Performance of Internal Cluster Validations Measures To Guide Evolutionary Clustering

Pranav Nerurkar[1], Madhav Chandane[2] and Sunil Bhirud[3]

[1] Dept. of CE & IT, VJTI, Mumbai,
pranavn91@upscfever.com
[2] Dept. of CE & IT, VJTI, Mumbai,
mmchandane@ce.vjti.ac.in
[3] Dept. of CE & IT, VJTI, Mumbai,
sgbhirud@ce.vjti.ac.in

Abstract. Clustering of data into coarse grained descriptions is a challenging task of unsupervised machine learning due to unavailability of spatial characteristics of data. Clustering is an NP-hard grouping problem and thus no deterministic polynomial time solution is possible for it. For this reasons there are advantages of using a meta heuristic (swarm intelligence) strategy to find the near global optimal solution to this problem. To effectively guide the agents of the swarm in the meta heuristic strategy a suitable cost function has to be decided in the beginning. Identification of a suitable cost function has an important role in success of the meta heuristic. This paper utilizes internal validation criteria as cost functions as they achieve the dual goals of Clustering which are Compactness and Separation. Out of the multiple internal validation criteria included in the literature, two are identified for this purpose viz. BetaCV and Dunn index. These were used as cost functions of the swarm optimizer meta heuristic (PSO-BCV and PSO-Dunn) for purpose of clustering. To demonstrate the validity of the proposed technique it was tested on real datasets belonging to various domains ranging from medical diagnosis to cellular biology. The results obtained on these datasets from PSO-BCV and PSO-Dunn were compared with other meta heuristics (Differential Evolution) as well as the traditional swarm optimizer based on distance based criteria (PSO). The analysis of the results highlighted the suitability of this approach.

Keywords: Evolutionary clustering, Swarm intelligence, Cluster analysis, Cluster validation, Optimization

1 Introduction

Unsupervised machine learning is considered a challenging task as no class labels are provided for optimizing the cost function of the learning. The labels are implied in the data and the learning technique focuses on finding these implicit labels using a data driven method. There is an absence of information related to spatial features of the data such as number of clusters, size, shapes, density etc.

Such a scenario poses difficulty in the job of effectively finding clusters in the data. In such cases domain knowledge of the data provides indispensable value to the task of clustering as it may substitute the unavailability of spatial characteristics of the data, thus aiding the process of clustering. Another approach which is relied upon is to use validation statistics to understand the presence or absence of clustering tendency and arrive at a “good” clustering solution [1]. Cluster validation measures (CVI) are useful as the techniques which are applied on the data might not be able to find natural groupings present in the data [2].

CVI are of three types namely internal cluster validation measures, external cluster validation measures and relative validation measures. The first type doesn’t rely on presence of ground truth labels to be provided with the data. It relies on statistics calculated from the clustering solution obtained by the unsupervised algorithm. It specifies a range bound value (cost) of the obtained clustering solution. This value is an indicator of whether the results of the clustering are suitable. Thus, such measures can be used to decide whether to draw inferences from the data or not [1] [2]. External validation indices are useful if ground truth labels are provided with the data, these find application in both classification (supervised learning) as well as clustering (unsupervised learning) [3]. Purity, Information gain, Rand index, Variation of information are examples of this type. Relative clustering measures are more suitable for the task of finding the optimal number of clusters in the data. As most of the real world data will not have class labels associated with it, internal validations indices could be more useful in assessment of clustering results.

Clustering algorithms present in the literature belong to different classes of families. Prominent among these are the partitioning based algorithms such as k-means, mini batch k-means, k-means++, partition around medoids among others [4]. A second type of algorithms belong to hierarchical clustering techniques. These include AGNES, DIANA, CHAMELEON, BIRCH and others which use efficient data structures such as “dendrograms” to arrange clusters in a hierarchy. Both these category of algorithms use heuristics to assign data points to their respective prototypes or centroids [5]. Both category of algorithms rely on heuristics for clustering the data. Since the search space for clustering is vast and more suitable approach for traversing through this space and arriving at a global maxima is by the use of evolutionary algorithms [6].

Evolutionary algorithms are biologically inspired meta-heuristics that rely on multiple unintelligent agents that traverse the search space in a parallel manner. The inherent structure of such algorithms lead to avoidance of local optima in the search space [6]. A second advantage of evolutionary algorithms is that they allow consideration of multiple solutions to the same data which is not the case seen in non evolutionary algorithms like K-Means, PAM [7]. However, the success of a suitable meta heuristic strategy depends on the selection of a suitable cost function. The role of the cost function is to evaluate the solutions generated by the meta heuristic at an iteration. This helps in determining which swarm particle is at the best position in that iteration. This information can be used by the meta-heuristic to align the individual swarm agents towards the current best position and determine their next course.

Use of internal validation criteria as a cost function has dual advantages. Firstly, it removes the need to test the validity of clustering results obtained by the algorithm. Secondly, as internal cluster validation indices rely on maximizing both separation and compactness, they might be able to avoid biases seen in single objective optimization strategies. In this paper it is argued that a swarm optimizer based on validation criteria would perform better than one on traditional distance based heuristics. The reasoning behind this intuition is that use of such indices would convert the problem of clustering into a dual-objective optimization problem. This would lead to better clustering as otherwise a technique based on a single objective would be biased towards detection of only a single type of clusters.

Section II covers theoretical aspects of the evolutionary strategy used and describes the internal validation measures used in this paper. Section III covers experimental work and presents results and their critical discussion, followed by a conclusion in Section IV.

2 Related Work

Literature review has two subsections. In the first the structure of a general swarm based meta-heuristic strategy is described. In the second part, popular internal cluster validation indices seen in the literature are described.

2.1 Structure of a Meta Heuristic

The mathematical model of a meta heuristic algorithm has structured sections with each section having a logical meaning. These sections are Problem definition, Parameters definition, Initialization and the Iterative procedure.

Cost Function The key component of the problem definition is the cost function which is to be optimized using the meta-heuristics. Cost function indicates the cost of an individual solution and thus evaluates its "goodness". The cost function used in this paper is described in Section III.

Decision Variables The decision variables for the purpose of clustering are the k centroids of the data. Each centroid can be represented as a n dimensional vector $\overrightarrow{R^n}$. Thus the total decision variables in the problem are $n * k$. The range of the decision variables is limited to be between the maximum $VarMax$ and minimum value $VarMin$ of the points in the data, this also achieves the objective of restricting the search space.

2.2 Parameters of Meta heuristics

The parameters of the meta heuristics are maximum iterations allowed for the model, the size of the swarm, coefficients of inertia w and its damping ratio w_{damp} , personal acceleration coefficient c_1 (assigned to every particle) and the social acceleration coefficient c_2 (assigned to the entire swarm). In addition, limits are set to restrict the velocities of individual particles to the range $VelMin$ to $VelMax$. Constrictions coefficients phi_1, phi_2, phi, chi are also defined as per the Eqn.1 and Eqn.2.2. These additional parameters have been introduced on the basis of the work done by Bratton and Kennedy in [8]. The values for these variables are set according to standard settings and may be tuned to improve results.

$$phi = phi_1 + phi_2 \tag{1}$$

$$chi = \frac{2}{phi - 2 + \sqrt{phi^2 - 4 * phi}} \tag{2}$$

$$w = chi \tag{3}$$

$$c_1 = phi_1 * chi \tag{4}$$

$$c_2 = phi_2 * chi \tag{5}$$

2.3 Initialization and the Iterative Procedure

Each particle in the swarm has components such as position x_{pos}, velocity x_{vel}, cost x_{cost}, best position x_{pbest} and best cost x_{cbest}. The Global best x_{gbest} is set initially to $-\infty$ or ∞ depending on whether the objective function is to be maximized or minimized. The positions of the particles are initialized using uniform distribution between $VarMax$ and $VarMin$. The velocities of the particles is initialized to zero. The values for the best position of a particle and the cost at the best position is currently set to the initial position and the cost at the initial position respectively.

The iterative procedure updates the values for the best position and best cost obtained so far by each particle in the swarm and also update the variable that records the global best cost and position obtained by the swarm together. The update rule for the particle velocity is given by Eqn. 6

$$x_{i+1}(vel) = w * x_i(vel) + c_1 * (x_{pbest} - x_{pos}) + c_2 * (x_{gbest} - x_{pos}) \tag{6}$$

The velocity limits are applied as in Eqn: 7 and 8 to prevent the particle velocity from increasing or decreasing beyond the thresholds.

$$x_{i+1}(vel) = max(x_{i+1}(vel), VelMin) \tag{7}$$

$$x_{i+1}(vel) = min(x_{i+1}(vel), VelMax) \tag{8}$$

The position of the particle is updated by the Eqn 9

$$x_{i+1}(pos) = x_i(pos) * x_{i+1}(vel) \tag{9}$$

If the position of the particle is outside the thresholds $VarMax, VarMin$ it is updated by the rule in Eqn: 10 and 11

$$x_{i+1}(pos) = max(x_{i+1}(pos), VarMin) \tag{10}$$

$$x_{i+1}(pos) = min(x_{i+1}(pos), VarMax) \tag{11}$$

If the cost of the particle at its new position is better than the cost of the particle at its old position, then the values of personal best cost x_{cbest} and position x_{pbest} of the particle are updated to the new values. The values of the global best cost and position of the entire swarm is updated at every iteration to store the best global value at every iteration.

2.4 Internal Validation Criteria

Calinski-Harabasz index [1] [9] It is the average inter and intra-cluster sum of squared distances. A high value indicates compact and well separated clusters.

$$\frac{\sum_i n_i d^2(c_i, c)/(NC-1)}{\sum_i \sum_{x \in c_i} d^2(x, c_i)/(n-NC)} \tag{12}$$

I Index [1] [9] It is the ratio of maximum distance between centroids and sum of distances of objects to the centroid of their cluster. Ideally should have a maximum value.

$$(\frac{1}{NC} * \frac{\sum_{x \in D} d(x, c))}{\sum_i \sum_{x \in c_i} d(x, c_i)} * max_{i,j} d(c_i, c_j))^P \tag{13}$$

Dunn's Indices [1] [9] Ratio of minimum inter cluster distance to the maximum intra cluster distance. Ideally should be high.

$$min_i(min_j(\frac{min_{x \in c_i, y \in c_j} d(x, y)}{max_k(max_{x,y \in c_k} d(x, y))})) \tag{14}$$

BetaCV [1] [9] Ratio of mean intra cluster distance to mean inter cluster distance. A smaller value indicates better clustering.

$$\frac{d_{intra}}{d_{inter}} \tag{15}$$

$$d_{intra} = avg_{x \in c_i, y \in c_i} d(x, y)) \tag{16}$$

$$d_{intra} = avg_{x \in c_i, y \in c_j} d(x, y)) \tag{17}$$

Silhouette index [1] [9] Higher value is ideal as it indicates how similar a value placed in a cluster is to other members its own cluster compared to members belonging to other clusters.

$$\frac{1}{NC} \sum_i \left(\frac{1}{n_i} \sum_{x \in C_i} \frac{b(x) - a(x)}{max[b(x) - a(x))]}\right) \tag{18}$$

Davies Bouldin index [1] [9] For every point the similarity value to each cluster is calculated. The highest of these values is given to the point. The DB-index of the data is the average of the values of all points. Smaller value indicates clusters are distinct from each other.

Xie Beni index [1] [9] It is ratio of minimum square distance between centroids and mean square distance between each point and its centroid. Lower value is ideal.

$$[\sum_i \sum_{x \in C_i} d^2(x, c_i)]/[n * min_{i, j \neq i} d^2(c_i, c_j)] \tag{19}$$

Where,

- D : Data-set
- C_i: i^{th} cluster
- n_i: number of points in C_i
- c_i: centroid of C_i
- P: number of attributes in D
- NC: number of clusters
- $d(x, y)$: L_2 norm distance between x and y.
- $a(x)$: average dissimilarity of x with all other data in same cluster.
- $b(x)$: lowest average dissimilarity of x to other clusters to which it doesn't belong to.

Other validation indices seen in the literature are symmetry distance based index $Sym - index$ and the composite density between and within-cluster index $CDbw$. $Sym - index$ is applicable to only datasets that are internally symmetric whereas $CDbw$ gives unstable results on some datasets as it may not find centroids for each cluster [9]. For these reasons these are excluded from this section. In the literature, no reference was found for demonstrating the use of internal cluster validation indices as cost functions.

3 Mathematical Model

Dunn's validation index is conceptually the simplest of the internal validation indices and suitable to be a cost function [10]. BetaCV was also a suitable candidate for a cost function as it considers mean intra cluster and inter cluster distances and shall be less sensitive to outliers in data compared to indices like Xie-Beni and I-index as they rely on minimum or maximum distance based criteria [9].

3.1 Cost Function - BetaCV and Dunn index

Algorithm 1: Compute Dunn Index

```
Result: Returns Dunn index of the data
distMat = euclidean distance matrix of data;
ind = column vector of cluster ids;
minInter = 10000;
maxIntra = -1;
for $i \leftarrow 1$ to $max - clust - 1$ do
    for $j \leftarrow (i + 1)$ to $max - clust$ do
        temp = minWeights(distMat, ind, i, j);
        if $temp < minInter$ then
            minInter = temp;
for $i \leftarrow 1$ to $max - clust$ do
    temp = maxWeights(distMat, ind, i, i);
    if $temp > maxIntra$ then
        maxIntra = temp;
result = minInter / maxIntra;
```

Algorithm 2: Compute BetaCV Index

```
Result: Returns BetaCV index of the data
distMat = pdist2(X, X);
ind = column vector of cluster ids;
Win = 0;
for $i \leftarrow 1$ to $max - clust$ do
    Win = Win + sumWeights(distMat, ind, i, i);
Wout = 0;
for $i \leftarrow 1$ to $max - clust - 1$ do
    for $j \leftarrow (i + 1)$ to $max - clust$ do
        Wout = Wout + sumWeights(distMat, ind, i, j);
Nin = 0;
for $i \leftarrow 1$ to $max - clust$ do
    n = sum(ind == i);
    Nin = Nin + (n * (n-1) / 2);
Nout = 0;
for $i \leftarrow 1$ to $max - clust - 1$ do
    for $j \leftarrow (i + 1)$ to $max - clust$ do
        Nout = Nout + sum(ind == i) * sum(ind == j);
result = (Win / Nin) / (Wout / Nout);
```

The above algorithms are used as cost functions to guide the Swarm based meta heuristic described in Section II.

4 Experimental Study

Real Datasets from fields such as Cellular Biology [EColi], Cancer detection [Wisc], Plant species [Iris], Wine types [Wine] and yeast families [Yeast] are used for the cluster analysis. Swarm optimizer using Dunn index (PSO-Dunn) and BetaCV (PSO-BCV) as cost functions are compared with the Particle Swarm Optimizer [PSO] that uses distance based heuristics [11] [8] [12] as well as other meta heuristic such as Differential Evolution [13]. Distance based statistics obtained from clusters such as separation index, widest within-cluster gap, average silhouette width and external validity criteria viz. Variation of Information index [14] and Corrected Rand index [15] are used as metrics to measure performance.

4.1 Dataset

The datasets used for the experiment are given below:

Table 1. Description of the datasets

Sr. No	Name	Size	Attributes	Classes
1	EColi	336	7	8
2	Wisc	628	9	2
3	Wine	178	13	3
4	Iris	150	4	3
5	Yeast	1484	8	9

4.2 Results

Distance based Statistics Separation index (SI): is computed based on the distances for every point to the closest point not in the same cluster. Lower value means clusters are well separated. As both Dunn and BetaCV require high inter-cluster distances, the clustering obtained from these has better SI on two of the five datasets than other meta heuristics.

Table 2. Separation Index

Name	EColi	Wisc	Wine	Iris	Yeast
PSO	0.08	5.52	24.64	0.41	0.06
PSO-Dunn	0.11	**4.82**	70.1	1.1	0.36
PSO-BCV	0.09	5.6	29	**0.41**	0.44
DE	0.09	5.52	29.45	0.28	0.13

Widest within-cluster gap (Wgap): Largest link in within-cluster minimum spanning tree is computed. Larger value indicates good clustering. BetaCV and Dunn index based clustering has higher $Wgap$ on three of the five datasets compared to other meta heuristics. As both measures improve compactness, $Wgap$ might increase as close points are clustered together.

Table 3. Widest within-cluster gap

Name	EColi	Wisc	Wine	Iris	Yeast
PSO	0.55	9.16	133.22	0.81	0.56
PSO-Dunn	**0.72**	**9.16**	133.2	0.73	0.58
PSO-BCV	0.54	**9.16**	133.2	**0.82**	0.5
DE	0.55	9.16	133.2	0.82	0.6

Silhouette value (SH): is a measure of how similar an object is to its own cluster compared to other clusters. The silhouette ranges from -1 to +1, where a high value indicates that the object is well matched to its own cluster and poorly matched to neighboring clusters. BetaCV and Dunn index based clustering has higher SH on four of the five datasets compared to other meta heuristics. This could be as both measures seek to improve compactness.

Table 4. Average silhouette width

Name	EColi	Wisc	Wine	Iris	Yeast
PSO	0.27	0.6	0.57	0.55	0.21
PSO-Dunn	0.23	0.53	**0.6**	0.51	0.34
PSO-BCV	**0.29**	0.6	0.57	**0.55**	**0.48**
DE	0.25	0.6	0.57	0.53	0.28

External Validation Criteria : Variation of information (VI): is a measure of the distance between two clusterings and a less value indicates good clustering. BetaCV and Dunn index based clustering has higher VI on two of the five datasets compared to other meta heuristics.

Table 5. Variation of Information

Name	EColi	Wisc	Wine	Iris	Yeast
PSO	1.11	0.31	1.23	0.51	2.72
PSO-Dunn	**1.1**	0.67	1.06	0.51	2.37
PSO-BCV	1.26	0.33	1.24	0.55	**2.35**
DE	1.56	0.31	1.26	0.46	2.38

Corrected Rand Index (CRI): represents the frequency of occurrence of "agreements" over the total pairs of points. Agreements indicates whether the randomly chosen points lie in the same cluster in both partitions. Ideally should be equal to 1 (Range: -1 to 1). BetaCV and Dunn index based clustering has higher CRI on one of the five datasets.

Table 6. Corrected Rand Index

Name	EColi	Wisc	Wine	Iris	Yeast
PSO	0.56	0.86	0.39	0.74	0.07
PSO-Dunn	**0.62**	0.45	0.28	0.56	0
PSO-BCV	0.58	0.84	0.37	0.73	0
DE	0.51	0.86	0.37	0.8	0.13

5 Conclusion

Internal cluster validation indices have advantages as cost functions as they seek to achieve dual objectives of separation and compactness. This means clustering can become a dual objective optimization. However this also introduces additional overhead to the task of clustering due to the higher computational cost compared to traditional distance based heuristics used in clustering. Meta heuristic approach is used for clustering due to the advantage it provides over traditional clustering algorithms in terms of lower computation cost, parallel execution, faster convergence, effective traversal of search space and the ability to avoid local optimal solutions. This paper utilizes two internal clustering validation criteria viz. BetaCV and Dunn index for guiding the agents of the meta heuristic to reach the global optima. Both criteria are chosen due to their conceptual simplicity and robustness against outliers compared to other indices. The results have shown that the strategy provided in this paper can achieve effective clustering and results comparable in performance with other meta heuristics.

References

[1] 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*, 43(3):982–994, 2013.

[2] Ariel E Baya and Pablo M Granitto. How many clusters: A validation index for arbitrary-shaped clusters. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 10(2):401–414, 2013.

[3] Gongde Guo, Lifei Chen, Yanfang Ye, and Qingshan Jiang. Cluster validation method for determining the number of clusters in categorical sequences. *IEEE transactions on neural networks and learning systems*, 2016.

[4] Rui Xu and Donald Wunsch. Survey of clustering algorithms. *IEEE Transactions on neural networks*, 16(3):645–678, 2005.

[5] Anil K Jain. Data clustering: 50 years beyond k-means. *Pattern recognition letters*, 31(8):651–666, 2010.

[6] Deepayan Chakrabarti, Ravi Kumar, and Andrew Tomkins. Evolutionary clustering. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 554–560. ACM, 2006.

[7] Eduardo Raul Hruschka, Ricardo JGB Campello, Alex A Freitas, et al. A survey of evolutionary algorithms for clustering. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 39(2):133–155, 2009.

[8] Daniel Bratton and James Kennedy. Defining a standard for particle swarm optimization. In *Swarm Intelligence Symposium, 2007. SIS 2007. IEEE*, pages 120–127. IEEE, 2007.

[9] Jiawei Han, Jian Pei, and Micheline Kamber. *Data mining: concepts and techniques.* Elsevier, 2011.

[10] Lori Dalton, Virginia Ballarin, and Marcel Brun. Clustering algorithms: on learning, validation, performance, and applications to genomics. *Current genomics*, 10(6):430–445, 2009.

[11] Russell Eberhart and James Kennedy. A new optimizer using particle swarm theory. In *Micro Machine and Human Science, 1995. MHS'95., Proceedings of the Sixth International Symposium on*, pages 39–43. IEEE, 1995.

[12] Maurice Clerc. *Particle swarm optimization*, volume 93. John Wiley & Sons, 2010.

[13] Rainer Storn and Kenneth Price. Differential evolution–a simple and efficient heuristic for global optimization over continuous spaces. *Journal of global optimization*, 11(4):341–359, 1997.

[14] Marina Meilă. Comparing clusterings—an information based distance. *Journal of multivariate analysis*, 98(5):873–895, 2007.

[15] Christian Hennig. How many bee species? a case study in determining the number of clusters. In *Data Analysis, Machine Learning and Knowledge Discovery*, pages 41–49. Springer, 2014.