JCSE International Journal of Computer Sciences and Engineering Open Access

Survey Paper Volume-5, Issue-5 E-ISSN: 2347-2693

Survey on A Connectivity and Density Dissimilarity Based Clustering

P. Khandelwal[1*], S. Saxena[2]

[1]Dept. of CSE, Rajasthan College of Engineering for Women (Rajasthan Technical University), Jaipur, India
[2]Dept. of CSE, Rajasthan College of Engineering for Women (Rajasthan Technical University), Jaipur, India

Corresponding Author: khandelwal262@gmail.com

Available online at: www.ijcseonline.org

Received: 11/Apr/2017, Revised: 23/Apr/2017, Accepted: 17/May/2017, Published: 30/May/2017

***Abstract*—** In an aeon where information is precious, all these data need to uncover the relations presented between a set of unlabeled dataset for the purposes of manifold - revise, explore, sort, store, analyze; arises all the more. A very feasible way to explore the relations between data is clustering, an unsupervised data mining technique. Clustering aims to group like data points together in clusters with no similarity between data points of different clusters and leaves behind outliers or points not belonging to any of the clusters. Clustering can be applied to all types of data with varying nature (numeric, categorical, mixed), and dimensions (low, high), however, methodologies and similarity measures that can be applied may vary accordingly. In this manuscript we will discuss about various technologies used for clustering of data like role of distance metrics in clustering, clustering using ensembles and dimensionality reduction/minimaization techniques for modeling complex data relations.

***Keywords*—**Clustering,DistanceMetricStyling,Ensembling,LargeDimensions.

I. INTRODUCTION

Clustering aims to group resembling data points together in clusters with no similarity between data points of different clusters and leaves behind outliers or points not belonging to any of the clusters. Clustering has always been an active area of research and has been applied to thousands of applications so far. A typical clustering algorithm can be separated into three simple steps.

1. Deducing similarity of a data point from the remaining data points under analysis using an appropriate similarity/dissimilarity measure.

2. Grouping the objects on the basis of the deduced similarity.

3. Evaluating the accuracy of the clustering decision obtained.

Clustering can be applied to all types of data with altering nature and dimensions, however, methodologies and similarity measures that can be applied may vary accordingly. The clustering algorithms can be separated into two parts first

- *Partitioning Approaches*-decomposing the input dataset into a set of disjoint clusters in a single iteration.
- *Hierarchical Approaches*- Cluster analysis using a hierarchical approach builds a hierarchical tree. Clusters are merged on the basis of their proximity or close to one another. This approach can further be subdivided into
    - ○ *Agglomerative or Bottom-up approach*- Each observation in this case starts in its cluster with pairs of clusters merging with one another as the one goes up the hierarchy.
    - ○ *Divisive or Top-Down*- The observations in this case too start with one cluster followed by performing splitting recursively while moving down the hierarchy.

A.Role of Distance Metric in Clustering

Distance Metrics have a major role to play in any clustering algorithm. While the primary motive is to deduce similarity between data points in terms of distance and different metrics are proposed for the same, the metrics highly influence the shape of cluster formation.

B.Ensembling

There may be conditions when a single clustering algorithm is not fit enough for the application addressed. Ensembling of different clustering algorithms to get as product a single desired clustering decision is very common. Not only does it obscure the limitations of one algorithm, the strengths of the different participating algorithms help in getting desired results.

C. *Handling large dimensions*

One of the major concerns in data analysis is absence of well-developed techniques or algorithms for representing contact representations of high dimensional data. The 'curse of dimensionality' issue is observed in many analysis works. According to the curse, data in only one dimension is relatively filled and adding a dimension involves stretching which decrease the density dramatically and make the distance measure go meaningless. As a result, various dimensionality reduction techniques have been proposed for clustering high dimensional data.

Based on the different techniques, the survey is divided into three disciplines according to different purposes in the mentioned proposal - clustering using ensembles, proposals involving different distance metrics for clustering and dimensionality reduction techniques for modelling of complex data relations

II. EFFECTIVE DISTANCE METRICS

Hinneburg and Klein[1] proposed the DENCLUE (DENsity using CLUstEring) algorithm based on the concept of influence of each data point in the clustering process. Pekalska et al[16] proposed classification in two dissimilarity based representation spaces. The objective behind their work is to handle pair wise proximate data, which is usually done via kernel or NN rule. Baya and Granitto[11] proposed a Penalized k-Nearest-Neighbor-Graph (PKNNG) distance metric for clustering micro-array datasets for applicability in gene expression data analysis. The basis of the distance metric is a 2-step procedure. The first step of the procedure involves construction of a k-Nearest Neighbor Graph keeping the value of k low. Edges penalized with weight are added to the graph in the second step of the procedure. This connects the components of the neighbor graph. The geodesic distance between all pairs of vertices in the graph is then calculated by use of Dijkstra's algorithm. Rodriguez and Liao[2] proposed the Delta Density Clustering (DDC) algorithm. Chen and He [10] define intensity as a measure of density of data points. Precisely, intensity gives a numerical value of indicating how a data point is placed in a field of its relation to other points in the same field. Intensity defined is based on the concepts of theory analysis and data field intensity theory. Using the defined intensity, clustering of mixed streaming data is done, mainly estimation of cluster centers. Cluster centers are points at a relatively large distance from objects having field intensity higher than the point. Mixed data is handled by categorizing the dataset on the basis of dominance of the attributes in the dataset into three types: Categorical dominant, Numeric dominant and balanced. Depending on the dominance, corresponding distance metrics are used to deduce similarity between data points.The Penalized K-Nearest Neighbor Graph (PKNNG) distance metric by Baya and Granitto [11] relies on connectivity between components of a k-Nearest Neighbor graph with components of the neighbor graphs. In situations when connectivity is more or complicated, the PKNNG metric cannot work well, a solution was proposed by Baya et al [17]. Baya et al's work aims at improving the limitations of PKNNG metric.

III. CLUSTERING USING ENSEMBLES

Use Cluster ensemble methods can be transformed into graph/hypergraph partitioning problem. Strehl and Ghosh [8] proposed a graph/hypergraph based method, considers a partition sharing most of the information with all the required partitions as the consensus partition. Measure of the amount of information shared by two partitions is done through Normalized Mutual Information (NMI). The next set of cluster ensemble methods are the relabeling and voting based methods. The first step of generation is the labeling correspondence problem and the final result consists of the consensus partition through a voting process. Fischer and Buhmann[9] proposed a similar method based on plurality voting for consensus partition and solution to the labeling correspondence problem is provided through maximum likelihood problem by use of Hungarian method. Fred and Jain [15] proposed the idea of Evidence Accumulation by multiple Clustering (EAC) for the purpose of combining multiple clustering results. For this, the first step is to produce a set of object partitions via use of different clustering algorithms; use of a single clustering algorithm with varying parameters or initialization methods or combination of clustering algorithms with different feature spaces. The concept of EAC views each partition as independent entity with combination of these partitions on the basis of a voting mechanism, for the purpose of generating a similarity matrix between the patterns. On the final partition is performed clustering using a hierarchical agglomerative clustering algorithm. The algorithm has basis from the split and merge technique of the k-means algorithm. Baya et al's work [11] merges the concepts of density with connectivity. Information aggregation is done using EAC similarity which though possess an over-fragmentation tendency, gives a great insight related to density among neighbors. The dissimilarity has discriminative properties that emphasize on having some dissimilarity between two non-neighboring samples so as to simplify the clustering process. The dissimilarity function is built using three blocks: two blocks for measuring density and the third for connectivity. Density is measured by EAC using a method based on the ensemble of k-means with HC-Ward algorithm or based on ensembles of k-NN algorithm. Third block measures connectivity using the PKNNG distance. At last, merging with MDS (Multi Dimensional Scaling) is done in order to find a simpler representation of the input data. Dimensionality reduction technique is imposed to find a low dimensional representation of data. All the four concepts of density, connectivity, scaling and dimensionality reduction help in obtaining clusters of desired form.

IV. MODELING COMPLEX DATA RELATIONS

A.Dimensionality Reduction

Cox and Cox [12] proposed Multi-Dimensional Scaling (MDS) as a means to visualize similarity levels of individual data points in a dataset. The usefulness of MDS arises in situations with unknown relationships between objects but possible estimation of the distance matrix.Tenenbaum et al [4] pointed out the limitations of PCA and MDS in treating non-linear structures in high dimensional data. The actual degrees of freedom or dimensions are observed to not be detected by the two methods. With various advantages of PCA and MDS like good efficiency, global optimality and convergence guarantee asymptotically to the desired and ground truth structure, the authors attempt to combine the flexibility of learning the non-linear degrees of freedom of the datasets. Using local metric information, the global geometry of the dataset can be effectively learned by the proposed 'Isomap' algorithm. Geodesic distances between data points, instead of the Euclidean distance used by PCA and MDS, are used to capture the intrinsic geometry of a dataset. Roweis and Saul [14] proposed the Locally Linear Embedding (LLE) clustering algorithm, able to construct low dimensional mappings of data of high-dimensionality. Belkin and Niyogi [7], inspired from the works by Tenenbaum et al[4] and Roweis and Saul[14], proposed a simple algorithm involving very few computations and sparse eigen value problem.

V. CONCLUSION

In this paper we analyzed different methods suggested by various researchers for density, connectivity, scaling and dimensionality reduction of complex data, apart from large volume, data exists in variety of forms and distributions. Hence, simple approaches like clustering also need to be adapted for such changes. Therefore we need to combine various density and connectivity based traditional and popular clustering methods with specialized techniques to give interesting clustering outputs which are of much use in some real life applications.

REFERENCES

[1] A. Hinneburg. and D. Keim, "*An efficient approach to clustering large multimedia databases with noise*", Proceedings of the 4th ACM SIGKDD Conference, US, pp. 58-65, 1998.

[2] A. Rodriguez, A. Laio, "*Clustering by fast search and find of density peaks*", Science, Vol. 344, Issue.6191, pp. 1492-1496, 2014.

[3] J. Y. Chen, H.H. He, "*A fast density-based data stream clustering algorithm with cluster centers self-determined for mixed data*", Information Sciences, Vol. 345, Issue.6, pp. 271-293, 2016.

[4] A. E. Bayá, P. M. Granitto, "*Clustering gene expression data with a penalized graph-based metric*", BMC Bioinformatics, Vol. 12, Issue.1,pp.10-21, 2011.

[5] G. Hinton, S. Roweis, "*Stochastic neighbor embedding*", Proceedings of International Conference on Advances in Neural Information Processing Systems, US, pp. 833-840, 2003.

[6] M. Belkin, P. Niyogi, "*Laplacian eigenmaps for dimensionality reduction and data representation*", Neural Computation, Vol.15, No. 6, pp. 1373-1396, 2003.

[7] A. Strehl, J. Ghosh, "*Cluster ensembles: A knowledge reuse framework for combining multiple partitions*", Journal of Machine Learning Research, vol. 3, Issue.5, pp. 583-617, 2002.

[8] B. Fischer, J. Buhmann, "*Bagging for path-based clustering*", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 25, no. 11, pp.1411-1415, 2003.

[9] S. Vega-Pons, J. Ruiz-Shulcloper, "*Clustering ensemble method for heterogeneous partitions*", Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications, Barlin, pp.481-488 2009.

[10] A. E. Bayá, M. G. Larese, P. M. Granitto, "*Clustering using PK-D: A connectivity and density dissimilarity*", Expert Systems with Applications, Vol. 51, Issue.1, pp. 151-160, 2016.

[11] T. F. Cox, M. A. A Cox, *"Multidimensional scaling (2nd ed.)"*, Chapman & Hall/CRC, USA, pp.1-220, 2000.

[12] S. W. Kim and R. Duin, "*An empirical comparison of kernel-based and dissimilarity-based feature spaces*", InJoint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR), Berlin, pp. 559-568, 2010.

[13] S. Roweis, L. Saul, "*Nonlinear dimensionality reduction by locally linear embedding*", Science, vol.290, no.5500, pp.2323-2326.

[14] A.Fred, A.K. Jain, "*Combining multiple clusterings using evidence accumulation"*, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 27, no. 6, 2005, pp. 835-850.

[15] E. Pekalska, P. Paclik, R.P.W. Duin, "*A generalized kernel approach to dissimilarity-based classification*", Journal of Machine Learning Research, vol. 2, Issue.1, pp. 175-211, 2002.

[16] E. Pekalska, R. Duin, "*Beyond traditional kernels: Classification in two dissimilarity-based representation spaces*", IEEE Transactions on Systems, Man and Cybernetics Part C : Applications and Reviews, vol. 38, no. 6, pp. 729-744, 2008.