_i)}{2}\ln(2\pi) + \sum_{i=1}^{k}\frac{1}{2}\ln|\mathbf{\Sigma}_i|\right] - \left[\frac{rank(\mathbf{\Sigma}_\mu)}{2} + \frac{rank(\mathbf{\Sigma}_\mu)}{2}\ln(2\pi) + \frac{1}{2}\ln|\mathbf{\Sigma}_\mu|\right] \\
&= \frac{1}{2k}\left\{\sum_{i=1}^{k}[1+\ln(2\pi)]\,rank(\mathbf{\Sigma}_i) + \sum_{i=1}^{k}\ln|\mathbf{\Sigma}_i|\right\} - \frac{1}{2}\left\{[1+\ln(2\pi)]\,rank(\mathbf{\Sigma}_\mu) + \ln|\mathbf{\Sigma}_\mu|\right\} \\
&= \frac{1}{2}\left(\frac{1}{k}\sum_{i=1}^{k}\ln|\mathbf{\Sigma}_i| - \ln|\mathbf{\Sigma}_\mu|\right) + \frac{1+\ln(2\pi)}{2}\left[\frac{1}{k}\sum_{i=1}^{k}rank(\mathbf{\Sigma}_i) - rank(\mathbf{\Sigma}_\mu)\right]
\end{aligned}
$$

B EVALUATION METRICS

In our evaluations we used three common external validation indices: Accuracy (ACC), Adjusted Rand Index (ARI) and Normalized Mutual Information (NMI). The ACC and NMI range between 0 and 1, and the ARI ranges between -1 and 1. Larger ACC, ARI and NMI indicates better clustering result.

ACC is computed as follows:

$$
\mathrm{ACC} = \frac{\sum_{i=1}^{N}\delta\left(y_i, \mathrm{map}\left(c_i\right)\right)}{N} \tag{12}
$$

where y_i is the true cluster label, c_i is the cluster label obtained by clustering, and $\delta(x, y)$ is an indicator function returning 0 ($x \neq y$) or 1 ($x = y$). $map(\cdot)$ transforms the cluster label c_i to its true cluster label by the Hungarian algorithm (Papadimitriou & Steiglitz, 1998). The larger the ACC value is, the better the clustering result.

ARI is computed as follows:

$$
\mathrm{ARI} = \frac{\mathrm{RI} - E[\mathrm{RI}]}{1 - E[\mathrm{RI}]}, \qquad \mathrm{RI} = \frac{a+b}{C_2^N} \tag{13}
$$

where C_2^N is the total number of possible pairs in the dataset, a is the number of pairs of samples that are in the same ground truth class and in the same cluster, b is the number of pairs of samples that are in different ground truth class and in different cluster. $E[\mathrm{RI}]$ is the expected RI. Larger ARI indicates better clustering result.

NMI is computed as follows:

$$\mathrm{NMI}(A, B) = \frac{I(A, B)}{\sqrt{H(A)H(B)}} \tag{14}$$

where A is the predicted labels and B the ground truth labels. I is the mutual information and H is the entropy.

C MODEL CONFIGURATION

BERT is based on *bert-base-uncased*, SBERT is based on *all-distilroberta-v1*, SimCSE is based on *unsup-simcse-roberta-base*, ViT is based on *vit-base-patch16-224-in21k*, Swin is based on *swin-base-patch4-window7-224* and BEiT is based on *beit-base-patch16-224-pt22k*, where BERT, SBERT, SimCSE, ViT and BEiT return 768 embedding size, and Swin returns 1024 embedding size.

D CLUSTERING EVALUATION RESULTS UNDER A SINGLE CLUSTERING ALGORITHM

As before, we first present the statistical results of each clustering algorithm and report them in Table 11. Obviously, compared with other 13 indices, our IP index has significant advance in achieving the best results under each clustering algorithm. Specifically, for all 90 cases, our IP index can produce 52 best results, while the second top index SD only has 13 best results in K-Means. Our IP index can produce 44 best results, while the second top index S only has 15 best results in GMM. Our index can produce 37 best results, while the second top index DBCV only has 17 best results in AHC. Although CH can produce 63 best results while our index only has 26 best results in DBSCAN, our index is better than CH in the other three clustering algorithms. Finally, IP index outperforms other indices greatly with respect to different domains, i.e., text and image. Then the specific results of each clustering algorithm can been in D.1, D.2, D.3 and D.4, respectively.

Table 11: Number of best clustering results based on counting over K-Means, GMM, AHC and DBSCAN

	SD	Dunn	I	XB	S	CH	DB	S_Dbw	CVNN	DCVI	DBCV	AIC	BIC	**IP**
							K-Means							
Text datasets	4	2	0	0	4	0	0	2	1	2	1	0	0	**30**
Image datasets	9	3	0	6	4	3	7	5	6	2	7	1	1	**22**
All datasets	13	5	0	6	8	3	7	7	7	4	8	1	1	**52**
							GMM							
Text datasets	1	0	0	0	6	0	0	3	0	0	0	10	0	**25**
Image datasets	3	0	0	2	9	0	2	2	1	1	4	1	0	**19**
All datasets	4	0	0	2	15	0	2	5	1	1	4	11	0	**44**
							AHC							
Text datasets	4	8	0	2	10	0	5	6	0	6	6	0	0	**15**
Image datasets	12	4	0	3	4	3	3	8	6	6	11	0	0	**22**
All datasets	16	12	0	5	14	3	8	14	6	12	17	0	0	**37**
							DBSCAN							
Text datasets	0	0	0	0	0	**33**	0	0	1	0	4	0	0	7
Image datasets	0	0	0	0	3	**30**	0	0	3	0	1	0	0	19
All datasets	0	0	0	0	3	**63**	0	0	4	0	5	0	0	26

D.1 THE CLUSTERING RESULTS ON K-MEANS

In this section, we only use K-Means to evaluate the clustering results. Since the random initialization nature of K-Means, the experimental results are averaged over five random runs for each validation index. The evaluation results based on external indices of ACC and ARI are shown in Table 12, 13, 14, 15, 16, and 17, the evaluation results in terms of NMI are shown in Table 18, 19, 20, 21, 22 and 23, where the best results are highlighted in bold. Moreover, the optimal k results each index select can be seen in Fig. 2, 3, 4, 5, 6 and 7. Obviously, for almost all cases, our IP index outperforms other indices and is close to the real k value represented in red dash line.

Table 12: BERT based K-Means clustering results on five text datasets.

	SearchSnippets		**Biomedical**		**StackOverflow**		**WebofScience**		**Yahoo!Answers**	
	ACC	ARI	ACC	ARI	ACC	ARI	ACC	ARI	ACC	ARI
SD	27.98±10.23	17.91±6.21	29.31±0.01	14.64±0.01	8.42±0.01	0.90±0.00	40.66±0.01	24.36±0.02	19.97±0.03	2.59±0.02
Dunn	27.39±12.76	17.35±7.89	12.79±0.76	7.4±0.70	7.65±3.05	1.47±2.91	8.48±0.44	6.24±0.27	**25.98±8.58**	**10.18±3.11**
I	33.63±0.03	12.22±0.06	9.66±0.00	2.52±0.01	6.29±0.00	0.17±0.00	29.39±0.01	17.99±0.01	12.90±0.03	0.70±0.01
XB	17.22±3.37	11.19±2.12	29.31±0.01	14.64±0.01	8.42±0.01	0.90±0.00	29.39±0.01	17.99±0.01	12.90±0.03	0.70±0.01
S	10.05±0.59	6.66±0.24	9.66±0.00	2.52±0.01	8.42±0.01	0.90±0.00	29.39±0.01	17.99±0.01	12.90±0.03	0.70±0.01
CH	33.63±0.03	12.22±0.06	9.66±0.00	2.52±0.01	6.29±0.00	0.17±0.00	29.39±0.01	17.99±0.01	12.90±0.03	0.70±0.01
DB	10.22±0.39	6.67±0.18	29.31±0.01	14.64±0.01	10.96±0.01	2.41±0.01	40.66±0.01	24.36±0.02	16.52±0.02	1.89±0.01
S_Dbw	9.85±0.51	6.60±0.15	11.38±0.43	6.58±0.23	12.92±0.66	**7.05±0.23**	8.20±0.39	6.09±0.12	12.59±1.53	5.19±0.56
CVNN	33.63±0.03	12.22±0.06	9.66±0.00	2.52±0.01	6.29±0.00	0.17±0.00	29.39±0.01	17.99±0.01	12.90±0.03	0.70±0.01
DCVI	10.18±0.93	6.70±0.41	11.59±0.29	6.74±0.30	13.14±0.65	6.98±0.35	8.60±0.21	6.22±0.19	11.91±0.94	5.03±0.27
DBCV	9.73±0.59	6.45±0.13	11.87±0.81	6.90±0.46	12.58±0.27	6.87±0.08	8.49±0.20	6.26±0.26	11.93±0.76	4.99±0.25
AIC	33.63±0.03	12.22±0.06	9.66±0.00	2.52±0.01	6.29±0.00	0.17±0.00	29.39±0.01	17.99±0.01	12.90±0.03	0.70±0.01
BIC	33.63±0.03	12.22±0.06	9.66±0.00	2.52±0.01	6.29±0.00	0.17±0.00	29.39±0.01	17.99±0.01	12.90±0.03	0.70±0.01
IP	**54.18±5.71**	**33.58±4.04**	**31.48±0.55**	**15.63±0.12**	**15.98±0.14**	4.09±0.04	**47.93±2.40**	**31.90±1.75**	19.97±0.03	2.59±0.02

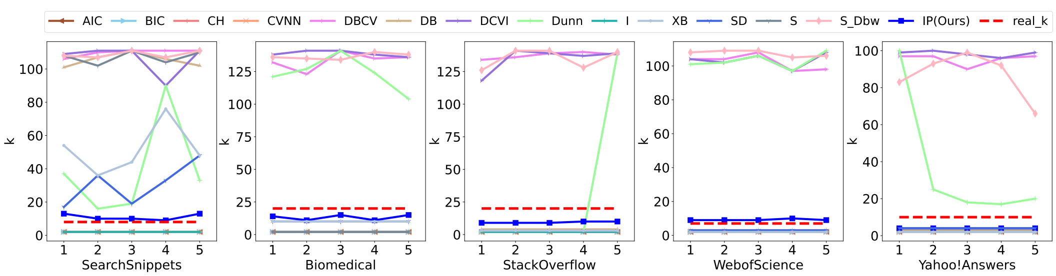


Figure 2: The optimal k value found by each index on BERT representations for text datasets.

Table 13: SBERT based K-Means clustering results on five text datasets.

	SearchSnippets		**Biomedical**		**StackOverflow**		**WebofScience**		**Yahoo!Answers**	
	ACC	ARI	ACC	ARI	ACC	ARI	ACC	ARI	ACC	ARI
SD	24.24±10.15	18.27±7.29	41.16±12.78	26.50±7.84	59.66±20.26	36.43±15.51	22.85±8.38	20.20±6.95	17.93±4.80	11.96±3.05
Dunn	16.94±4.66	13.09±3.74	21.37±6.59	14.58±4.54	28.78±13.07	17.78±11.57	15.82±9.60	13.22±8.97	22.08±11.56	13.23±4.27
I	34.87±0.01	13.81±0.01	9.90±0.01	4.49±0.01	13.47±2.01	6.32±0.96	32.96±0.00	24.40±0.00	16.73±0.02	4.27±0.02
XB	23.68±10.79	17.73±7.92	41.23±4.3	24.88±3.90	31.16±20.08	16.51±13.09	35.72±16.36	29.44±11.80	12.49±1.07	8.44±0.75
S	11.10±0.36	8.31±0.28	47.26±1.79	**29.9±0.98**	69.32±1.48	**58.87±1.17**	24.83±3.17	22.11±2.29	11.46±0.48	7.73±0.33
CH	34.87±0.01	13.81±0.01	9.90±0.01	4.49±0.01	9.87±0.00	5.42±0.00	32.96±0.00	24.40±0.00	16.73±0.02	4.27±0.02
DB	10.94±0.19	8.21±0.16	16.88±0.39	11.30±0.39	53.35±17.02	30.44±12.35	29.03±3.31	25.34±2.50	11.41±0.36	7.54±0.19
S_Dbw	10.92±0.40	8.23±0.03	16.42±0.75	10.84±0.29	20.32±0.86	19.40±1.21	9.94±0.27	8.28±0.17	11.27±0.50	7.54±0.26
CVNN	34.87±0.01	13.81±0.01	9.90±0.01	4.49±0.01	20.76±5.32	9.43±2.33	32.96±0.00	24.40±0.00	23.42±0.01	10.14±0.04
DCVI	10.79±0.39	8.18±0.20	17.04±1.47	11.52±0.99	63.17±12.44	40.40±9.5	10.79±0.66	8.52±0.16	12.09±0.99	7.87±0.50
DBCV	10.88±0.13	8.26±0.18	16.96±0.93	11.40±0.78	55.40±5.40	43.65±8.77	10.67±0.80	8.66±0.42	11.36±0.25	7.52±0.13
AIC	34.87±0.01	13.81±0.01	9.90±0.01	4.49±0.01	9.87±0.00	5.42±0.00	32.96±0.00	24.40±0.00	16.73±0.02	4.27±0.02
BIC	34.87±0.01	13.81±0.01	9.90±0.01	4.49±0.01	9.87±0.00	5.42±0.00	32.96±0.00	24.40±0.00	16.73±0.02	4.27±0.02
IP	**55.68±4.06**	**42.57±2.72**	**47.34±1.16**	28.59±0.93	**72.61±1.62**	50.25±3.00	**58.60±1.81**	**44.50±1.46**	**50.14±2.26**	**27.07±1.71**

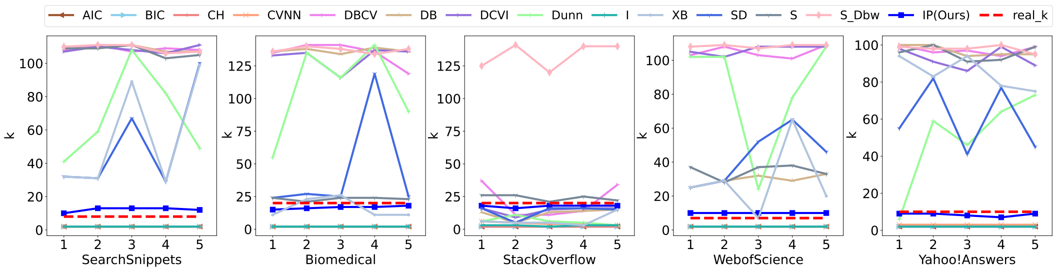


Figure 3: The optimal k value found by each index on SBERT representations for text datasets.

Table 14: SimCSE based K-Means clustering results on five text datasets.

	SearchSnippets		**Biomedical**		**StackOverflow**		**WebofScience**		**Yahoo!Answers**	
	ACC	ARI	ACC	ARI	ACC	ARI	ACC	ARI	ACC	ARI
SD	10.77±0.80	7.35±0.69	21.06±8.20	9.47±6.35	61.42±2.06	46.07±0.99	**49.01±1.77**	**34.49±1.44**	17.29±1.72	8.74±0.83
Dunn	24.24±13.22	15.72±8.27	23.10±1.61	11.34±1.41	38.60±14.71	28.25±6.46	22.47±17.29	16.78±11.43	30.93±11.3	12.3±3.89
I	30.44±0.01	6.29±0.01	9.75±0.00	3.21±0.00	9.21±0.02	3.10±0.02	31.69±0.00	21.95±0.00	17.70±0.04	4.22±0.02
XB	11.25±1.54	7.49±0.94	17.39±0.01	6.63±0.01	60.40±3.08	45.69±1.47	43.59±0.00	28.19±0.01	11.22±1.66	5.49±0.97
S	10.52±0.62	6.82±0.44	9.75±0.00	3.21±0.00	63.36±1.61	**46.99±0.71**	31.69±0.00	21.95±0.00	17.70±0.04	4.22±0.02
CH	30.44±0.01	6.29±0.01	9.75±0.00	3.21±0.00	9.21±0.02	3.10±0.02	31.69±0.00	21.95±0.00	17.70±0.04	4.22±0.02
DB	10.82±0.67	6.98±0.19	14.76±1.06	8.93±0.54	63.25±3.11	46.71±1.33	43.59±0.00	28.19±0.01	10.26±0.25	4.89±0.10
S_Dbw	10.32±0.42	6.75±0.09	13.67±0.28	8.14±0.20	19.57±0.55	17.19±0.29	8.64±0.62	6.98±0.24	10.27±0.5	5.07±0.26
CVNN	30.44±0.01	6.29±0.01	17.39±0.01	6.63±0.01	9.21±0.02	3.10±0.02	31.69±0.00	21.95±0.00	17.70±0.04	4.22±0.02
DCVI	10.26±0.31	6.65±0.05	14.47±0.28	8.79±0.21	21.49±1.03	18.53±1.13	9.04±0.19	7.13±0.17	10.38±0.5	5.07±0.21
DBCV	10.04±0.22	6.67±0.20	13.99±0.74	8.44±0.52	21.71±1.54	18.57±1.72	9.32±0.48	7.17±0.23	10.41±0.46	5.06±0.24
AIC	30.44±0.01	6.29±0.01	9.75±0.00	3.21±0.00	9.21±0.02	3.10±0.02	31.69±0.00	21.95±0.00	17.70±0.04	4.22±0.02
BIC	30.44±0.01	6.29±0.01	9.75±0.00	3.21±0.00	9.21±0.02	3.10±0.02	31.69±0.00	21.95±0.00	17.70±0.04	4.22±0.02
IP	**50.33±0.82**	**30.74±0.99**	**37.96±0.19**	**20.06±0.10**	**68.74±1.38**	46.44±1.36	48.98±3.73	33.74±2.07	**37.03±0.15**	**14.75±0.16**

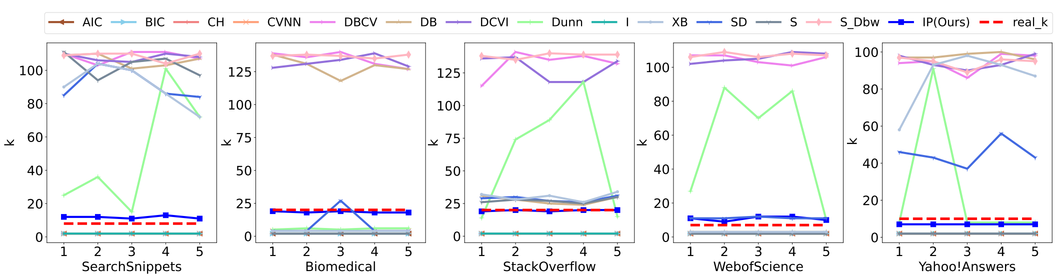


Figure 4: The optimal k value found by each index on SimCSE representations for text datasets.

Table 15: ViT based K-Means clustering results on five image datasets.

	CIFAR-10		**MNIST**		**FashionMNIST**		**ImageNet-10**		**CINIC-10**	
	ACC	ARI	ACC	ARI	ACC	ARI	ACC	ARI	ACC	ARI
SD	**76.07±0.96**	58.84±1.63	31.17±0.23	11.71±0.29	33.40±0.01	21.72±0.02	80.06±12.91	71.44±20.37	17.66±1.71	13.29±1.35
Dunn	23.10±4.86	10.71±3.00	29.19±0.89	10.64±0.31	35.25±2.63	20.33±0.51	57.15±38.37	51.34±42.69	29.28±3.54	20.03±2.69
I	19.55±0.00	8.52±0.00	15.87±0.00	1.90±0.00	19.81±0.00	11.31±0.00	23.68±5.32	9.90±4.10	19.02±0.00	7.60±0.00
XB	23.10±4.86	10.71±3.00	28.78±0.03	10.5±0.03	19.81±0.00	11.31±0.00	92.21±5.75	89.12±6.89	19.02±0.00	7.60±0.00
S	68.07±2.29	57.42±0.96	15.87±0.00	1.90±0.00	30.14±0.00	18.93±0.00	89.13±2.84	88.55±1.85	53.86±1.07	**41.14±0.69**
CH	19.55±0.00	8.52±0.00	15.87±0.00	1.90±0.00	19.81±0.00	11.31±0.00	92.21±5.75	89.12±6.89	19.02±0.00	7.60±0.00
DB	23.10±4.86	10.71±3.00	30.82±0.07	11.44±0.40	19.81±0.00	11.31±0.00	68.95±40.62	60.95±48.98	19.02±0.00	7.60±0.00
S_Dbw	18.60±4.88	16.85±4.55	7.50±0.21	4.41±0.07	8.92±0.24	6.16±0.07	66.31±16.12	50.61±24.24	15.33±1.65	11.99±1.1
CVNN	23.10±4.86	10.71±3.00	15.87±0.00	1.90±0.00	19.81±0.00	11.31±0.00	52.64±8.71	31.27±9.35	19.02±0.00	7.60±0.00
DCVI	19.70±4.89	15.08±0.63	7.61±0.33	4.38±0.13	8.89±0.11	6.21±0.19	84.74±30.79	82.49±31.77	15.05±0.76	11.73±0.43
DBCV	17.36±1.05	15.35±0.69	7.73±0.26	4.39±0.12	9.10±0.31	6.21±0.13	**92.36±8.83**	**91.64±7.17**	14.62±0.76	11.38±0.39
AIC	19.55±0.00	8.52±0.00	15.87±0.00	1.90±0.00	19.81±0.00	11.31±0.00	19.80±0.01	8.40±2.17	19.02±0.00	7.60±0.00
BIC	19.55±0.00	8.52±0.00	15.87±0.00	1.90±0.00	19.81±0.00	11.31±0.00	19.80±0.01	8.40±2.17	19.02±0.00	7.60±0.00
IP	75.87±0.03	**58.89±0.09**	**31.37±0.36**	**11.73±0.33**	**39.75±1.08**	**21.87±0.23**	72.14±13.20	58.87±20.65	**63.39±0.96**	40.94±2.68

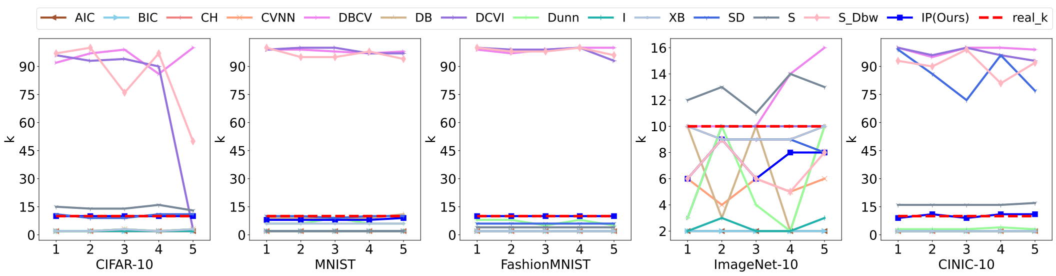


Figure 5: The optimal k value found by each index on ViT representations for image datasets.

Table 16: Swin based K-Means clustering results on five image datasets.

	CIFAR-10		**MNIST**		**FashionMNIST**		**ImageNet-10**		**CINIC-10**	
	ACC	ARI	ACC	ARI	ACC	ARI	ACC	ARI	ACC	ARI
SD	82.98±5.17	79.22±3.93	**27.69±0.03**	9.13±0.02	30.85±0.00	20.65±0.00	**99.84±0.00**	**99.64±0.00**	63.12±3.42	51.65±2.45
Dunn	81.19±6.46	78.84±6.64	22.27±0.60	**10.56±0.10**	19.99±0.00	13.57±0.00	83.78±15.16	82.85±16.46	35.64±13.51	29.44±11.03
I	20.00±0.00	17.35±0.00	19.47±0.01	4.94±0.00	19.99±0.00	13.57±0.00	19.81±0.06	4.56±0.07	19.89±0.00	14.32±0.01
XB	86.76±0.02	82.11±0.06	21.12±3.69	5.78±1.89	34.27±0.01	**20.76±0.03**	**99.84±0.00**	**99.64±0.00**	56.69±15.68	44.02±12.84
S	91.44±1.36	88.85±1.13	19.47±0.01	4.94±0.00	24.65±0.00	15.22±0.00	88.86±2.29	90.75±1.70	65.29±2.47	**52.94±2.19**
CH	20.00±0.00	17.35±0.00	19.47±0.01	4.94±0.00	19.99±0.00	13.57±0.00	**99.84±0.00**	**99.64±0.00**	19.89±0.00	14.32±0.01
DB	86.76±0.02	82.11±0.06	**27.69±0.03**	9.13±0.02	30.85±0.00	20.65±0.00	**99.84±0.00**	**99.64±0.00**	**68.87±0.11**	52.62±0.17
S_Dbw	20.07±0.51	19.90±0.41	8.32±0.37	4.38±0.32	9.05±0.44	6.20±0.13	**99.84±0.00**	**99.64±0.00**	20.72±3.52	16.93±3.28
CVNN	86.76±0.02	82.11±0.06	19.47±0.01	4.94±0.00	19.99±0.00	13.57±0.00	**99.84±0.00**	**99.64±0.00**	27.01±3.98	19.72±3.02
DCVI	20.66±0.86	20.28±1.29	8.19±0.34	4.25±0.12	9.70±0.68	6.62±0.28	89.81±14.18	89.18±15.15	18.52±2.05	14.81±1.31
DBCV	21.14±1.04	20.88±1.02	8.37±0.34	4.35±0.19	9.34±0.29	6.44±0.08	**99.84±0.00**	**99.64±0.00**	20.43±1.39	16.30±1.25
AIC	20.00±0.00	17.35±0.00	19.47±0.01	4.94±0.00	19.99±0.00	13.57±0.00	19.81±0.06	4.56±0.07	19.89±0.00	14.32±0.01
BIC	20.00±0.00	17.35±0.00	19.47±0.01	4.94±0.00	19.99±0.00	13.57±0.00	19.81±0.06	4.56±0.07	19.89±0.00	14.32±0.01
IP	**95.38±0.02**	**90.12±0.05**	27.30±0.01	8.60±0.01	**35.18±0.30**	20.10±0.67	**99.84±0.00**	**99.64±0.00**	68.05±1.77	51.80±1.74

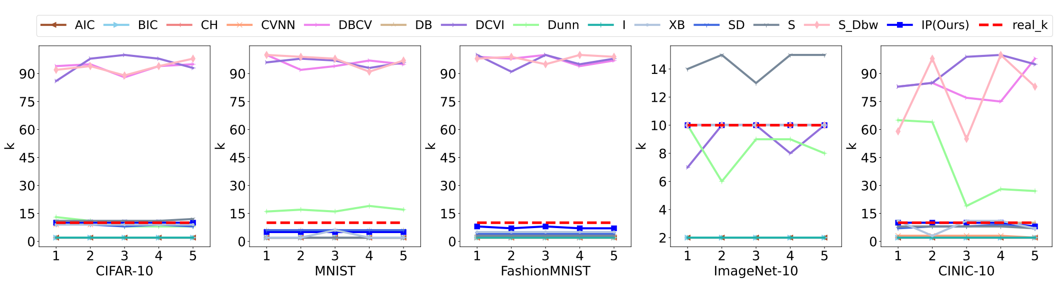


Figure 6: The optimal k value found by each index on Swin representations for image datasets.

Table 17: BEiT based K-Means clustering results on five image datasets.

	CIFAR-10		**MNIST**		**FashionMNIST**		**ImageNet-10**		**CINIC-10**	
	ACC	ARI	ACC	ARI	ACC	ARI	ACC	ARI	ACC	ARI
SD	21.49±2.10	6.26±0.84	**20.38±0.00**	5.66±0.00	30.89±0.00	20.06±0.00	32.44±5.43	16.65±2.08	18.10±0.91	5.83±0.07
Dunn	8.54±1.36	3.67±0.55	9.92±0.73	**6.57±0.32**	13.45±1.27	10.33±0.71	12.80±2.60	6.25±1.25	7.31±0.42	2.64±0.19
I	17.21±0.00	4.97±0.01	19.77±0.01	5.72±0.00	19.67±0.00	11.79±0.00	23.77±0.00	7.23±0.00	18.03±0.07	2.97±0.01
XB	22.20±0.02	6.24±0.01	19.77±0.01	5.72±0.00	19.67±0.00	11.79±0.00	23.77±0.00	7.23±0.00	18.76±0.02	3.39±0.01
S	17.21±0.00	4.97±0.01	19.77±0.01	5.72±0.00	19.67±0.00	11.79±0.00	17.37±0.03	4.08±0.03	17.09±0.00	3.42±0.00
CH	17.21±0.00	4.97±0.01	**20.38±0.00**	5.66±0.00	19.67±0.00	11.79±0.00	23.77±0.00	7.23±0.00	17.09±0.00	3.42±0.00
DB	12.92±1.37	5.44±0.61	19.77±0.01	5.72±0.00	19.67±0.00	11.79±0.00	23.77±0.00	7.23±0.00	18.76±0.02	3.39±0.01
S_Dbw	7.63±0.33	3.29±0.15	9.73±0.48	6.46±0.29	12.51±0.58	9.86±0.53	11.64±1.24	6.77±0.63	7.04±0.47	2.63±0.20
CVNN	17.21±0.00	4.97±0.01	**20.38±0.00**	5.66±0.00	**41.49±2.10**	**23.91±0.88**	31.62±0.03	12.25±0.03	17.09±0.00	3.42±0.00
DCVI	7.66±0.16	3.30±0.10	9.48±0.27	6.42±0.12	12.13±0.90	9.51±0.59	11.44±0.52	6.68±0.31	7.14±0.43	2.56±0.16
DBCV	7.76±0.24	3.33±0.08	9.22±0.29	6.22±0.11	12.15±0.83	9.37±0.35	11.06±0.25	6.58±0.15	7.25±0.63	2.59±0.15
AIC	17.21±0.00	4.97±0.01	**20.38±0.00**	5.66±0.00	19.67±0.00	11.79±0.00	17.37±0.03	4.08±0.03	17.09±0.00	3.42±0.00
BIC	17.21±0.00	4.97±0.01	**20.38±0.00**	5.66±0.00	19.67±0.00	11.79±0.00	17.37±0.03	4.08±0.03	17.09±0.00	3.42±0.00
IP	**23.79±0.04**	**7.16±0.07**	**20.38±0.00**	5.66±0.00	37.99±0.53	23.16±0.19	**37.35±1.50**	**18.47±1.13**	**23.64±0.06**	**8.37±0.03**

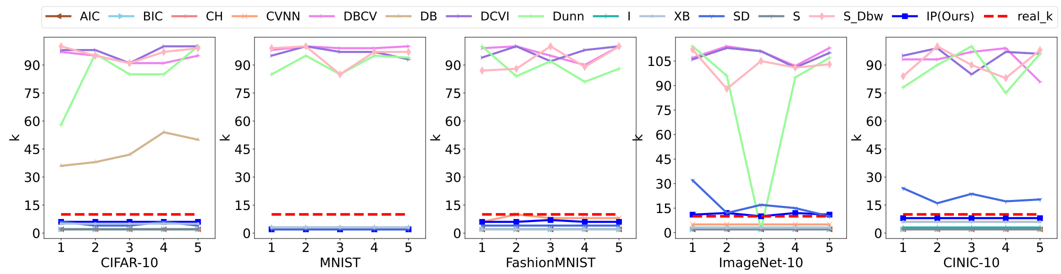


Figure 7: The optimal k value found by each index on BEiT representations for image datasets.

Table 18: BERT based K-Means clustering results in terms of NMI on five text datasets.

	SearchSnippets	**Biomedical**	**StackOverflow**	**WebofScience**	**Yahoo!Answers**
SD	37.51±0.89	26.6±0.02	2.51±0.01	36.26±0.07	6.9±0.02
Dunn	37.43±0.69	28.84±0.16	5.93±12.07	35.78±0.17	19.03±0.82
I	16.68±0.1	8.21±0.02	0.53±0.0	28.63±0.01	1.85±0.01
XB	36.82±0.24	26.6±0.02	2.51±0.01	28.63±0.01	1.85±0.01
S	37.5±0.35	8.21±0.02	2.51±0.01	28.63±0.01	1.85±0.01
CH	16.68±0.1	8.21±0.02	0.53±0.0	28.63±0.01	1.85±0.01
DB	37.05±0.05	26.6±0.02	6.99±0.01	36.26±0.07	4.79±0.02
S_Dbw	37.47±0.47	**28.95±0.1**	28.42±0.46	35.96±0.25	19.87±0.26
CVNN	16.68±0.1	8.21±0.02	0.53±0.0	28.63±0.01	1.85±0.01
DCVI	37.41±0.56	29.01±0.17	28.34±0.59	35.67±0.28	**20.08±0.37**
DBCV	37.30±0.37	28.95±0.14	**28.44±0.37**	35.72±0.20	19.98±0.20
AIC	16.68±0.10	8.21±0.02	0.53±0.00	28.63±0.01	1.85±0.01
BIC	16.68±0.10	8.21±0.02	0.53±0.00	28.63±0.01	1.85±0.01
IP	**42.31±3.01**	27.09±0.23	10.56±0.12	**40.83±0.21**	6.90±0.02

Table 19: SBERT based K-Means clustering results in terms of NMI on five text datasets.

	SearchSnippets	**Biomedical**	**StackOverflow**	**WebofScience**	**Yahoo!Answers**
SD	46.56±2.27	42.18±1.56	67.69±13.39	49.52±2.64	34.19±0.64
Dunn	45.25±1.54	39.61±0.95	49.53±12.21	46.2±3.58	32.72±3.21
I	20.16±0.04	14.4±0.02	30.84±3.87	40.34±0.0	7.83±0.03
XB	46.34±2.46	38.98±2.85	49.97±14.62	51.89±3.15	33.06±0.21
S	43.57±0.42	**42.56±1.03**	68.85±0.79	50.31±0.63	33.2±0.39
CH	20.16±0.04	14.4±0.02	24.15±0.01	40.34±0.0	7.83±0.03
DB	43.5±0.44	38.79±0.18	65.69±12.34	51.5±0.95	33.06±0.34
S_Dbw	43.42±0.22	38.73±0.14	56.35±0.45	44.22±0.32	33.15±0.14
CVNN	20.16±0.04	14.4±0.02	41.61±6.51	40.34±0.0	17.17±0.05
DCVI	43.26±0.16	38.91±0.34	71.36±5.24	44.27±0.22	33.04±0.2
DBCV	43.59±0.55	39.14±0.34	67.50±2.55	44.32±0.39	32.99±0.26
AIC	20.16±0.04	14.40±0.02	24.15±0.01	40.34±0.00	7.83±0.03
BIC	20.16±0.04	14.40±0.02	24.15±0.01	40.34±0.00	7.83±0.03
IP	**53.45±1.26**	42.16±1.04	**72.94±0.62**	**54.34±0.82**	**34.80±1.88**

Table 20: SimCSE based K-Means clustering results in terms of NMI on five text datasets.

	SearchSnippets	**Biomedical**	**StackOverflow**	**WebofScience**	**Yahoo!Answers**
SD	37.06±0.53	20.81±6.61	58.68±0.86	**44.6±0.26**	**23.82±0.3**
Dunn	37.78±0.9	23.05±1.48	52.1±1.7	41.89±2.24	22.03±0.79
I	10.4±0.01	10.21±0.01	9.67±0.06	35.76±0.0	7.92±0.04
XB	36.84±0.36	17.85±0.03	58.69±0.9	43.53±0.02	23.13±0.31
S	36.69±0.44	10.21±0.01	59.28±0.35	35.76±0.0	7.92±0.04
CH	10.4±0.01	10.21±0.01	9.67±0.06	35.76±0.0	7.92±0.04
DB	37.13±0.5	32.08±0.18	59.12±0.9	43.53±0.02	23.3±0.14
S_Dbw	37.06±0.19	32.02±0.28	51.4±0.16	39.39±0.12	23.33±0.24
CVNN	10.4±0.01	17.85±0.03	9.67±0.06	35.76±0.0	7.92±0.04
DCVI	37.02±0.29	**32.1±0.21**	51.55±0.17	39.55±0.18	23.27±0.39
DBCV	37.12±0.67	31.92±0.25	51.25±0.50	39.55±0.19	23.12±0.24
AIC	10.40±0.01	10.21±0.01	9.67±0.06	35.76±0.00	7.92±0.04
BIC	10.40±0.01	10.21±0.01	9.67±0.06	35.76±0.00	7.92±0.04
IP	**39.79±0.57**	31.55±0.08	**59.22±0.79**	44.50±1.09	20.93±0.11

Table 21: ViT based K-Means clustering results in terms of NMI on five image datasets.

	CIFAR-10	**MNIST**	**FashionMNIST**	**ImageNet-10**	**CINIC-10**
SD	**71.26±0.6**	20.38±0.42	33.31±0.02	87.49±7.95	45.4±0.57
Dunn	34.92±6.81	18.6±0.58	33.15±1.3	67.19±30.54	42.84±3.18
I	29.94±0.0	3.53±0.0	25.14±0.0	35.96±9.22	26.43±0.01
XB	34.92±6.81	18.34±0.04	25.14±0.0	**93.75±2.19**	26.43±0.01
S	67.94±1.29	3.53±0.0	29.67±0.0	90.82±0.92	55.92±0.55
CH	29.94±0.0	3.53±0.0	25.14±0.0	**93.75±2.19**	26.43±0.01
DB	34.92±6.81	19.86±0.37	25.14±0.0	70.62±35.5	26.43±0.01
S_Dbw	53.7±2.06	23.77±0.33	33.31±0.13	78.58±10.88	44.83±0.45
CVNN	34.92±6.81	3.53±0.0	25.14±0.0	68.42±8.85	26.43±0.01
DCVI	50.72±4.67	24.09±0.28	33.29±0.11	87.89±18.48	44.79±0.26
DBCV	52.80±0.56	**24.18±0.24**	33.20±0.19	93.22±4.06	44.76±0.12
AIC	29.94±0.00	3.53±0.00	25.14±0.00	31.51±4.42	26.43±0.01
BIC	29.94±0.00	3.53±0.00	25.14±0.00	31.51±4.42	26.43±0.01
IP	70.24±0.07	20.74±0.42	**34.88±0.71**	82.93±8.36	**56.89±0.73**

Table 22: SWin based K-Means clustering results in terms of NMI on five image datasets.

	CIFAR-10	**MNIST**	**FashionMNIST**	**ImageNet-10**	**CINIC-10**
SD	87.75±1.38	14.19±0.03	34.89±0.0	**99.47±0.0**	**62.9±1.44**
Dunn	86.77±2.03	17.9±0.11	33.58±0.0	93.33±6.56	53.86±3.46
I	40.25±0.0	9.45±0.01	33.58±0.0	23.7±0.31	30.65±0.03
XB	88.77±0.05	10.41±2.14	**35.76±0.08**	**99.47±0.0**	58.06±8.9
S	89.9±0.59	9.45±0.01	29.48±0.01	94.61±0.47	63.88±0.51
CH	40.25±0.0	9.45±0.01	33.58±0.0	**99.47±0.0**	30.65±0.03
DB	88.77±0.05	14.19±0.03	34.89±0.0	**99.47±0.0**	62.41±0.08
S_Dbw	63.45±0.26	**20.07±0.18**	33.42±0.29	**99.47±0.0**	50.18±1.46
CVNN	88.77±0.05	9.45±0.01	33.58±0.0	**99.47±0.0**	39.91±5.19
DCVI	63.44±0.55	19.76±0.11	33.63±0.2	95.77±5.36	49.21±0.4
DBCV	63.64±0.52	19.85±0.26	33.63±0.24	**99.47±0.00**	49.78±0.60
AIC	40.25±0.00	9.45±0.01	33.58±0.00	23.70±0.31	30.65±0.03
BIC	40.25±0.00	9.45±0.01	33.58±0.00	23.70±0.31	30.65±0.03
IP	**90.14±0.02**	13.69±0.02	34.10±0.33	**99.47±0.00**	61.97±0.97

Table 23: BEiT based K-Means clustering results in terms of NMI on five image datasets.

	CIFAR-10	**MNIST**	**FashionMNIST**	**ImageNet-10**	**CINIC-10**
SD	11.58±0.76	13.48±0.0	36.27±0.0	**28.71±0.33**	12.86±0.44
Dunn	17.76±0.33	30.53±0.15	**40.68±0.32**	24.77±8.03	14.55±0.26
I	9.21±0.02	15.36±0.0	24.09±0.0	20.93±0.0	7.99±0.06
XB	11.03±0.02	15.36±0.0	24.09±0.0	20.93±0.0	7.97±0.02
S	9.21±0.02	15.36±0.0	24.09±0.0	10.4±0.01	6.93±0.0
CH	9.21±0.02	13.48±0.0	24.09±0.0	20.93±0.0	6.93±0.0
DB	16.92±0.11	15.36±0.0	24.09±0.0	20.93±0.0	7.97±0.02
S_Dbw	17.8±0.15	30.74±0.41	40.66±0.25	28.66±0.1	**14.82±0.12**
CVNN	9.21±0.02	13.48±0.0	39.98±1.29	23.58±0.03	6.93±0.0
DCVI	**17.99±0.12**	**30.84±0.37**	40.53±0.18	28.55±0.08	14.73±0.2
DBCV	17.77±0.17	**30.84±0.45**	40.57±0.24	28.47±0.18	14.75±0.17
AIC	9.21±0.02	13.48±0.00	24.09±0.00	10.40±0.01	6.93±0.00
BIC	9.21±0.02	13.48±0.00	24.09±0.00	10.40±0.01	6.93±0.00
IP	12.39±0.06	13.48±0.00	39.24±0.21	28.50±0.62	13.69±0.10

D.2 THE CUSTERING RESULTS ON GMM

In this section, we only use GMM to evaluate the clustering results. Since the random initialization nature of GMM, the experimental results are averaged over five random runs for each validation index. The evaluation results based on external indices of ACC and ARI are shown in 24, 25, 26, 27, 28 and 29, the evaluation results in terms of NMI are shown in Table 30, 31, 32, 33, 34 and 35 where the best results are highlighted in bold and the optimal k value each index selecet is provided in Fig. 8, 9, 10, 11, 12 and 13. Obviously, for almost all cases, our IP index outperforms other indices and is close to the real k value represented in red dash line.

Table 24: BERT based GMM clustering results on five text datasets.

	SearchSnippets		Biomedical		StackOverflow		WebofScience		Yahoo!Answers	
	ACC	ARI	ACC	ARI	ACC	ARI	ACC	ARI	ACC	ARI
SD	27.19±7.62	17.25±4.15	29.45±0.05	14.73±0.15	13.98±1.96	3.48±0.66	40.88±0.01	24.69±0.01	19.76±1.95	3.06±1.00
Dunn	30.72±15.00	15.60±6.96	14.31±8.35	5.56±5.12	7.18±1.94	0.59±0.98	12.54±5.19	8.99±3.76	26.13±3.34	9.76±3.00
I	34.65±0.11	14.06±0.27	9.82±0.00	2.94±0.00	6.32±0.00	0.15±0.00	29.42±0.01	18.43±0.06	12.23±0.26	0.24±0.05
XB	24.46±10.60	15.42±6.13	26.04±3.78	12.38±2.60	8.01±0.00	0.65±0.00	29.42±0.01	18.43±0.06	12.23±0.26	0.24±0.05
S	9.87±0.52	6.58±0.22	9.82±0.00	2.94±0.00	8.01±0.00	0.65±0.00	29.42±0.01	18.43±0.06	12.23±0.26	0.24±0.05
CH	34.65±0.11	14.06±0.27	9.82±0.00	2.94±0.00	6.32±0.00	0.15±0.00	29.42±0.01	18.43±0.06	12.23±0.26	0.24±0.05
DB	9.84±0.76	6.62±0.34	27.38±3.58	13.55±2.17	10.73±0.01	2.34±0.01	40.88±0.01	24.69±0.01	14.78±2.58	1.25±0.97
S_Dbw	9.44±0.51	6.35±0.29	11.66±0.38	6.78±0.26	12.54±0.46	6.71±0.26	8.23±0.20	5.93±0.10	13.52±3.18	5.66±1.31
CVNN	34.65±0.11	14.06±0.27	14.88±3.35	6.65±2.36	6.32±0.00	0.15±0.00	29.42±0.01	18.43±0.06	20.80±2.99	3.82±1.98
DCVI	9.63±0.52	6.44±0.24	11.78±0.42	6.85±0.27	12.73±0.40	6.85±0.28	8.73±0.80	6.30±0.33	12.04±0.35	4.98±0.14
DBCV	9.73±0.58	6.57±0.35	11.98±0.65	6.99±0.52	12.59±0.39	6.73±0.23	8.68±0.23	6.25±0.21	12.51±0.60	5.22±0.19
AIC	27.84±1.16	18.61±0.62	24.50±1.69	13.88±1.33	**22.34±0.48**	**9.99±0.30**	40.88±0.01	24.69±0.01	**29.52±1.72**	**11.86±0.62**
BIC	34.65±0.11	14.06±0.27	9.82±0.00	2.94±0.00	6.32±0.00	0.15±0.00	29.42±0.01	18.43±0.06	12.23±0.26	0.24±0.05
IP	**50.68±2.20**	**30.71±2.89**	**31.96±0.11**	**15.86±0.19**	18.07±2.63	5.05±1.00	**45.54±2.47**	**29.92±1.96**	20.54±2.65	3.70±1.80

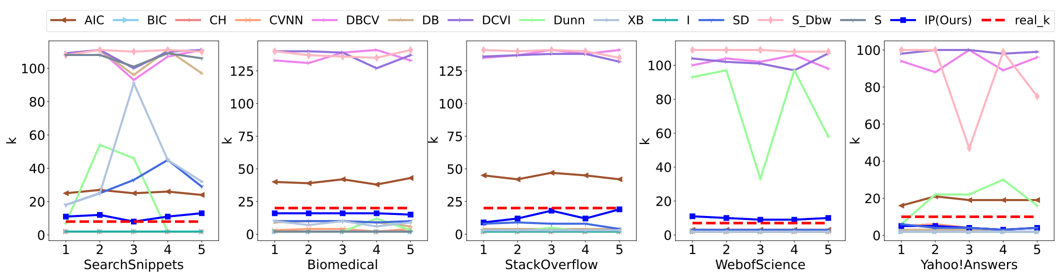


Figure 8: The optimal k value found by each index on BERT representations for text datasets.

Table 25: SBERT based GMM clustering results on five text datasets.

	SearchSnippets		Biomedical		StackOverflow		WebofScience		Yahoo!Answers	
	ACC	ARI	ACC	ARI	ACC	ARI	ACC	ARI	ACC	ARI
SD	26.25±8.66	20.43±6.14	35.15±13.42	22.66±7.93	68.51±4.92	44.19±1.52	29.92±7.39	25.65±4.88	16.96±3.60	11.36±2.20
Dunn	47.67±17.86	31.03±15.55	22.96±6.51	14.98±4.38	16.34±8.99	12.31±9.13	48.45±9.46	35.36±6.95	28.41±15.81	14.68±10.31
I	35.67±0.01	15.15±0.02	9.90±0.00	4.88±0.00	10.81±2.08	6.48±1.47	32.96±0.06	22.36±5.31	17.64±0.01	5.83±0.02
XB	24.44±13.63	18.61±9.88	38.32±5.06	23.23±3.83	56.17±19.87	35.29±13.00	38.96±15.02	32.11±11.81	16.10±4.46	10.63±2.87
S	10.73±0.61	8.18±0.27	**47.98±1.09**	**29.86±0.65**	70.54±3.54	**55.39±2.68**	50.09±18.15	40.88±13.47	11.71±0.55	7.71±0.36
CH	35.67±0.01	15.15±0.02	9.90±0.00	4.88±0.00	14.48±6.51	9.72±5.59	35.50±5.75	26.14±3.15	17.64±0.01	5.83±0.02
DB	10.87±0.60	8.28±0.18	17.30±1.17	11.52±0.54	66.22±9.54	42.33±5.46	30.57±5.56	26.44±4.10	11.44±0.66	7.53±0.22
S_Dbw	10.54±0.50	8.07±0.14	16.37±0.68	10.88±0.37	19.58±0.56	18.80±0.48	10.59±0.58	8.40±0.11	11.38±0.57	7.50±0.19
CVNN	35.67±0.01	15.15±0.02	9.90±0.00	4.88±0.00	14.46±6.47	8.43±3.52	37.25±9.65	27.80±6.87	17.64±0.01	5.83±0.02
DCVI	10.83±0.92	8.34±0.60	17.07±1.12	11.30±0.36	60.63±7.88	39.93±5.34	11.59±1.12	9.33±0.98	12.10±0.72	8.00±0.63
DBCV	10.61±0.51	8.11±0.08	17.41±0.76	11.56±0.74	59.76±17.46	43.85±12.50	11.22±0.96	8.90±0.44	11.39±0.56	7.54±0.20
AIC	23.92±1.16	18.75±0.70	31.33±1.09	21.43±1.16	35.87±2.84	35.11±2.28	23.08±1.53	20.95±1.02	25.25±0.99	16.81±0.57
BIC	35.67±0.01	15.15±0.02	9.90±0.00	4.88±0.00	9.88±0.02	5.96±0.42	32.96±0.06	22.36±5.31	17.64±0.01	5.83±0.02
IP	**57.30±4.07**	**43.66±2.36**	46.85±1.94	27.87±1.68	**72.37±1.53**	48.45±4.56	**55.19±4.24**	**42.03±3.05**	**45.03±3.29**	**23.31±2.54**

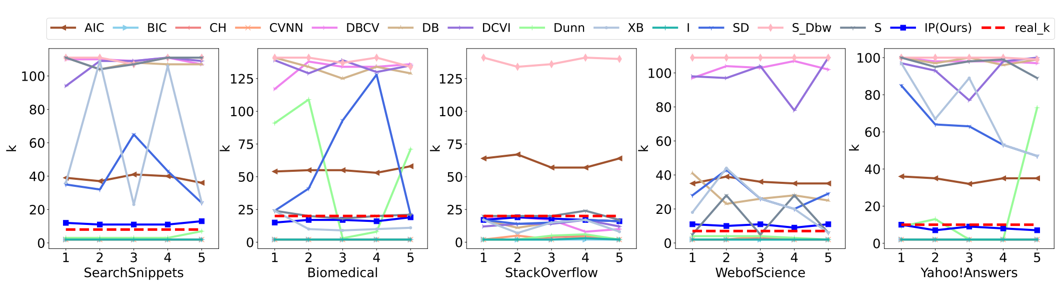


Figure 9: The optimal k value found by each index on SBERT representations for text datasets.

Table 26: SimCSE based GMM clustering results on five text datasets.

	SearchSnippets		Biomedical		StackOverflow		WebofScience		Yahoo!Answers	
	ACC	ARI	ACC	ARI	ACC	ARI	ACC	ARI	ACC	ARI
SD	15.83±4.84	10.15±3.17	32.91±4.88	18.41±3.51	61.59±6.52	43.41±5.43	55.10±9.34	36.36±4.02	16.35±2.34	8.28±1.13
Dunn	33.02±17.80	14.36±11.62	18.74±8.19	8.88±5.44	26.20±21.59	15.46±16.05	32.04±0.01	22.47±0.00	21.70±7.96	6.84±4.06
I	31.38±0.07	9.01±0.08	9.79±0.01	2.99±0.05	7.96±0.39	1.15±0.30	32.04±0.01	22.47±0.00	18.14±0.02	5.02±0.03
XB	16.25±8.42	10.77±5.87	30.73±4.27	16.05±2.12	63.69±3.62	45.64±0.57	39.03±6.38	26.00±3.23	16.02±2.16	8.14±1.12
S	10.49±1.13	6.76±0.43	9.79±0.01	2.99±0.05	62.76±3.02	**46.48±1.35**	32.04±0.01	22.47±0.00	18.14±0.02	5.02±0.03
CH	31.38±0.07	9.01±0.08	9.79±0.01	2.99±0.05	12.76±0.07	6.87±0.16	32.04±0.01	22.47±0.00	18.14±0.02	5.02±0.03
DB	10.03±0.58	6.53±0.18	14.11±0.94	8.77±0.60	62.63±3.41	45.37±1.95	47.91±6.26	32.42±5.57	10.03±0.40	4.92±0.18
S_Dbw	10.06±0.60	6.50±0.14	13.53±0.61	8.39±0.31	19.82±0.58	17.14±0.33	8.70±0.33	6.85±0.06	9.98±0.43	4.85±0.19
CVNN	31.38±0.07	9.01±0.08	13.63±2.10	5.86±0.69	12.76±0.07	6.87±0.16	32.04±0.01	22.47±0.00	18.14±0.02	5.02±0.03
DCVI	10.82±0.51	6.94±0.21	13.51±0.90	8.39±0.47	19.81±0.95	17.15±0.51	9.06±0.59	7.16±0.33	9.89±0.47	4.88±0.16
DBCV	10.39±0.41	6.61±0.18	13.94±0.79	8.64±0.45	19.65±0.67	17.01±0.29	9.50±0.93	7.42±0.55	11.92±1.91	5.91±0.98
AIC	24.42±0.68	16.24±0.67	26.95±0.12	16.24±0.18	41.91±1.33	34.70±1.16	**55.65±3.13**	**38.40±0.60**	22.73±1.23	11.32±0.63
BIC	31.38±0.07	9.01±0.08	9.79±0.01	2.99±0.05	7.96±0.39	1.15±0.30	32.04±0.01	22.47±0.00	18.14±0.02	5.02±0.03
IP	**48.29±3.91**	**28.32±2.62**	**37.99±0.19**	**20.13±0.27**	**67.49±2.51**	46.21±1.64	46.41±3.61	31.83±1.66	**35.25±2.82**	**13.63±2.11**

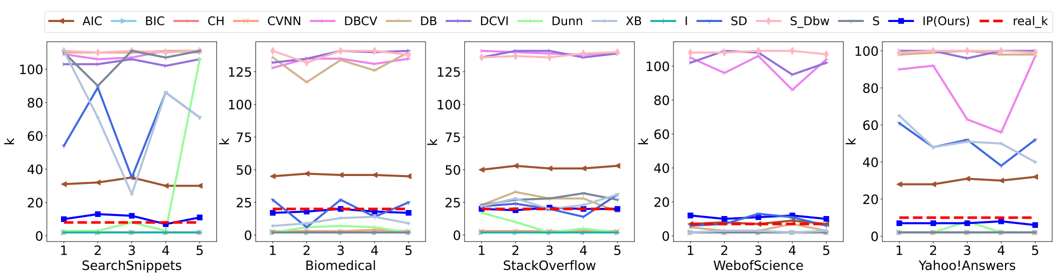


Figure 10: The optimal k value found by each index on SimCSE representations for text datasets.

Table 27: ViT based GMM clustering results on five image datasets.

	CIFAR-10		MNIST		FashionMNIST		ImageNet-10		CINIC-10	
	ACC	ARI	ACC	ARI	ACC	ARI	ACC	ARI	ACC	ARI
SD	69.11±7.08	53.05±7.04	29.68±1.74	10.71±1.32	32.48±1.42	20.68±1.42	82.27±11.64	74.07±17.78	29.36±12.92	20.61±7.21
Dunn	47.27±22.57	37.26±20.70	27.63±0.78	9.12±0.86	36.40±2.84	20.57±0.59	29.66±14.00	20.69±18.36	31.04±4.50	22.81±4.38
I	21.56±4.01	13.31±3.64	16.02±0.00	1.99±0.00	19.87±0.00	11.66±0.00	23.54±5.34	7.83±5.13	19.50±0.26	11.93±2.41
XB	47.29±18.92	34.67±18.36	28.50±0.94	10.13±0.95	19.87±0.00	11.66±0.00	88.33±10.28	83.08±14.68	33.74±8.85	21.73±8.83
S	67.83±3.34	55.12±2.36	16.02±0.00	1.99±0.00	30.11±0.05	18.94±0.02	**93.07±4.07**	**91.18±4.83**	56.74±1.70	**41.36±0.87**
CH	19.78±0.12	12.16±2.03	16.02±0.00	1.99±0.00	19.87±0.00	11.66±0.00	91.79±8.84	87.97±12.49	19.50±0.26	11.93±2.41
DB	56.61±21.38	44.40±21.24	30.00±1.93	11.01±1.52	19.87±0.00	11.66±0.00	55.09±40.42	43.88±48.49	20.44±8.11	12.69±3.87
S_Dbw	17.48±2.12	16.01±2.14	7.37±0.25	4.23±0.15	8.80±0.28	6.19±0.08	68.18±25.51	61.39±24.20	16.34±2.99	12.91±2.31
CVNN	31.12±16.21	19.98±12.77	16.02±0.00	1.99±0.00	19.87±0.00	11.66±0.00	54.54±17.79	37.92±24.98	21.15±3.77	13.71±5.17
DCVI	20.83±5.34	19.66±5.79	7.38±0.29	4.23±0.15	8.87±0.32	6.29±0.18	51.29±25.36	45.91±28.08	15.33±1.13	12.02±0.88
DBCV	16.33±0.37	14.75±0.39	7.49±0.28	4.34±0.12	8.97±0.38	6.35±0.15	81.05±19.39	79.76±19.90	15.65±0.94	12.07±0.80
AIC	28.21±1.97	26.67±2.46	13.55±0.94	7.57±0.50	16.94±0.72	11.68±0.51	33.01±3.63	36.74±4.01	27.31±1.26	21.82±0.62
BIC	19.78±0.12	12.16±2.03	16.02±0.00	1.99±0.00	19.87±0.00	11.66±0.00	19.74±0.18	6.44±2.80	19.50±0.26	11.93±2.41
IP	**73.60±3.53**	**56.39±3.39**	**30.78±1.05**	**12.05±0.25**	**38.90±1.80**	**21.32±0.61**	80.96±9.90	73.42±16.84	**59.47±4.88**	38.70±5.14

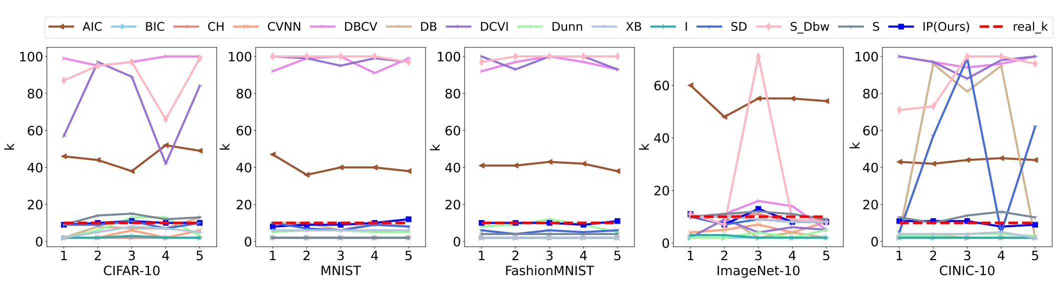


Figure 11: The optimal k value found by each index on ViT representations for image datasets.

Table 28: Swin based GMM clustering results on five image datasets.

	CIFAR-10		**MNIST**		**FashionMNIST**		**ImageNet-10**		**CINIC-10**	
	ACC	ARI	ACC	ARI	ACC	ARI	ACC	ARI	ACC	ARI
SD	74.07±21.13	68.91±25.07	**28.46±0.55**	**10.12±0.60**	30.82±0.04	20.60±0.09	71.76±26.80	62.11±34.91	64.82±6.10	50.99±4.16
Dunn	39.90±20.25	36.86±22.86	20.17±1.05	7.16±2.97	14.74±5.20	10.06±3.56	43.87±25.01	39.57±29.21	19.89±0.00	14.65±0.01
I	23.84±5.40	14.03±6.86	19.49±0.00	4.99±0.00	17.47±2.30	6.73±6.26	29.87±17.30	18.27±27.22	19.89±0.00	14.65±0.01
XB	72.37±16.35	68.17±19.63	19.49±0.00	4.99±0.00	34.26±0.02	**20.77±0.01**	83.72±11.50	79.47±16.85	55.96±16.19	43.67±13.05
S	**83.33±8.91**	**81.36±7.57**	19.49±0.00	4.99±0.00	24.65±0.02	15.29±0.02	**89.65±9.85**	**91.17±8.50**	50.51±15.02	43.66±9.85
CH	23.87±8.67	19.32±8.78	19.49±0.00	4.99±0.00	22.79±2.56	14.62±0.94	88.53±13.27	84.43±19.69	19.89±0.00	14.65±0.01
DB	82.76±8.17	78.84±5.87	28.32±0.59	9.92±0.79	30.82±0.04	20.60±0.09	69.77±29.93	64.12±36.75	68.13±1.34	**53.41±0.74**
S_Dbw	22.27±2.93	21.81±4.03	8.48±0.77	4.45±0.37	9.10±0.15	6.25±0.05	60.25±34.43	54.26±43.70	20.37±4.74	16.66±4.43
CVNN	58.54±18.89	51.50±22.25	19.49±0.00	4.99±0.00	22.79±2.56	14.62±0.94	61.81±35.55	51.92±43.60	19.89±0.00	14.65±0.01
DCVI	21.19±1.21	20.73±1.67	8.33±0.53	4.46±0.30	9.35±0.53	6.47±0.37	59.82±24.42	57.67±31.36	17.27±0.42	13.75±0.43
DBCV	22.19±2.53	21.36±2.61	8.37±0.61	4.43±0.34	9.49±0.37	6.50±0.23	88.31±10.26	87.30±11.75	17.52±1.38	13.92±1.18
AIC	43.81±2.50	47.01±3.22	16.28±0.72	8.50±0.46	21.25±1.51	13.79±0.80	50.88±6.23	55.04±6.82	35.38±1.05	29.84±0.77
BIC	19.94±0.13	13.46±6.12	19.49±0.00	4.99±0.00	17.47±2.30	6.73±6.26	19.93±0.10	10.43±7.96	19.89±0.00	14.65±0.01
IP	80.96±8.35	76.12±7.75	28.42±0.74	9.84±0.73	**35.63±0.63**	20.43±0.60	71.76±26.80	62.11±34.91	**68.57±2.84**	52.89±1.34

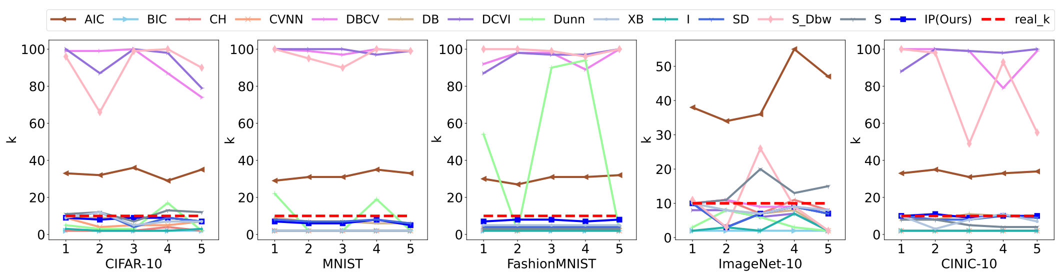


Figure 12: The optimal k value found by each index on Swin representations for image datasets.

Table 29: BEiT based GMM clustering results on five image datasets.

	CIFAR-10		**MNIST**		**FashionMNIST**		**ImageNet-10**		**CINIC-10**	
	ACC	ARI	ACC	ARI	ACC	ARI	ACC	ARI	ACC	ARI
SD	21.48±2.12	6.24±0.86	20.41±0.00	5.71±0.00	30.93±0.00	20.18±0.00	32.67±1.49	**16.38±0.58**	17.33±1.18	5.65±0.30
Dunn	14.57±3.94	4.83±0.51	13.62±6.34	8.06±2.58	37.01±13.94	22.12±7.14	22.53±5.43	7.72±4.03	8.17±3.26	2.92±1.07
I	17.44±0.01	5.17±0.01	23.39±3.32	8.00±2.10	19.78±0.00	12.05±0.01	22.81±2.87	6.98±2.41	18.37±0.07	3.09±0.10
XB	21.75±1.02	6.12±0.28	20.15±0.36	5.72±0.01	19.78±0.00	12.05±0.01	22.81±2.87	6.98±2.41	18.77±0.02	3.40±0.01
S	17.44±0.01	5.17±0.01	20.15±0.36	5.72±0.01	19.78±0.00	12.05±0.01	21.54±3.49	6.20±2.59	17.24±0.01	3.53±0.01
CH	17.44±0.01	5.17±0.01	20.41±0.00	5.71±0.00	19.78±0.00	12.05±0.01	24.10±0.07	8.07±0.07	17.24±0.01	3.53±0.01
DB	15.85±2.58	6.69±1.09	22.24±3.11	7.12±1.92	19.78±0.00	12.05±0.01	24.93±6.12	9.04±5.71	18.77±0.02	3.40±0.01
S_Dbw	7.26±0.38	3.15±0.10	9.25±0.25	6.20±0.16	12.37±0.42	9.49±0.42	11.13±0.60	6.51±0.33	6.82±0.28	2.45±0.05
CVNN	17.44±0.01	5.17±0.01	23.94±3.22	7.91±2.00	**38.93±3.04**	23.02±1.09	22.49±5.22	8.22±1.54	17.24±0.01	3.53±0.01
DCVI	7.49±0.24	3.24±0.04	9.23±0.27	6.23±0.12	12.30±0.66	9.40±0.27	11.30±0.79	6.55±0.39	6.81±0.27	2.45±0.06
DBCV	7.73±0.71	3.30±0.17	9.32±0.42	6.31±0.19	12.31±0.43	9.46±0.32	11.39±0.77	6.57±0.30	8.21±1.97	2.91±0.61
AIC	13.15±0.49	5.37±0.20	17.76±0.61	**10.41±0.43**	22.93±1.37	17.30±0.93	16.72±0.77	9.69±0.43	11.92±0.72	4.10±0.19
BIC	17.44±0.01	5.17±0.01	20.41±0.00	5.71±0.00	19.78±0.00	12.05±0.01	18.29±0.53	5.62±2.08	17.24±0.01	3.53±0.01
IP	**24.71±0.84**	**8.01±0.77**	**23.99±3.27**	8.06±2.17	37.42±1.10	**23.07±0.12**	**34.54±1.97**	14.80±1.99	**23.42±0.56**	**7.56±0.93**

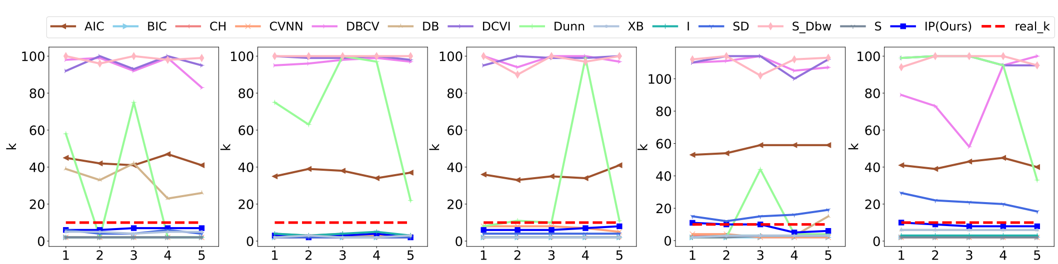


Figure 13: The optimal k value found by each index on BEiT representations for image datasets.

Table 30: BERT based GMM clustering results in terms of NMI on five text datasets.

	SearchSnippets	**Biomedical**	**StackOverflow**	**WebofScience**	**Yahoo!Answers**
SD	37.58±0.78	26.71±0.09	9.22±1.41	36.58±0.02	7.46±1.89
Dunn	29.92±9.06	13.73±7.16	1.66±2.67	36.36±0.80	17.02±4.09
I	19.73±0.38	9.93±0.03	0.46±0.00	29.34±0.09	1.01±0.08
XB	37.21±0.62	23.74±3.16	1.95±0.00	29.34±0.09	1.01±0.08
S	37.28±0.53	9.93±0.03	1.95±0.00	29.34±0.09	1.01±0.08
CH	19.73±0.38	9.93±0.03	0.46±0.00	29.34±0.09	1.01±0.08
DB	37.18±0.56	25.25±2.76	6.83±0.02	36.58±0.02	3.31±2.19
S_Dbw	37.14±0.59	**29.06±0.16**	**28.23±0.34**	35.57±0.20	**20.10±0.24**
CVNN	19.73±0.38	16.23±3.96	0.46±0.00	29.34±0.09	8.02±2.56
DCVI	37.19±0.54	29.04±0.16	28.19±0.50	35.71±0.29	20.00±0.19
DBCV	37.25±0.49	28.99±0.22	28.21±0.43	35.65±0.27	19.98±0.16
AIC	37.90±1.35	28.20±0.42	22.61±0.90	36.58±0.02	19.31±0.93
BIC	19.73±0.38	9.93±0.03	0.46±0.00	29.34±0.09	1.01±0.08
IP	**39.72±3.34**	27.29±0.22	10.24±0.75	**39.99±0.86**	7.90±2.39

Table 31: SBERT based GMM clustering results in terms of NMI on five text datasets.

	SearchSnippets	**Biomedical**	**StackOverflow**	**WebofScience**	**Yahoo!Answers**
SD	47.39±1.81	40.99±1.46	**73.45±1.32**	51.19±1.30	33.89±0.88
Dunn	38.06±12.30	35.70±7.21	35.38±14.75	49.13±4.77	24.87±13.05
I	22.76±0.08	15.77±0.01	26.03±3.54	38.61±5.12	10.69±0.04
XB	46.07±2.45	37.98±2.70	67.64±8.79	52.77±3.41	33.68±1.00
S	43.33±0.50	**42.35±1.02**	70.06±1.78	**55.55±3.59**	32.87±0.24
CH	22.76±0.08	15.77±0.01	33.03±11.52	42.54±3.66	10.69±0.04
DB	43.39±0.44	39.06±0.23	72.45±3.47	51.63±1.18	32.94±0.32
S_Dbw	43.24±0.41	38.94±0.27	56.16±0.14	44.10±0.16	32.99±0.28
CVNN	22.76±0.08	15.77±0.01	31.81±10.45	43.09±4.91	10.69±0.04
DCVI	43.36±0.51	39.02±0.30	70.78±2.69	44.66±0.47	32.99±0.35
DBCV	43.32±0.35	39.07±0.44	67.94±6.02	44.43±0.24	32.87±0.20
AIC	47.28±0.69	40.99±0.49	61.24±0.70	49.89±0.57	**34.92±0.26**
BIC	22.76±0.08	15.77±0.01	24.34±0.41	38.61±5.12	10.69±0.04
IP	**53.67±0.74**	41.49±1.37	73.36±1.38	53.75±1.52	31.05±2.63

Table 32: SimCSE based GMM clustering results in terms of NMI on five text datasets.

	SearchSnippets	**Biomedical**	**StackOverflow**	**WebofScience**	**Yahoo!Answers**
SD	36.56±0.74	30.05±3.48	56.74±3.70	45.44±1.32	23.67±0.08
Dunn	24.91±11.65	18.64±8.52	25.18±22.26	36.68±0.03	11.93±5.44
I	14.31±0.07	9.59±0.23	3.42±0.87	36.68±0.03	9.49±0.05
XB	37.34±1.95	28.08±1.78	58.44±0.42	40.85±3.82	23.62±0.27
S	36.53±0.43	9.59±0.23	58.79±0.62	36.68±0.03	9.49±0.05
CH	14.31±0.07	9.59±0.23	17.38±0.35	36.68±0.03	9.49±0.05
DB	36.64±0.49	31.99±0.16	58.22±1.31	44.47±1.17	23.28±0.17
S_Dbw	36.52±0.41	31.98±0.15	51.26±0.25	39.13±0.16	23.17±0.28
CVNN	14.31±0.07	15.47±1.66	17.38±0.35	36.68±0.03	9.49±0.05
DCVI	36.58±0.27	31.95±0.18	51.31±0.17	39.33±0.26	23.23±0.17
DBCV	36.49±0.45	32.01±0.17	51.22±0.20	39.43±0.33	23.34±0.30
AIC	37.81±0.82	**32.03±0.22**	55.67±0.49	**45.72±1.02**	**23.95±0.36**
BIC	14.31±0.07	9.59±0.23	3.42±0.87	36.68±0.03	9.49±0.05
IP	**37.53±2.35**	31.62±0.16	**59.04±1.18**	44.09±0.67	19.74±2.35

D.3 THE CUSTERING RESULTS ON AHC

In this section, we only use AHC to evaluate the clustering results. The evaluation results based on external indices of ACC, ARI and NMI are shown in Table 36, 37, 38, 39, 40 and 41, where the best results are highlighted in bold. Moreover, the optimal k each index select, i.e. opt_k, and the true cluster number in each dataset, i.e. $dataset\text{-}k$, are provided in table. Obviously, for almost all cases, our IP index outperforms other indices and is close to the real k value.

Table 33: ViT based GMM clustering results in terms of NMI on five image datasets.

	CIFAR-10	**MNIST**	**FashionMNIST**	**ImageNet-10**	**CINIC-10**
SD	68.10±2.73	19.27±2.05	31.80±2.09	88.97±5.79	48.07±2.57
Dunn	54.67±16.57	16.84±1.01	33.67±1.32	45.84±22.36	44.33±3.45
I	34.39±5.86	3.69±0.00	26.08±0.02	31.42±11.87	29.52±1.71
XB	54.94±14.94	17.74±1.14	26.08±0.02	91.94±4.53	43.49±9.78
S	68.52±1.01	3.69±0.00	29.65±0.06	**93.39±2.09**	55.65±0.73
CH	31.93±1.09	3.69±0.00	26.08±0.02	93.29±3.84	29.52±1.71
DB	60.54±17.25	19.38±2.10	26.08±0.02	58.61±36.52	41.16±8.23
S_Dbw	53.23±1.10	**23.85±0.17**	33.36±0.22	83.84±9.27	45.17±1.16
CVNN	42.37±15.17	3.69±0.00	26.08±0.02	68.65±15.96	31.99±6.20
DCVI	54.70±2.32	23.77±0.11	33.44±0.15	69.08±22.06	44.67±0.44
DBCV	52.47±0.29	23.82±0.24	33.41±0.31	87.92±8.31	44.53±0.24
AIC	57.69±0.95	22.90±0.32	33.89±0.59	72.87±1.03	48.74±0.35
BIC	31.93±1.09	3.69±0.00	26.08±0.02	27.37±6.07	29.52±1.71
IP	**69.58±1.44**	20.59±0.70	**34.09±0.54**	88.17±4.96	**54.43±2.33**

Table 34: SWin based GMM clustering results in terms of NMI on five image datasets.

	CIFAR-10	**MNIST**	**FashionMNIST**	**ImageNet-10**	**CINIC-10**
SD	81.29±11.95	15.46±0.89	34.84±0.11	81.69±22.68	61.36±2.92
Dunn	60.62±15.95	13.06±4.80	33.67±0.37	63.54±26.27	31.41±0.02
I	37.72±11.88	9.56±0.00	15.88±16.19	40.07±26.62	31.41±0.02
XB	81.82±7.77	9.56±0.00	**35.70±0.16**	92.86±5.52	57.20±8.49
S	**86.39±2.44**	9.56±0.00	29.63±0.02	95.33±3.89	57.42±6.34
CH	43.57±11.96	9.56±0.00	31.23±2.17	94.35±6.23	31.41±0.02
DB	86.36±2.29	15.16±1.04	34.84±0.11	78.83±30.79	**63.05±1.19**
S_Dbw	64.05±1.55	19.84±0.34	33.34±0.22	69.50±33.65	50.15±1.84
CVNN	74.76±10.36	9.56±0.00	31.23±2.17	69.56±33.83	31.41±0.02
DCVI	63.69±0.82	19.86±0.27	33.47±0.19	75.52±28.87	48.98±0.52
DBCV	63.83±1.10	**19.87±0.28**	33.54±0.22	**95.25±3.81**	48.98±0.70
AIC	73.79±1.01	18.35±0.23	34.68±0.31	79.93±2.47	54.39±0.22
BIC	35.08±9.18	9.56±0.00	15.88±16.19	32.40±11.47	31.41±0.02
IP	85.57±2.73	15.05±0.82	34.10±0.34	81.69±22.68	62.36±0.71

Table 35: BEiT based GMM clustering results in terms of NMI on five image datasets.

	CIFAR-10	**MNIST**	**FashionMNIST**	**ImageNet-10**	**CINIC-10**
SD	11.60±0.72	13.57±0.00	36.40±0.01	28.39±0.65	13.03±0.29
Dunn	12.74±4.29	29.10±2.33	41.33±1.25	20.15±7.62	14.58±0.57
I	9.61±0.02	18.87±3.23	24.75±0.02	20.27±3.51	8.32±0.10
XB	11.07±0.01	14.29±0.98	24.75±0.02	20.27±3.51	7.97±0.03
S	9.61±0.02	14.29±0.98	24.75±0.02	17.97±5.46	7.44±0.02
CH	9.61±0.02	13.57±0.00	24.75±0.02	21.83±0.02	7.44±0.02
DB	17.11±0.27	17.48±3.79	24.75±0.02	21.61±5.15	7.97±0.03
S_Dbw	17.72±0.12	**30.82±0.28**	40.57±0.13	28.51±0.15	14.81±0.16
CVNN	9.61±0.02	17.54±3.63	39.58±1.34	17.84±5.44	7.44±0.02
DCVI	17.68±0.08	30.77±0.24	40.49±0.19	28.53±0.23	**14.82±0.13**
DBCV	**17.73±0.05**	30.72±0.23	40.55±0.14	**28.53±0.15**	14.60±0.39
AIC	16.88±0.28	27.92±0.64	**41.39±0.36**	28.51±0.29	13.84±0.12
BIC	9.61±0.02	13.57±0.00	24.75±0.02	13.19±1.59	7.44±0.02
IP	13.62±1.10	17.37±3.49	39.52±0.45	25.39±1.40	13.03±0.86

Table 36: BERT based AHC clustering results on five text datasets.

	SearchSnippets - 8				**Biomedical - 20**				**StackOverflow - 20**				**WebofScience - 7**				**Yahoo!Answers - 10**			
	ACC	ARI	NMI	opt_k	ACC	ARI	NMI	opt_k	ACC	ARI	NMI	opt_k	ACC	ARI	NMI	opt_k	ACC	ARI	NMI	opt_k
SD	13.08	8.07	38.17	98	25.61	11.87	22.53	7	10.86	1.56	7.69	4	42.58	25.51	**40.41**	3	18.5	2.77	6.42	3
Dunn	19.21	12.5	39.08	57	12.26	6.93	**27.95**	141	13.8	**7.16**	**28.78**	141	20.48	13.75	36.2	50	**29.73**	**9.68**	16.78	15
I	37.27	12.84	24.83	3	9.49	2.77	8.89	2	6.57	0.18	0.64	2	31.52	21.39	34.22	2	18.3	2.75	6.38	4
XB	13.08	8.07	38.17	98	28.52	13.24	25.08	10	10.86	1.56	7.69	4	31.52	21.39	34.22	2	18.5	2.77	6.42	3
S	30.79	8.77	18.2	2	9.49	2.77	8.89	2	8.57	0.7	3.12	3	31.52	21.39	34.22	2	15.76	2.06	3.84	2
CH	30.79	8.77	18.2	2	9.49	2.77	8.89	2	6.57	0.18	0.64	2	31.52	21.39	34.22	2	15.76	2.06	3.84	2
DB	10.93	7.06	37.96	110	28.52	13.24	25.08	10	10.86	1.56	7.69	4	42.58	25.51	**40.41**	3	19.25	2.56	8.9	6
S_Dbw	10.75	7.02	37.92	111	12.9	7.22	27.84	126	13.8	**7.16**	**28.78**	141	9.69	6.5	35.2	109	20.66	8.12	19.12	45
CVNN	49.09	25.14	37.89	5	12.88	5.13	13.94	3	6.57	0.18	0.64	2	31.52	21.39	34.22	2	15.76	2.06	3.84	2
DCVI	10.75	7.02	37.92	111	12.26	6.93	**27.95**	141	13.8	**7.16**	**28.78**	141	9.69	6.5	35.2	109	14.18	5.64	**19.63**	100
DBCV	10.75	7.02	37.92	111	12.26	6.93	**27.95**	141	13.8	**7.16**	**28.78**	141	9.69	6.5	35.2	109	14.18	5.64	**19.63**	100
AIC	30.79	8.77	18.2	2	9.49	2.77	8.89	2	6.57	0.18	0.64	2	31.52	21.39	34.22	2	15.76	2.06	3.84	2
BIC	30.79	8.77	18.2	2	9.49	2.77	8.89	2	6.57	0.18	0.64	2	31.52	21.39	34.22	2	15.76	2.06	3.84	2
IP	**64.6**	**33.29**	**43.96**	7	**28.8**	**13.93**	25.38	15	**14.49**	3.95	10.54	7	**53.72**	**31.44**	39.73	7	18.5	2.77	6.42	3

Table 37: SBERT based AHC clustering results on five text datasets.

	SearchSnippets - 8				**Biomedical - 20**				**StackOverflow - 20**				**WebofScience - 7**				**Yahoo!Answers - 10**			
	ACC	ARI	NMI	opt_k	ACC	ARI	NMI	opt_k	ACC	ARI	NMI	opt_k	ACC	ARI	NMI	opt_k	ACC	ARI	NMI	opt_k
SD	35.57	25.65	49.98	32	**32.46**	**19.31**	33.88	42	33.66	10.43	50.1	8	30.12	24.23	47.3	29	29.65	16.86	32.17	28
Dunn	36.09	15.29	28.18	2	9.79	3.03	11.07	2	9.55	3.05	19.76	2	32.66	23.76	38.21	2	18.08	5.49	11.96	2
I	36.09	15.29	28.18	2	13.73	5.06	16.87	3	13.4	3.29	25.39	3	32.66	23.76	38.21	2	18.08	5.49	11.96	2
XB	14.95	9.61	44.41	111	17	6.06	19.86	4	33.66	10.43	50.1	8	56.46	**44.25**	**52.39**	13	17.66	10.13	31.23	74
S	**71.99**	**53.75**	**58.41**	10	9.79	3.03	11.07	2	**62.84**	**46.21**	60.04	26	**62**	42.84	50.05	6	**48.68**	**24.18**	**33.95**	11
CH	36.09	15.29	28.18	2	9.79	3.03	11.07	2	9.55	3.05	19.76	2	32.66	23.76	38.21	2	18.08	5.49	11.96	2
DB	14.95	9.61	44.41	111	17.76	10.74	**34.44**	141	26.01	7.29	42.83	6	56.46	**44.25**	**52.39**	13	13.53	7.44	30.59	99
S_Dbw	14.95	9.61	44.41	111	17.76	10.74	**34.44**	138	51.75	17.02	59.97	13	11.96	8.44	41.59	109	13.53	7.44	30.57	100
CVNN	36.09	15.29	28.18	2	9.79	3.03	11.07	2	21.72	5.74	37.4	5	32.66	23.76	38.21	2	18.08	5.49	11.96	2
DCVI	14.95	9.67	44.42	110	17.76	10.74	**34.44**	141	26.93	24.17	53.42	141	11.96	8.44	41.59	109	13.53	7.44	30.57	100
DBCV	14.95	9.61	44.41	111	17.76	10.74	**34.44**	141	26.93	24.17	53.42	141	11.96	8.44	41.59	109	13.53	7.44	30.57	100
AIC	36.09	15.29	28.18	2	9.79	3.03	11.07	2	9.55	3.05	19.76	2	32.66	23.76	38.21	2	18.08	5.49	11.96	2
BIC	36.09	15.29	28.18	2	9.79	3.03	11.07	2	9.55	3.05	19.76	2	32.66	23.76	38.21	2	18.08	5.49	11.96	2
IP	61.01	33.17	50.8	5	24.4	8.8	25.89	8	54.99	19.35	**61.4**	14	**62**	42.84	50.05	6	44.1	20.28	30.29	7

Table 38: SimCSE based AHC clustering results on five text datasets.

	SearchSnippets - 8				**Biomedical - 20**				**StackOverflow - 20**				**WebofScience - 7**				**Yahoo!Answers - 10**			
	ACC	ARI	NMI	opt_k	ACC	ARI	NMI	opt_k	ACC	ARI	NMI	opt_k	ACC	ARI	NMI	opt_k	ACC	ARI	NMI	opt_k
SD	12.1	6.77	37	111	16.26	4.07	16.28	4	35.94	26.34	50.39	76	50.25	33.08	44.24	10	**25.96**	4.37	17.99	4
Dunn	**48.7**	**26.19**	**40.03**	12	17.69	10.48	29.97	112	8.97	2.31	15.85	2	28.92	19.04	30.28	2	16.36	1.74	8.25	2
I	43.23	14.67	28.97	4	12.68	3.59	12.91	3	12.86	4.38	25.03	3	28.92	19.04	30.28	2	21.54	3.21	13.79	3
XB	12.1	6.77	37	111	16.26	4.07	16.28	4	8.97	2.31	15.85	2	40.13	24.49	38.35	3	16.36	1.74	8.25	2
S	12.1	6.77	37	111	9.7	2.97	9.91	2	54.28	**36.8**	**53.09**	30	28.92	19.04	30.28	2	16.36	1.74	8.25	2
CH	32.21	8.82	15.97	2	9.7	2.97	9.91	2	8.97	2.31	15.85	2	28.92	19.04	30.28	2	16.36	1.74	8.25	2
DB	12.1	6.77	37	111	16.26	4.07	16.28	4	**57.99**	23.79	54.46	19	50.25	33.08	44.24	10	9.55	4.41	21.78	100
S_Dbw	12.1	6.77	37	111	14.44	8.69	**30.09**	141	24.81	19.33	48.93	141	10.79	7.28	38.48	109	10.3	**4.65**	**21.84**	95
CVNN	32.21	8.82	15.97	2	9.7	2.97	9.91	2	8.97	2.31	15.85	2	28.92	19.04	30.28	2	21.54	3.21	13.79	3
DCVI	12.1	6.81	37.01	110	14.44	8.69	**30.09**	141	24.81	19.33	48.93	141	10.79	7.28	38.48	109	9.55	4.41	21.78	100
DBCV	12.1	6.81	37.01	110	14.44	8.69	**30.09**	141	24.81	19.33	48.93	141	10.79	7.28	38.48	109	9.55	4.41	21.78	100
AIC	32.21	8.82	15.97	2	9.7	2.97	9.91	2	8.97	2.31	15.85	2	28.92	19.04	30.28	2	16.36	1.74	8.25	2
BIC	32.21	8.82	15.97	2	9.7	2.97	9.91	2	8.97	2.31	15.85	2	28.92	19.04	30.28	2	16.36	1.74	8.25	2
IP	38.82	13.99	25.74	3	**28.86**	**13.73**	25.62	11	36.31	15.93	43.76	10	**52.45**	**34.44**	**44.56**	9	21.54	3.21	13.79	3

Table 39: ViT based AHC clustering results on five image datasets.

	CIFAR-10 - 10				**MNIST - 10**				**FashionMNIST - 10**				**ImageNet-10 - 10**				**CINIC-10 - 10**			
	ACC	ARI	NMI	opt_k	ACC	ARI	NMI	opt_k	ACC	ARI	NMI	opt_k	ACC	ARI	NMI	opt_k	ACC	ARI	NMI	opt_k
SD	26.44	10.03	36.34	3	**31.07**	**11.74**	19.67	9	30.38	16.69	**35.1**	4	87.52	82.8	91.19	9	**61.4**	35.04	**52.82**	10
Dunn	50.37	38.95	56.87	6	29.57	10.74	17.82	6	30.38	16.69	**35.1**	4	19.85	9.56	33.86	2	18.43	6.32	23.59	2
I	18.47	6.02	24.6	2	23.52	5.43	9.22	4	19.79	9.91	28.48	2	29.65	16.54	49.2	3	18.43	6.32	23.59	2
XB	26.44	10.03	36.34	3	21.82	3.78	8.22	3	19.79	9.91	28.48	2	87.52	82.8	91.19	9	18.43	6.32	23.59	2
S	60.92	49.9	63.19	16	16.88	2.69	5.33	2	19.79	9.91	28.48	2	**91.85**	**92.36**	**92.99**	12	57.36	**35.96**	52.69	9
CH	18.47	6.02	24.6	2	16.88	2.69	5.33	2	19.79	9.91	28.48	2	87.52	82.8	91.19	9	18.43	6.32	23.59	2
DB	26.44	10.03	36.34	3	30.01	11.45	19.61	8	19.79	9.91	28.48	2	87.52	82.8	91.19	9	18.43	6.32	23.59	2
S_Dbw	22.16	18.31	52.76	81	8.81	4.76	**24.04**	99	9.96	6.26	32.55	99	87.52	82.8	91.19	9	17.01	12.04	43.28	100
CVNN	26.44	10.03	36.34	3	16.88	2.69	5.33	2	22.97	11.36	27.83	3	59.23	49.23	78.3	6	18.43	6.32	23.59	2
DCVI	50.37	38.95	56.87	6	8.81	4.71	24.02	100	9.96	6.16	32.52	100	39.58	25.55	61.52	4	17.06	12.18	43.31	99
DBCV	20.17	16.16	51.81	100	8.81	4.71	24.02	100	9.96	6.16	32.52	100	**91.85**	**92.36**	**92.99**	12	17.01	12.04	43.28	100
AIC	18.47	6.02	24.6	2	16.88	2.69	5.33	2	19.79	9.91	28.48	2	19.85	9.56	33.86	2	18.43	6.32	23.59	2
BIC	18.47	6.02	24.6	2	16.88	2.69	5.33	2	19.79	9.91	28.48	2	19.85	9.56	33.86	2	18.43	6.32	23.59	2
IP	**70.92**	**51.77**	**64.26**	10	29.57	10.74	17.82	6	**38.57**	**22.17**	34.24	8	87.52	82.8	91.19	9	46.74	28.84	48.71	7

D.4 THE CUSTERING RESULTS ON DBSCAN

In this section, we only use DBSCAN to evaluate the clustering results. The evaluation results based on external indices of ACC, ARI and NMI are shown in Table 42, 43, 44, 45, 46 and 47, where the best results are highlighted in bold and Top-3 best results are underlined. Moreover, the optimal k each index select, i.e. opt_k, and the true cluster number in each dataset, i.e. $dataset\text{-}k$, are provided in table. Our index is either on par or slightly better than competing indices.

Table 40: Swin based AHC clustering results on five image datasets.

	CIFAR-10 - 10				**MNIST - 10**				**FashionMNIST - 10**				**ImageNet-10 - 10**				**CINIC-10 - 10**			
	ACC	ARI	NMI	opt_k	ACC	ARI	NMI	opt_k	ACC	ARI	NMI	opt_k	ACC	ARI	NMI	opt_k	ACC	ARI	NMI	opt_k
SD	85.92	80.08	86.43	9	19.36	4.09	7.85	3	31.8	**21.96**	**37.75**	4	**99.66**	**99.25**	**98.94**	10	**64.77**	47.09	**58.26**	12
Dunn	19.94	18.12	42.4	2	23.54	6.38	9.96	4	19.97	13.79	34.16	2	39.93	18.49	57.59	4	19.35	14.62	33.03	2
I	19.94	18.12	42.4	2	16.93	2.01	4.46	2	19.97	13.79	34.16	2	29.93	10.73	43.02	3	19.35	14.62	33.03	2
XB	85.92	80.08	86.43	9	16.93	2.01	4.46	2	34.43	20.54	36.51	5	**99.66**	**99.25**	**98.94**	10	27.79	18.64	43.35	3
S	89	83.9	85.83	11	16.93	2.01	4.46	2	19.97	13.79	34.16	2	91.03	91.92	94.97	14	43.62	38.8	52.97	5
CH	19.94	18.12	42.4	2	16.93	2.01	4.46	2	19.97	13.79	34.16	2	**99.66**	**99.25**	**98.94**	10	19.35	14.62	33.03	2
DB	85.92	80.08	86.43	9	16.93	2.01	4.46	2	34.43	20.54	36.51	5	**99.66**	**99.25**	**98.94**	10	27.79	18.64	43.35	3
S_Dbw	46.3	46.79	72.51	34	8.66	4.35	**19.86**	99	9.81	6.36	33.21	100	**99.66**	**99.25**	**98.94**	10	20.7	15.57	48.79	86
CVNN	76.7	72.92	83.99	8	16.93	2.01	4.46	2	19.97	13.79	34.16	2	**99.66**	**99.25**	**98.94**	10	27.79	18.64	43.35	3
DCVI	21.2	19.46	62.25	100	8.66	4.34	**19.86**	100	9.81	6.36	33.21	100	39.93	18.49	57.59	4	18.74	14.08	48.21	100
DBCV	19.94	18.12	42.4	2	8.66	4.34	**19.86**	100	9.81	6.36	33.21	100	**99.66**	**99.25**	**98.94**	10	18.74	14.08	48.21	100
AIC	19.94	18.12	42.4	2	16.93	2.01	4.46	2	19.97	13.79	34.16	2	19.97	9.85	35.28	2	19.35	14.62	33.03	2
BIC	19.94	18.12	42.4	2	16.93	2.01	4.46	2	19.97	13.79	34.16	2	19.97	9.85	35.28	2	19.35	14.62	33.03	2
IP	**93.62**	**86.56**	**87.01**	10	**24.03**	**7.17**	12.3	8	**35.03**	21.09	35.54	6	**99.66**	**99.25**	**98.94**	10	61.49	**47.21**	57.78	8

Table 41: BEiT based AHC clustering results on five image datasets.

	CIFAR-10 - 10				**MNIST - 10**				**FashionMNIST - 10**				**ImageNet-10 - 10**				**CINIC-10 - 10**			
	ACC	ARI	NMI	opt_k	ACC	ARI	NMI	opt_k	ACC	ARI	NMI	opt_k	ACC	ARI	NMI	opt_k	ACC	ARI	NMI	opt_k
SD	20.67	6.21	9.42	5	20.87	6.04	12.07	2	34.74	21.87	38.85	5	16.54	3.92	11.21	2	17.19	3.82	6.93	2
Dunn	10.19	4.14	16.66	79	11.17	**6.93**	**32.16**	100	15.48	11.7	41.01	84	20.45	4.67	16.25	3	7.55	2.34	**13.6**	100
I	16.96	3.96	7.56	2	22.61	5.9	18.2	4	19.9	12.47	25.97	2	20.45	4.67	16.25	3	17.88	3.5	8.46	3
XB	17.59	3.85	7.53	3	20.87	6.04	12.07	2	19.9	12.47	25.97	2	20.45	4.67	16.25	3	17.19	3.82	6.93	2
S	16.96	3.96	7.56	2	20.87	6.04	12.07	2	19.9	12.47	25.97	2	20.45	4.67	16.25	3	17.19	3.82	6.93	2
CH	16.96	3.96	7.56	2	20.87	6.04	12.07	2	19.9	12.47	25.97	2	16.54	3.92	11.21	2	17.19	3.82	6.93	2
DB	8.44	3.37	17.15	98	22.61	5.9	18.2	4	19.9	12.47	25.97	2	20.45	4.67	16.25	3	9.22	2.79	12.96	73
S_Dbw	8.16	3.31	**17.2**	100	11.17	6.93	**32.16**	100	13.48	10.15	40.73	97	12.78	6.87	**28.54**	113	7.55	2.34	13.58	99
CVNN	16.96	3.96	7.56	2	20.87	6.04	12.07	2	**42.72**	**23.03**	**42.06**	10	16.54	3.92	11.21	2	17.19	3.82	6.93	2
DCVI	8.16	3.31	**17.2**	100	11.17	**6.93**	**32.16**	100	12.6	9.71	40.64	100	12.78	6.87	**28.54**	114	7.55	2.34	**13.6**	100
DBCV	8.16	3.31	**17.2**	100	11.17	**6.93**	**32.16**	100	12.6	9.71	40.64	100	12.78	6.87	**28.54**	114	15.06	4.36	12.15	29
AIC	16.96	3.96	7.56	2	20.87	6.04	12.07	2	19.9	12.47	25.97	2	16.54	3.92	11.21	2	17.19	3.82	6.93	2
BIC	16.96	3.96	7.56	2	20.87	6.04	12.07	2	19.9	12.47	25.97	2	16.54	3.92	11.21	2	17.19	3.82	6.93	2
IP	**25.23**	**7.05**	12.6	8	**23.85**	6.48	18.36	3	37.62	20.75	38.08	7	**32.19**	**15.9**	26.49	14	**21.17**	**4.69**	9.88	7

Table 42: BERT based DBSCAN clustering results on five text datasets.

	SearchSnippets - 8				**Biomedical - 20**				**StackOverflow - 20**				**WebofScience - 7**				**Yahoo!Answers - 10**			
	ACC	ARI	NMI	opt_k	ACC	ARI	NMI	opt_k	ACC	ARI	NMI	opt_k	ACC	ARI	NMI	opt_k	ACC	ARI	NMI	opt_k
SD	21.56	0	0.01	2	5	0	0.01	2	5.01	0	0.01	2	17.6	0	0.02	2	10.01	0	0.02	2
Dunn	21.56	0	0.01	2	5	0	0.01	2	5.01	0	0.01	2	17.61	0	0.02	2	10.04	0	0.05	2
I	21.56	0	0.01	2	5	0	0.01	2	5.01	0	0.01	2	17.6	0	0.02	2	10.01	0	0.02	2
XB	21.56	0	0.01	2	5	0	0.01	2	5.01	0	0.01	2	17.6	0	0.02	2	10.01	0	0.02	2
S	21.56	0	0.01	2	5	0	0.01	2	5.01	0	0.01	2	17.6	0	0.02	2	10.01	0	0.02	2
CH	**24.63**	0.29	**5.39**	2	9.22	**1.45**	5.27	2	**6.45**	**0.28**	**0.91**	2	**25.86**	**3.49**	**14.52**	4	11.86	0.18	0.94	2
DB	21.56	0	0.01	2	5	0	0.01	2	5.01	0	0.01	2	17.6	0	0.02	2	10.01	0	0.02	2
S_Dbw	21.64	0.01	0.18	3	5	0	0.01	2	5.01	0	0.01	2	17.6	0	0.02	2	10.05	0.01	0.07	3
CVNN	21.67	0.01	1.01	14	5.34	0	0.66	2	5.62	0.02	0.43	2	17.85	0.02	0.46	2	10.04	0	0.05	2
DCVI	21.56	0	0.01	2	5	0	0.01	2	5.01	0	0.01	2	17.6	0	0.02	2	10.01	0	0.02	2
DBCV	23.41	**0.5**	1.13	3	8.28	0.74	3.87	6	5.95	0.06	0.58	3	19.43	0.14	0.8	3	10.01	0	0.02	2
AIC	21.56	0	0.03	2	5.01	0	0.03	2	5.01	0	0.01	2	17.6	0	0.02	2	10.02	0	0.04	2
BIC	21.56	0	0.03	2	5.01	0	0.03	2	5.01	0	0.01	2	17.6	0	0.02	2	10.02	0	0.04	2
IP	21.74	-0.38	4.34	2	**9.23**	1.36	**7.2**	2	5.38	0.01	0.39	2	18.38	0.05	0.34	2	**13.65**	**0.95**	**2.86**	2

Table 43: SBERT based DBSCAN clustering results on five text datasets.

	SearchSnippets - 8				Biomedical - 20				StackOverflow - 20				WebofScience - 7				Yahoo!Answers - 10			
	ACC	ARI	NMI	opt_k	ACC	ARI	NMI	opt_k	ACC	ARI	NMI	opt_k	ACC	ARI	NMI	opt_k	ACC	ARI	NMI	opt_k
SD	21.89	0.06	1.26	13	6.24	0.03	3.8	38	5.11	0	0.38	7	18.92	-0.02	6.77	12	10.4	0	1.22	14
Dunn	21.58	0.01	0.07	2	5	0	0.02	2	5.01	0	0.01	2	17.62	0	0.02	2	10.01	0	0.02	2
I	21.55	0	0.01	2	5.1	0	0.19	2	5.01	0	0.01	2	17.62	0	0.02	2	10.01	0	0.02	2
XB	21.55	0	0.01	2	5.04	0	0.09	3	5.01	0	0.01	2	17.62	0	0.02	2	10.01	0	0.02	2
S	21.56	0	0.02	2	5	0	0.02	2	5.01	0	0.02	2	17.62	0	0.03	2	10.01	0	0.02	2
CH	**24.46**	**2.18**	**4.37**	2	**13.08**	**4.39**	**17.73**	3	**26.05**	**4.78**	**37.69**	9	**42.56**	**11.19**	**35.83**	7	**21.33**	**3.23**	**14.61**	5
DB	21.55	0	0.01	2	5.04	0	0.09	3	5.01	0	0.01	2	17.62	0	0.02	2	10.01	0	0.02	2
S_Dbw	21.6	0	0.09	2	5.46	0	1.46	20	5.28	0	1.13	17	17.62	0	0.02	2	10.04	0	0.07	3
CVNN	21.87	0.01	0.63	4	6.1	0.03	2.27	7	5.13	0	0.36	4	22.97	0.51	14.77	6	10.04	0	0.04	2
DCVI	21.55	0	0.01	2	6.24	0.03	3.8	38	5.01	0	0.01	2	17.62	0	0.02	2	10.01	0	0.02	2
DBCV	21.49	-0.03	0.74	3	12.8	0.76	13.8	6	21.17	2.26	29.04	18	18.42	0.25	0.86	3	16.78	0.34	12.52	22
AIC	21.56	0	0.03	2	5	0	0.02	2	5.01	0	0.02	2	17.62	0	0.03	2	10.03	0	0.06	2
BIC	21.56	0	0.03	2	5	0	0.02	2	5.01	0	0.02	2	17.62	0	0.03	2	10.03	0	0.06	2
IP	21.9	0.12	1.13	2	12.24	3.13	16.18	3	6.17	0.08	2.39	2	19.09	0.62	2.28	2	11.96	0.24	1.97	2

Table 44: SimCSE based DBSCAN clustering results on five text datasets.

	SearchSnippets - 8				Biomedical - 20				StackOverflow - 20				WebofScience - 7				Yahoo!Answers - 10			
	ACC	ARI	NMI	opt_k	ACC	ARI	NMI	opt_k	ACC	ARI	NMI	opt_k	ACC	ARI	NMI	opt_k	ACC	ARI	NMI	opt_k
SD	21.65	0.05	1.03	11	8.86	0.25	7.34	12	6.64	0.02	3.18	10	17.62	0	0.02	2	10.01	0	0.02	2
Dunn	21.56	0	0.01	2	5	0	0.02	2	5.02	0	0.03	2	17.62	0	0.02	2	10.01	0	0.02	2
I	21.56	0	0.01	2	5	0	0.02	2	5.04	0	0.08	2	17.62	0	0.02	2	10.01	0	0.02	2
XB	21.56	0	0.01	2	5	0	0.02	2	5.04	0	0.08	2	17.62	0	0.02	2	10.01	0	0.02	2
S	21.56	0	0.01	2	5.03	0	0.08	2	5.02	0	0.06	2	17.62	0	0.02	2	10.01	0	0.02	2
CH	**26.29**	**3.09**	5.79	2	9.18	**2.27**	7.02	2	7.54	0.37	1.77	2	**26.46**	**6.05**	**15.68**	3	**14.78**	**0.95**	**6.26**	2
DB	21.56	0	0.01	2	5.1	0	0.43	9	5.04	0	0.08	2	17.62	0	0.02	2	10.01	0	0.02	2
S_Dbw	21.6	0	0.09	2	6.1	0.03	2.85	22	5.45	0	1.12	14	17.62	0	0.02	2	10.04	0.01	0.07	3
CVNN	21.65	-0.01	0.52	4	9.15	0.34	7.9	5	8.86	0.21	**7.41**	4	18.13	0.04	1.01	2	10.03	0	0.04	2
DCVI	21.56	0	0.01	2	5	0	0.02	2	6.33	0.01	3.89	42	17.62	0	0.02	2	10.01	0	0.02	2
DBCV	21.04	0.27	**13.25**	114	8.91	1.93	6.03	3	**9.29**	**1.65**	5.86	3	18.9	0.27	0.83	3	10.09	0	0.2	4
AIC	21.56	0	0.01	2	5	0	0.02	2	5.02	0	0.03	2	17.62	0	0.02	2	10.01	0	0.02	2
BIC	21.56	0	0.01	2	5	0	0.02	2	5.02	0	0.03	2	17.62	0	0.02	2	10.01	0	0.02	2
IP	21.6	-0.29	1.88	2	**11.31**	0.99	**12.65**	3	5.4	0.01	0.31	2	25.9	5.66	15.16	3	14.72	0.92	6.19	2

Table 45: ViT based DBSCAN clustering results on five image datasets.

	CIFAR-10 - 10				MNIST - 10				FashionMNIST - 10				ImageNet-10 - 10				CINIC-10 - 10			
	ACC	ARI	NMI	opt_k	ACC	ARI	NMI	opt_k	ACC	ARI	NMI	opt_k	ACC	ARI	NMI	opt_k	ACC	ARI	NMI	opt_k
SD	10.49	0	1.39	7	11.34	0	0.02	2	10.01	0	0.02	2	34.34	5.11	36.03	9	10.01	0	0.02	2
Dunn	10.01	0	0.02	2	11.34	0	0.02	2	10.01	0	0.02	2	10.03	0	0.06	2	10.01	0	0.02	2
I	10.01	0	0.02	2	11.34	0	0.02	2	10.01	0	0.02	2	10.04	0	0.08	2	10.01	0	0.02	2
XB	10.01	0	0.02	2	11.34	0	0.02	2	10.01	0	0.02	2	10.04	0	0.08	2	10.01	0	0.02	2
S	10.01	0	0.02	2	11.34	0	0.02	2	10.02	0	0.04	2	**75.64**	**62.69**	**79.51**	10	10.01	0	0.02	2
CH	32.68	**16.62**	40.09	5	**15.55**	**0.82**	**3.99**	2	**19.68**	**3.19**	**12.92**	3	75.63	59.05	78.06	9	**38.18**	**13.33**	**42.86**	7
DB	10.01	0	0.02	2	11.34	0	0.02	2	10.01	0	0.02	2	10.04	0	0.08	2	10.01	0	0.02	2
S_Dbw	11.52	0.02	4.32	16	11.34	0	0.02	2	10.01	0	0.02	2	20.41	1.38	18.56	14	10.03	0	0.08	2
CVNN	11.09	0.02	2.75	7	12.95	0.12	2.46	2	12.82	0.34	4.56	3	15.5	0.43	10.18	4	10.86	0.03	1.71	2
DCVI	10.01	0	0.02	2	11.34	0	0.02	2	10.01	0	0.02	2	10.03	0	0.06	2	10.01	0	0.02	2
DBCV	11.46	0.13	1.51	3	14.66	1.24	3.17	3	15.69	1.79	4.44	3	67.96	45.14	71.31	10	28.71	4.25	30.43	7
AIC	10.01	0	0.03	2	11.35	0	0.03	2	10.01	0	0.03	2	10.03	0	0.06	2	10.02	0	0.04	2
BIC	10.01	0	0.03	2	11.35	0	0.03	2	10.01	0	0.03	2	10.03	0	0.06	2	10.02	0	0.04	2
IP	37.38	11.91	**44.58**	8	14.52	0.43	3.28	2	12.48	0.39	4.11	2	53.08	17.88	56.16	8	25.01	3.63	26.38	5

Table 46: Swin based DBSCAN clustering results on five image datasets.

	CIFAR-10 - 10				**MNIST - 10**				**FashionMNIST - 10**				**ImageNet-10 - 10**				**CINIC-10 - 10**			
	ACC	ARI	NMI	opt_k	ACC	ARI	NMI	opt_k	ACC	ARI	NMI	opt_k	ACC	ARI	NMI	opt_k	ACC	ARI	NMI	opt_k
SD	10.01	0	0.02	2	15.31	0.78	8.24	3	10.01	0	0.02	2	56.1	21.67	59.71	10	10.01	0	0.02	2
Dunn	10.01	0	0.02	2	11.37	0	0.04	2	10.01	0	0.02	2	10.02	0	0.03	2	10.01	0	0.02	2
I	10.01	0	0.02	2	11.37	0	0.04	2	10.01	0	0.02	2	10.02	0	0.03	2	10.01	0	0.02	2
XB	10.01	0	0.02	2	11.37	0	0.04	2	10.01	0	0.02	2	10.02	0	0.03	2	10.01	0	0.02	2
S	10.01	0	0.02	2	11.37	0	0.04	2	10.04	0	0.08	2	10.02	0	0.03	2	10.04	0	0.08	2
CH	**33.74**	**18.58**	**42.17**	5	**18.98**	**3.28**	5.89	2	**19.23**	**3.09**	**16.7**	3	**73.78**	**43.41**	**73.46**	10	14.56	1.27	2.48	2
DB	10.01	0	0.02	2	11.15	-0.01	0.38	2	10.01	0	0.02	2	10.02	0	0.03	2	10.01	0	0.02	2
S_Dbw	13.62	0.13	7.21	11	13.77	0.31	6.63	4	10.52	0	1.16	7	13.92	0.2	8.88	13	10.03	0	0.08	2
CVNN	12.9	0.11	5.5	5	13.63	0.29	6.37	3	11.31	0.03	2.25	4	12.82	0.09	5.55	6	14.87	0.27	9.23	8
DCVI	10.01	0	0.02	2	11.38	0	0.07	2	10.01	0	0.02	2	10.12	0	0.27	2	10.01	0	0.02	2
DBCV	23.44	1.51	23.05	15	16.18	1.23	**9.44**	3	13.68	0.36	7.42	3	57.78	22.57	60.58	10	22.38	1.95	24.05	12
AIC	10.01	0	0.03	2	11.37	0	0.04	2	10.02	0	0.04	2	10.02	0	0.03	2	10.04	0	0.08	2
BIC	10.01	0	0.03	2	11.37	0	0.04	2	10.02	0	0.04	2	10.02	0	0.03	2	10.04	0	0.08	2
IP	**33.74**	**18.58**	**42.17**	5	16.74	1.48	8.57	2	**19.23**	**3.09**	**16.7**	3	**73.78**	**43.41**	**73.46**	10	**26.19**	**4.29**	**28.26**	13

Table 47: BEiT based DBSCAN clustering results on five image datasets.

	CIFAR-10 - 10				**MNIST - 10**				**FashionMNIST - 10**				**ImageNet-10 - 10**				**CINIC-10 - 10**			
	ACC	ARI	NMI	opt_k	ACC	ARI	NMI	opt_k	ACC	ARI	NMI	opt_k	ACC	ARI	NMI	opt_k	ACC	ARI	NMI	opt_k
SD	10.01	0	0.02	2	11.36	0	0.02	2	10.01	0	0.02	2	10.01	0	0.02	2	10.01	0	0.02	2
Dunn	10.01	0	0.02	2	11.38	0	0.06	2	10.01	0	0.02	2	10.01	0	0.02	2	10.03	0	0.08	2
I	10.01	0	0.02	2	11.36	0	0.02	2	10.01	0	0.02	2	10.01	0	0.02	2	10.01	0	0.02	2
XB	10.01	0	0.02	2	11.36	0	0.02	2	10.01	0	0.02	2	10.01	0	0.02	2	10.01	0	0.02	2
S	10.01	0	0.02	2	11.36	0	0.02	2	10.01	0	0.02	2	10.01	0	0.02	2	10.02	0	0.04	2
CH	**14.26**	**1.24**	2.71	2	11.77	0	2.11	2	25.11	**15.58**	**33.44**	3	**17.05**	**2.29**	**8.19**	2	**13.08**	0.54	**2.69**	2
DB	10.01	0	0.02	2	11.36	0	0.02	2	10.01	0	0.02	2	10.01	0	0.02	2	10.01	0	0.02	2
S_Dbw	10.01	0	0.02	2	10.58	-0.04	1.72	10	12.08	0.23	6.15	20	11.46	0.05	3.18	12	10.03	0	0.08	2
CVNN	10.18	0	0.36	2	**20.73**	**2.42**	**14.48**	3	21.9	2.85	22.8	6	14.45	0.77	5.73	3	11.32	0.07	0.99	2
DCVI	10.01	0	0.02	2	11.36	0	0.02	2	10.01	0	0.02	2	10.01	0	0.02	2	10.01	0	0.02	2
DBCV	12.28	0.23	1.62	4	12.05	0.05	1.42	3	10.18	0	0.19	3	13.15	0.39	3.34	4	10.01	0	0.03	2
AIC	10.01	0	0.03	2	11.36	0	0.02	2	10.01	0	0.03	2	10.02	0	0.03	2	10.01	0	0.03	2
BIC	10.01	0	0.03	2	11.36	0	0.02	2	10.01	0	0.03	2	10.02	0	0.03	2	10.01	0	0.03	2
IP	13.03	0.55	2.73	2	**20.73**	**2.42**	**14.48**	3	**26.65**	15.04	31.26	3	14.01	0.65	5.23	2	13.04	**0.57**	1.95	2

E Evaluation Results on UCI Datasets

In this section, we compare our index with other 13 indices on five real-world datasets from UCI Repository (Frank, 2010). Table 48 lists the basic statistics of these datasets. The evaluation results based on external indices of ACC, ARI and NMI are shown in Table 49, where the best results are highlighted in bold and the optimal k each index select, i.e. opt_k, and the true cluster number in each dataset, i.e. $dataset\text{-}k$, are provided. Our index is either on par or slightly better than competing indices.

Table 48: Statistics of UCI datasets.

Dataset	Samples	Attributes	Classes
Wine	178	13	3
Satimage	6,435	36	6
Ecoli	336	7	8
ControlChart	600	60	6
Letter	20,000	16	26

Table 49: Clustering results in terms of ACC and ARI on five UCI datasets.

	Wine - 3				**Satimage - 6**				**Ecoli - 8**				**ControlChart - 6**				**Letter - 26**			
	ACC	ARI	NMI	opt_k	ACC	ARI	NMI	opt_k	ACC	ARI	NMI	opt_k	ACC	ARI	NMI	opt_k	ACC	ARI	NMI	opt_k
SD	**100**	**100**	**100**	2	99.72	99.56	99.18	3	79.63	75.43	78.02	2	**100**	**100**	**100**	2	**83.7**	**77.01**	81.24	2
Dunn	69.23	59.72	74.63	2	99.51	99.22	98.87	4	**83.91**	80.1	78.57	3	71.94	59.43	70.04	2	69.29	65.8	**88.24**	18
I	**100**	**100**	**100**	2	**100**	**100**	**100**	3	60.87	44.64	63.09	4	**100**	**100**	**100**	2	49.32	16.53	41.27	5
XB	**100**	**100**	**100**	2	**100**	**100**	**100**	3	81.63	**86.46**	**80.94**	2	**100**	**100**	**100**	2	33.92	14.59	69.3	52
S	**100**	**100**	**100**	2	**100**	**100**	**100**	3	82.61	84.81	80.41	2	**100**	**100**	**100**	2	33.92	14.59	69.3	52
CH	97.85	95.98	94.03	2	99.56	99.28	98.77	3	79.63	75.43	78.02	2	80.99	69.04	74.05	2	33.92	14.59	69.3	52
DB	**100**	**100**	**100**	2	100	100	100	3	81.63	**86.46**	**80.94**	2	**100**	**100**	**100**	2	49.32	16.53	41.27	5
S_Dbw	**100**	**100**	**100**	2	66.81	51.02	64.19	3	79.63	75.43	78.02	2	**100**	**100**	**100**	2	49.32	16.53	41.27	5
CVNN	**100**	**100**	**100**	2	**100**	**100**	**100**	3	81.63	**86.46**	**80.94**	2	**100**	**100**	**100**	2	74.58	0	0	2
DCVI	**100**	**100**	**100**	2	**100**	**100**	**100**	3	81.63	**86.46**	**80.94**	2	**100**	**100**	**100**	2	49.32	16.53	41.27	5
DBCV	**100**	**100**	**100**	2	**100**	**100**	**100**	2	82.61	84.81	80.41	2	**100**	**100**	**100**	2	**83.7**	**77.01**	81.24	2
AIC	60.11	40.26	50.3	2	23.75	0.02	0.15	2	44.18	3.39	10.11	2	33.96	14.47	39.49	2	4.09	0	0.04	2
BIC	60.11	40.26	50.3	2	23.75	0.02	0.15	2	44.18	3.39	10.11	2	33.96	14.47	39.49	2	4.09	0	0.04	2
IP	**100**	**100**	**100**	2	**100**	**100**	**100**	2	82.61	84.81	80.41	2	**100**	**100**	**100**	2	**83.7**	**77.01**	81.24	2

F TF-IDF BASED CLUSTERING RESULTS

Table 50: TF-IDF based clustering results in terms of ACC and ARI on five text datasets.

	SearchSnippets - 8				**Biomedical - 20**				**StackOverflow - 20**				**WebofScience - 7**				**Yahoo!Answers - 10**			
	ACC	ARI	NMI	opt_k	ACC	ARI	NMI	opt_k	ACC	ARI	NMI	opt_k	ACC	ARI	NMI	opt_k	ACC	ARI	NMI	opt_k
SD	70.09	55.18	81.01	6	30.77	1.32	30.4	10	**64.53**	**30.59**	**65.53**	15	17.8	0	0.05	2	21.14	0.24	9.04	16
Dunn	**86.76**	**80.73**	**81.3**	3	**87.17**	**74.01**	68.15	2	**64.53**	**30.59**	**65.53**	15	**91.52**	**87.06**	**86.94**	6	21.36	0.41	1.21	2
I	62.5	31.93	56.09	2	31.31	-2.1	14.99	2	58.08	25.7	57.4	8	18.01	0	0.09	3	21.71	0.41	2.81	3
XB	62.5	31.93	56.09	2	31.31	-2.1	14.99	2	51.77	21.39	50.17	6	17.8	0	0.05	2	21.71	0.41	2.81	3
S	14.37	9.43	53.9	167	17.14	8.16	31.47	135	30.01	16.06	39.78	132	11.46	7.05	35.63	103	16.88	0.99	16.02	47
CH	62.5	31.93	56.09	2	31.31	-2.1	14.99	2	58.57	29.24	66	31	32.55	23.71	38.76	2	21.14	0.24	9.04	16
DB	62.5	31.93	56.09	2	30.77	1.32	30.4	10	58.08	25.7	57.4	8	17.8	0	0.05	2	**21.71**	0.41	2.81	3
S_Dbw	62.5	31.93	56.09	2	31.31	-2.1	14.99	2	34.06	5.7	25.48	2	17.8	0	0.05	2	**21.71**	0.41	2.81	3
CVNN	19.76	-0.3	10.42	51	31.31	-2.1	14.99	2	51.77	21.39	50.17	6	17.8	0	0.05	2	**21.71**	0.41	2.81	3
DCVI	62.5	31.93	56.09	2	31.31	-2.1	14.99	2	58.08	25.7	57.4	8	17.8	0	0.05	2	**21.71**	0.41	2.81	3
DBCV	14.37	9.43	53.9	167	86.54	72.81	**68.46**	2	20.22	12.88	56.06	168	18.01	0	0.09	3	10.05	0	0.03	2
AIC	33.64	10.13	14.25	2	8.64	1.3	3.85	2	6.97	0.6	2.49	2	30.97	15.34	24.8	2	16.63	**2.79**	5.07	2
BIC	33.64	10.13	14.25	2	8.64	1.3	3.85	2	6.97	0.6	2.49	2	17.8	0	0.05	2	16.63	**2.79**	5.07	2
IP	33.64	10.13	14.25	2	8.64	1.3	3.85	2	6.97	0.6	2.49	2	39.17	18.1	31.19	3	16.21	1.79	3.41	3

In this section, we evaluate ther clustering results on five text datasets. Text representations are obtained by computing TF-IDF features on the 1500 most frequently occurring word stems. As shown in Table 50, our method degrades when the data is represented without the assumption of normal distribution.