Review of Different Automatic Evaluation of Clusters Based on a Cluster Validation Indexes

S. A. Magdum, H. A. Tirmare

Abstract— **Clustering is an important technique in data mining. In clustering process the given data points partitioned into a group based on some similarity measures. There are number of clustering algorithms presented and applied in many fields, such as image analysis, data mining, pattern recognition and Bioinformatics. For most clustering algorithms, the number of clusters is one of the essential parameters. However, this parameter is initially not available for most of the data sets. Thus, this paper intends to overcome this problem by proposing different algorithm for automatic clustering without knowing the value of k that is the number of clusters.**

Index Terms— **Genetic Algorithms (GA), Automatic clustering, Validation, Particle Swarm Optimization (PSO).**

I. Introduction

Data mining is a process used for finding the relationship and patterns that exist in large database. The objective of data mining is to identify valid, useful, and understandable correlations and patterns in existing data. Also the term "data mining" is primarily used by statisticians, researchers, and the MIS and business communities [11]. The term Knowledge Discovery in Databases (KDD) is generally used to refer to the overall process of discovering useful knowledge from data, where data mining is a particular step in this process.

Clustering is an unsupervised classification technique which partitions a set of objects in such a way that objects in the same clusters are more similar to one another than the objects in different clusters according to certain predefined criterion [6]. Clustering algorithms can be classified into two types: hierarchical and partitional [6]. Hierarchical clustering starts grouping objects using bottom-up or top-down approaches and forms a tree like structure. The Hierarchical algorithms are either agglomerative or divisive. In agglomerative type the clustering starts with considering each object has its own cluster and combine them into a single cluster based on some objective function. Divisive algorithm takes a group of objects as one cluster and then partition the objects into subgroups based on similarity measures.

Determining the appropriate number of clusters from a given data set is an important consideration in clustering [1]. A procedure for determining the optimal number of clusters is shown in Fig.1. Given the data set *X*, a specific clustering algorithm and a fixed range of number of clusters [*Mmin*, *Mmax*], the basic procedure involves [9]:

1. Repeat a clustering algorithm successively for the number of clusters *M* from a predefined range [*Mmin*, *Mmax*].
2. Obtain the clustering results (partitions *P* and centroids *C*) and calculate the validity index value for each.
3. Select the *M** for which the partitioning provides the best result, according to the validity index.
4. Compare the *M** with external information if available.

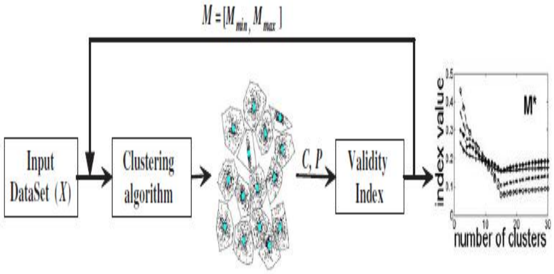

Figure 1: Determining the number of clusters in cluster validity analysis.

In order to find the proper value of k and validation of obtaining partition requires some validation techniques. The process of estimating how well a partition fits the structure underlying the data is known as cluster validation [4]. This validation can be of two type's internal or external validation. For external validation a prior knowledge of a data set is required, but in practice it is difficult to have prior information of the dataset [7]. In internal validation technique we don't require prior knowledge. Moreover, external validation validates a partition by comparing it with the correct partition. And internal validation validates a partition by examining just the partitioned data. When the correct partition is available the usual approach is to compare it with the partition proposed by the clustering algorithm based on one of the many indices that compare data partitions; e.g. Rand, Adjusted Rand, Jaccard, Fowlkes–Mallows, Variation of Information [8]. On the other hand, when a correct partition is not available the indexes used for validating clusters are focused on the compactness and separation of the clusters; e.g., Dunn index [7], CH index [10], Devies -Bouldin [4].

This paper giver review of the recent work presented for automatic evaluation of clusters based on a different cluster validation index (CVI).

Manuscript received April, 2015.

***S.A.Magdum**, Department of Technology, Shivaji University., Kolhapur, India, Mobile No 07387242202*
***H.A.Tirmare**, Department of Technology, Shivaji University Kolhapur, India, Mobile No 09225802907*

II. AUTOMATIC CLUSTERING TECHNIQUES

A. A Two stage genetic algorithm for automatic clustering:

This algorithm is proposed by Hong he and Yonghong Tan which determines the proper number of clusters and partitions from a given dataset [5]. The two stage genetic clustering algorithms (TGCA) work in two stages; first one is selection operation and second is mutation operation. The TGCA uses variable length chromosomes that allow the algorithm to effectively search for optimum value for cluster centers and the number of clusters. The chromosomes are represented by the cluster center based represented and length of chromosomes is the total size of the population.

The initial value of cluster centroids is decided by the attribute filter method. This method carefully selects seeds for clusters and it uses Euclidean distance to compare objects with the centroids. This genetic algorithm considers two fitness functions one as squared Euclidean distance and Calinski Harabasz (CH) index. The index is based on internal cluster cohesion and external cluster isolation. The internal cohesion is calculated by the Within Group Sum of Square Distance (WGSD).

$$\mathrm{WGSD} = \textstyle\sum_{i=1}^{k} \sum_{X=ci} d(X, Zi)^2$$

The external isolation is calculated by the Between Groups sum of Squared Euclidean Distance (BGSD).

$$\mathrm{BGSD} = \textstyle\sum_{i=1}^{k} bi\, d(Zi, Ztot)^2$$

Where, Z*tot* is the center of all objects. The CH index is defined as,

$$CH_K = \frac{BGSD}{K-1} \frac{M-K}{WGSD}$$

Where, M is the total objects in the dataset.

K=k [$k_{min,}$ k_{max}] min=2, max=√M.

B. A symmetry based multiobjective clustering:

This technique is proposed by Sriparna Saha, and Sanghamitra Bandyopadhyay which not only produce the clusters automatically, but also gives optimization strategy [1]. This method assigns the data points to different clusters based on newly developed point symmetry based distance rather than Euclidean Distance. Here two cluster validity index are used, one is based on Euclidean Distance as XB-index and another is pointing symmetry based cluster validity index Sym-Index to determine the appropriate number of clusters in a dataset.

Initially K centers are calculated by using

$$Ki = (rand * mod\ (Kmax - 1)) + 2$$

Where, rand () function returns an integer and Kmax is the upper bound of number of clusters.

The XB index (Xie-Beni) is a ratio of within the cluster compactness to between cluster separations.

$$XB = \frac{\sigma(Z;X)}{Sep(Z)} = \frac{\sum_{i=1}^{k} (\sum_{k=1}^{ni} d^2\, e\ (\bar{Z\imath}, \bar{X}_k)}{n(\min_{i \neq j}(||\bar{Z\imath} - \bar{Z}_j||)2)}$$

In Sym-Index cluster separation is measured between any two clusters centers and it is given as,

$$\mathrm{Sym}(t) = \frac{1}{k} \times \frac{1}{\epsilon_k} \times D_k$$

Where,

$$\epsilon_k = \textstyle\sum_{i=1}^{k} E_i \ ; \quad E_i = \sum_{j=1}^{ni} d_{ps}(\overline{X_J}, \bar{Z}_i)$$

$$D_K = max_{ij=1}^{k} ||\bar{Z\imath} - \overline{Z_J}||$$

When the partitioning is compact and good, the total deviation (σ) should be low while the minimal separation (*Sep*) between any two cluster centers should be high. Thus, the objective is therefore to minimize the XB-index for achieving the proper clustering. *DK* is the maximum Euclidean distance between two cluster centers among all centers. $d_{ps} = (\bar{x_j^i}, \bar{z_i})$ is the point symmetry based distance [12] between the *j*th point of the *i*th cluster, xij, and the cluster center *zi* and *ni* is the total number of points in the *i*th cluster.

C. A generalized Automatic clustering Algorithm:

This is another technique introduced recently by Sriparna Saha, and Sanghamitra Bandyopadhyay [3]. In this method each cluster is divided into several small hyper spherical sub clusters and the centers of all these small sub-clusters are encoded in a string to represent the whole clustering. These sub clusters are merged appropriately to form global clusters. Three objective functions, one reflecting the total compactness of the partitioning based on the Euclidean distance, the other reflecting the total symmetry of the clusters, and the last reflecting the cluster connectedness, are considered here.

To achieve three objectives three different cluster validity indexes are used here. To find the amount of symmetry present in a particular partition is given by Sym-Index which is described in previous method. To find out total compactness of giving partitions in terms of the Euclidean distance I-index is used. And to quantify total connectedness of the clusters is given by the Con - Index.

Con-Index: It is connectivity based cluster validity index. The connectivity between a set of points is measured using Relative Neighborhood Graph (RNG). The distance between a pair of points is measured in the following way.

- Construct the relative neighborhood graph of the whole data set.
- The distance between any two points, **x** and **y**, denoted $d_{short}(x, y)$

The $d_{short} = (x, y)$ is given as,

$$d_{short}\ (x, y) = min_i^p\, max_j^{n\ edge\ j}\ w(ed_j^i)$$

Where *p* are the number of possible paths between two points, *w* denotes the weight of the edge along the shortest path between x and y.

The Con-Index is defined as,

$$con = \frac{\sum_{i=1}^{k} \sum_{j=1}^{n_k} d_{short}(\bar{m}_i, \bar{x}_j^i)}{n \times min_{i,j=1}^{k}\ \Lambda_{i \neq j}\, d_{short}(\bar{m}_i, \bar{m}_j)}$$

Where d_{short} is the shortest distance between x and y along the relative neighborhood graph. The $d_{short}(\bar{m}_i, \bar{x}_j^i)$ gives the cluster connectedness between ith cluster and jth point and $d_{short}(\bar{m}_i, \bar{m}_j)$ shows cluster separation.

I- Index: This is Euclidean distance based cluster validity index. It is represented as

$$I(K) = (\frac{1}{K} \times \frac{\epsilon\ 1}{\epsilon\ k} \times D_K)^p$$

Where k is number of clusters and D*k* measures the maximum separation between two clusters over all possible pairs of clusters, which try to reduce I-Index as K is increased. The value of K for which I-Index takes its maximum value is considered as the appropriate number of clusters.

D. Automatic Clustering using Improved Differential Evolution:

Differential Evolution (DE) is one of the fast and efficient global search heuristics. The Swagatam et al. has proposed a new method for automatic clustering by using Improved Differential Evolution [2]. The algorithm is modified to improve convergence speed of the classical DE. This method proposes a solution to automatic clustering of large unlabeled data sets. It is one of the agglomerative types of algorithm which combines the partition according to some similarity measure. Here two cluster validity indexes are used; DB Index and CS Index.

DB Index: this measure is a function of the ratio of the sum of within cluster scatter to between cluster separations. Within ith cluster scatter is given as,

$$S_{i,q} = \Big[\frac{1}{N_i} \sum_{\overline{X \in C_i}} \big\|\vec{X} - \overrightarrow{mi}^{1/q}$$

Where m_i is the ith cluster, q>= 1 selected independently and N_i is the number of elements in the ith cluster C_i. The distance between ith and jth cluster is given as,

$$d_{ij,t} = \|\overrightarrow{m_i} - \overrightarrow{m_j}\|\, t$$

The $R_{i,qt}$ is defined as,

$$R_{i,qt} = \max_{j\in K\ j\neq i} \left\{ \frac{S_{i,q} + S_{j,q}}{d_{ij,t}} \right\}$$

Finally, DB measure is defined as,

$$DB(K) = \frac{1}{K} \sum_{i=1}^{k} R_{i,qt}$$

The smallest DB (K) index indicates a valid optimal partition.

CS-Measure: Chou et al. have proposed CS measure for evaluating the validity of a clustering scheme. Before applying the CS measure, the centroid of a cluster is computed by averaging the data vectors that belong to that cluster using

$$\vec{m_i} = \frac{1}{N_i} \sum_{x_j \in c_i} \vec{x_i}$$

A distance metric between any two data points $\vec{X_i}$ and $\vec{X_j}$ is denoted by. $d(\vec{X_i}, \vec{X_j})$ Then, the CS measure can be defined as

$$CS(K) = \frac{\sum_{i=1}^{k} \left[\frac{1}{N_i} \sum_{\vec{X}_i \in c_i} \max_{\vec{X}_q \in c_i} \{ d(\vec{X_i}, \vec{X_q}) \} \right]}{\sum_{i=1}^{K} \left[\min_{j \in K, j \neq i} \{ d(\overrightarrow{m_i}, \overrightarrow{m_q}) \} \right]}$$

According to Chou et al., the CS measure is more efficient in tackling clusters of different densities and/or sizes than the other popular validity measures, the price being paid in terms of high computational load with increasing *K* and *n*.

E. An Automatic Clustering Technique for Optimal Clusters:

Automatic Merging for Optimal Clusters (AMOC) which aims to generate nearly optimal clusters for the given datasets automatically. The AMOC is an extension to standard k-means with a two phase iterative procedure combining certain validation techniques in order to find optimal clusters with automation of merging of clusters. A number of criteria have been proposed to measure how close the obtained clustering is to a ground truth clustering, such as Rand index [13].

Rand Index (RI):

Rand's Index [12] was motivated by standard classification problems in which the result of a classification scheme has to be compared to a correct classification. For Rand, comparing two clusters was just a natural extension of this problem which has a corresponding extension of the performance measure: instead of counting single elements he counts correctly classified pairs of elements. Thus, the Rand Index is defined by

$$R(C, C') = \frac{a+b}{a+b+c+d} = \frac{a+b}{\binom{n}{2}}$$

Where a + b can be considered as the number of agreements between X and Y (two clusters) and c + d as the number of disagreements between X and Y.

Rand Index has a value between 0 and 1 with 0 indicating that the two data clusters do not agree on any pair of points and 1 indicating that the data clusters are exactly the same.

III. Conclusion

Clustering is an unsupervised classification scheme where prior knowledge of the data set is not always available. The cluster validity indexes plays important role in automatic clustering to find out validity of produced clusters. Almost all validity indexes relay on inter cluster separation and intra cluster similarity measures. These indexes can be used in different configuration of data sets depending upon their usefulness. Not all validity indexes are useful in all cases, depending upon information available the use of internal or external validity indexes in deciding.

Acknowledgment

Our thanks to the Department of Technology, Shivaji University for allowing us to go ahead with this system. My thanks to all persons who directly or indirectly helped me to complete this work.

REFERENCES

[1]. Sriparna Saha, Sanghamitra Bandyopadhyay, A symmetry based multiobjective clustering technique for automatic evolution of clusters, Pattern Recognition 43 (2010) 738 – 751.

[2]. Swagatam Das, Ajith Abraham, Amit Konar, Automatic Clustering Using an Improved Differential Evolution Algorithm, IEEE transactions on systems, man, and cybernetics—part a: systems and humans, vol. 38, no. 1, January 2008.

[3]. Sriparna Saha, Sanghamitra Bandyopadhyay, A generalized automatic clustering algorithm in a multiobjective framework, Applied Soft Computing 13 (2013) 89–108.

[4]. Olatz Arbelaitz, Ibai Gurrutxaga , Javier Muguerza Jesu´s M.Pe´ rez, In˜igo Perona, An extensive comparative study of cluster validity indices, Pattern Recognition 46 (2013) 243–256.

[5]. Hong He, Yonghong Tan, A two-stage genetic algorithm for automatic clustering, Neurocomputing 81 (2012) 49-59.

[6]. Jain, A. K., Murty, M. N., Flynn, P. J. (1999). Data clustering: A review. ACM Computing Surveys, 31, 264–323.

[7]. Eréndira Rendón, Itzel Abundez, Alejandra Arizmendi and Elvia M. Quiroz, Internal versus External cluster validation indexes, international journal of computers and communications, Issue 1, Volume 5, 2011.

[8]. V. Batagelj, M. Bren, Comparing resemblance measures, Journal of Classification 12 (1995) 73–90.

[9]. Qinpei Zhao and Pasi Fränti, Senior Member "Centroid Ratio for a Pairwise Swap Clutering Algorithm" IEEE transactions on knowledge and data engineering, VOL. 26, NO. 5, MAY 2014.

[10]. T. Calinski and J. Harabasz, "A dendrite method for cluster analysis," Communication in statistics 3, 1–27 (1974).

[11]. Joyce Jackson "Data Mining: A Conceptual Overview" Communications of the Association for Information Systems Volume 8, 2002

[12]. Rand, William M.: Objective Criteria for the Evaluation of Clustering Methods. Journal of the American Statistical Association, 66(336):846–850, 1971.

[13]. K. Karteeka Pavan, Allam Appa Rao, A.V. Dattatreya Rao "An Automatic Clustering Technique for Optimal Clusters" (2011).

Surekha Magdum received the B.E. degree in Computer Science and Engineering from Shivaji University, India in 2012. She pursuing the M.Tech degree in computer science and technology from Shivaji University, Kolhapur, India in 2015.Her current research interests include clustering algorithm and multimedia processing.

Hemant Tirmare received the B.E. and M.E. degrees in Computer Science and Engineering from Shivaji University, India in 2000 and 2010 respectively. Since 2013, he has been a Assistant Professor with the same university. He has published 03 journals paper and more than 23 workshops are attended. His current research interests include clustering algorithms, Network Security.