Related Works

E. Goetz

I. Introduction

Although this document is not an exhaustive discussion of the field of unsupervised learning, it is an academic survey of serious clustering concepts and research findings. Understanding the following publications and their conclusions is crucial to understanding the design of this analysis and the current state of research in this area. As this project expands, this list of references will continue to grow.

References

[HV01] Maria Halkidi and Michalis Vazirgiannis. Clustering validity assessment: Finding the optimal partitioning of a data set. In *Proceedings - IEEE International Conference on Data Mining, ICDM*, pages 187–194, 02 2001.

In their publication, Halkidi and Vazirgiannis define a new validity index for clustering algorithms. Observing the criteria of these algorithms— they note— has always proved difficult as humans cannot perceive more than three dimensions. Resultantly, an entire branch of clustering research has been dedicated to developing verification methods. Current methods are divided into three categories: external basis, internal basis, and relative basis. But nearly all of these methods depend on *a priori* knowledge of the number of clusters that exist and they are devoid of any geometric relationship to the data. Therefore, the authors define a new relative, algorithm-independent validity index. This index, S_Dbw, measures cluster compactness with intra-cluster variance as well as cluster separation in terms of inter-cluster density. After defining these concepts mathematically, Halkidi and Vazirgiannis define S_Dbw as the sum of intra-cluster variance and inter-cluster density. Consequently, as the index of a clustering algorithm decreases, the separation between clusters and the compactness of clusters increases. Two mathematical proofs show S_Dbw is minimized when the correct number of clusters is used and when the optimal partitions are found. The authors briefly discuss the time complexity of S_Dbw (which is actually dependent on three variables) before evaluating the validity of their index. Using synthetic data sets, they show the ability of their algorithm to select the correct number of clusters and the optimal partition, illustrating their theoretical proofs. Subsequently, they compare their own algorithm to the RS, RMSSDT, DB, and SD indices. This comparison reveals that only S_Dbw was able to select the correct number of clusters in all situations. Close observation of the one synthetic data set reveals a potential grouping of either two or three. Because S_Dbw found very close indexes for both groupings, authors therefore suggest that S_Dbw may be modified to weigh the importance of variance or density higher than that of its counterpart. Despite the great promise of their index, the authors admit that S_Dbw does not work properly for concave geometry (like rings) and extremely curved geometry.

[JH09] Tong Jianhua and Tan Hongzhou. Clustering validity based on the improved s_dbw index. *Journal of Electronics (China)*, 26(2), February 2009.

Eight years after S_Dbw was introduced by Halkidi and Vazirgiannis, Tong Jianhua and Tan Hongzhou devise a modified version of S_Dbw, called S_Dbw$_{\text{new}}$, designed to improve the accuracy of the first index. The original S_Dbw index, referred to as S_Dbw*, finds the neighborhood of a cluster by utilizing a hyper-sphere. Jianhua and Hongzhou contend that this prevents S_Dbw* from properly indexing non-circular clusters. Instead, they suggest that the neighborhood of a given cluster should be measured by a margin point instead of a midpoint. Using a margin point allows for a more accurate level of inter-cluster density when two clusters have vastly different sizes and densities with a relatively small distance separating them. Additionally, S_Dbw$_{\text{new}}$ modifies the calculation of intra-cluster variance. Previously, the average scattering of all clusters was calculated while including the fraction $\frac{n-n_i}{n}$ where n represents the number of data points in a given data set and n_i represents the number of data points belonging to a given cluster. But two clusters that have the same variance will not be reported as such if one of their data set is larger than the other's ($\frac{n-n_i}{n}$ always increases as n increases). To prevent this from affecting results, the authors change the scatter function of S_Dbw so that its relation to size is flat. The authors then compare the validity of their new index to the validity of the old index by evaluating the indices given to the clusters found by the K-Means, DBSCAN and SOM clustering algorithms when they were given four data sets. S_Dbw$_{\text{new}}$ preformed perfectly— correctly identifying the number of clusters in all cases. But S_Dbw* failed to identify the correct number of clusters when the clusters were close together and when they possessed different shapes. In other cases, both indexes gave similar results. Concluding their paper, Halkidi and Vazirgiannis suggest the adaptation of S_Dbw* is important for when clusters are non-circular as well as when they differ in size and density.

[LLX+10] Yanchi Liu, Zhongmou Li, Hui Xiong, Xuedong Gao, and Junjie Wu. Understanding of internal clustering validation measures. In *2010 IEEE International Conference on Data Mining*, 2010.

Although many clustering validity indices have been proposed, little study has examined how the indices compare to one another. In order to learn about the strengths and weaknesses of different indices, Yanchi Liu, Zhongmou Li, Hui Xiong, Xuedong Gao and Junjie Wu analyze eleven different internal validation measures— as internal indices possess the ability to choose the best clustering algorithm, to select the optimal cluster number, and to use unclassified data. Discussing internal validation, the authors note that internal indices are generally based on two criteria: compactness and separation. Compactness is usually measured by variance, maximum pairwise distance, average pairwise distance or center-based distance, whereas separation is often measured with minimum pairwise distance and density. These measures are utilized by the indices under study, which consist of: the Root-mean-square standard deviation (RMSSTD), R-squared, the Modified Hubert statistic, the Calinski-Harabasz index, the I index, Dunn's index, the Silhouette index, the Davies-Bouldin index, the Xie-Beni index, the modified Davies-Bouldin index, the modified Xie-Beni index and the original S_Dbw index. Other internal validation methods exist, but they possess poor performance and or are designed to handle specific types of clustering. Consequently, they were excluded from analysis. To evaluate the effectiveness of the validation measures, the authors consider five data aspects: monotonicity, noise, density, subclusters and skewed distributions. Monotonicity is problematic in the case of RMSSTD, R-squared and the Modified Hubert statistic. All three of these functions continuously increase or decrease as the number of clusters grows. Finding the correct number of clusters is therefore looking for shift point in the curve of the given function. Unfortunately, settling on a shift point is difficult and subjective, lending little credence to these measures. Both

Dunn's index and the Calinski-Harabasz index are affected by noise. Given that Dunn's uses minimum pairwise distance, noise can dramatically change the both the inter-cluster separation value and the final value given for a set of input. And the Calinski-Harabasz index suffers from a proportional relation between increased noise and decreased final value. Only the I index is affected by density— as density decreases, it tends to merge clusters in order to minimize the sum of distances between points and their cluster center. When confronted with subclusters, Dunn's index, Silhouette index, Davies-Bouldin index, the SD validity index, and the Xie-Beni index could not give the correct number of clusters. In all cases, the value of inter-cluster separation was maximized when subclusters were considered to be one large cluster. Considering the last aspect of data set distribution, skewed data sets, only proved problematic for the Calinski-Harabasz index. Its equation shares the same basis as the K-means algorithm, which also fails to handle skew. Resultantly, the paper concludes S_Dbw is the only index to perform well in all considered data set aspects.

[Mac05] David J. C. MacKay. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, 4 edition, March 2005.

While discussing learning algorithms, David J. C. MacKay devotes a chapter of his book to understanding K-means. Before he delves into the specifics of the clustering algorithm, he emphasizes the importance of clustering given its predictive power, aid in communication (through information compression), the discovery of difficult-to-define objects and its ability to model learning processes in neural systems. Having justified his discussion, he proceeds to introduce K-means as an algorithm that puts data points (with any number of dimensions) into K clusters. It conducts this process by first initializing the K means, or the means of each cluster. Often, this is done randomly. Once initialized, the algorithm then iterates over a two-step process consisting of assignments and updates. During assignment, all data points in the data set are given to the nearest mean. While in the update step, the means are updated to the mean of the data points assigned to them. This process is repeated until there is no change in the mean's value. This is a guaranteed convergence. Unfortunately, the initial points selected by the algorithm greatly affect the resulting clusters. Further the algorithm suffers from several design flaws; K-means cannot account for the weight nor the breadth of a cluster, it cannot represent cluster size nor shape, and it traditionally can only assign a point to a single cluster. The last problem is easily fixable. Soft K-means changes the responsibility of a mean for a point from necessarily being a one or zero to acting as a fraction of one. Moreover, the algorithm switches assignment from being based on an absolute minimum value to a soft minimal value. Yet soft K-means cannot fix all of the issues associated with K-means and there remains much to improve.

[PKG14] Matthew Kyan Paisarn, Muneesawang Kambiz, and Jarrah Ling Guan. *Unsupervised learning*. John Wiley & Sons, Inc., 1 edition, May 2014.

Unsupervised Learning: A Dynamic Approach introduces its reader to several basic aspects of learning in chapter two, especially in regards to clustering. It defines clustering as the process of grouping similar or related entities. Following from this definition, it claims the ability of any algorithm to cluster is linked to the distance method it chooses to assess similarity. To be considered as a possible distance metric, an equation must satisfy symmetry, non-negativity, the identity of indiscernibles, and the triangle identity. Many metrics meet these criteria, including the following metrics: Euclidean, Euclidean squared, wrighted Euclidean, standardized Euclidean, Mahalanobis, Manhattan, Chebychev, Minkowski, Cosine, and Pearson correlation. Euclidean distance uses a hyper-sphere to find points of equal similarity, but many of its alternatives present a weighted hyper-ellipsoid, designed to allow for non-spherical clusters. Mahalanobis distance extends this idea. Manhattan distance uses the sum of orthogonal distances to find similar data points, a method which is modified by Chebychev distance and Minkowski distance. Cosine distances considers the angle between vectors. Pearson considers whether vectors have similar variations. If any of these distance metrics are selected, a clustering algorithm can then be described in terms of its cluster prototypes, its class membership function and its method of partitioning (hard or soft). From this information, algorithms have been found to fall into one of several approaches. One of the most popular approaches is using iterative mean-squared error. Best exemplified by K-Means, this approach initially defines K cluster prototypes arbitrarily. Then the prototypes are refined using nearest neighbor allocation until no significant change occurs to the prototypes. Unsurprisingly, this method is sensitive to initialization. A second approach to clustering exists and is called the mixture decomposition approach. This method assumes that the given data has a finite mixture of probability distributions and uses this assumption to find possible Gaussian mixtures. Considering a bottom-up approach, the book introduces a third clustering tactic: the agglomerative hierarchical approach. This approach starts every data point as its own cluster, then it merges the clusters using a proximity matrix. The hierarchy of connections can later be cut off at an appropriate value. Introduced fourth, the graph-theoredic approach treats data points as nodes on a graph connected by edges to all other points. The largest edges are then removed to minimize an energy function and identify clusters. There is also a fifth evolutionary approach that evolves a population to achieve an optimal clustering solution. Growing and producing new hyperboxes with self-adaptive parameters allows for the creation of possible clustering solutions. Finally, the book discusses its sixth and last approach: neural networks. Using neural networks allows one to identify nonlinear relations through the use of competitive learning or learning vector quantization. After exploring these approaches, the text highlights the importance of determining cluster validity through cost function-based indices, density-based indices, or geometric-based indices. Cost function-based indices attempt to minimize an error or heuristic and include indices such as the Figure of Merit, the Bayes Information Criterion, and the I-index. Density-based cluster validity indices capture the amount of overlap between clusters in a solution. Such methods include the Partition Coefficient, the Classification Entropy Index and the Xie Beni Index. Geometric-based indices consider the relationship between and within clusters. They attempt to maximize the separation of clusters and clusters' density. Examples of this basis include: Dunn's index, Calinski-Harabasz index and Geometric Index. The chapter concludes by reiterating the importance of indices.