Using Categorical Attributes for Clustering

Avli Saxena, Manoj Singh
Gurukul Institute of Engineering and Technology, Kota (Rajasthan), India

Abstract

The traditional clustering algorithms focused on clustering numeric data by exploiting the inherent geometric properties of the dataset for calculating distance functions between the points to be clustered. The distance based approach did not fit into clustering real life data containing categorical values. The focus of research then shifted to clustering such data and various categorical clustering algorithms are proposed till date. The clustering of categorical data turns complex because of the absence of a natural order on the individual domains, high dimensionality of data and the existence of subspace clusters in the categorical datasets. This survey focuses on the shortcomings of categorical data and the recent developments in the direction of using data with categorical attributes for clustering

Keywords: *Data Analysis, Clustering, Categorical Data, ROCK.*

1. Introduction

Clustering is an unsupervised form of learning in data mining with Classification as the supervised learning approach. The process of clustering starts by taking as input a dataset and grouping the similar data points in clusters until all the data points are grouped. The similarity between data points is calculated through a similarity/distance measure. The very first of the proposed clustering algorithms concentrated on clustering numeric data through the use of derived ideas from statistics and geometry. With changing requirements and time, it was observed that the real life data contains categorical values and not numeric values and hence limited the scope of the existing clustering algorithms to numeric data only. Categorical data is different from numeric data in the sense that it groups the data into categories and not any numeric values. For example, a set of {male, female, children} can be used to categorize a group of people. This set cannot be clustered based on the distance between the people present. Hence, no distance based conventional clustering approaches were found useful in this direction opening further a new area of research. The initial proposals first converted the categorical data into corresponding numeric data followed by clustering this data according to the traditional clustering approach of distance. This approach was proposed seeing the ease in computations involving operations on numeric data. The earlier notions of statistics and geometry could not be applied to categorical data due to some limitations of the categorical data. With time, the researchers proposed clustering methods that can directly be applied to categorical data [1,2,3,4,5,6]. This paper provides a brief overview of some of the classic categorical data clustering methods and the recent trends in the same.

2. Limitations of Categorical Data

The data containing categorical attributes pose a number of challenges on the existing clustering methods due to the following reasons.

- ***No natural order***
  The traditional similarity measures are based on the co-occurrence of attribute values. Some others like the Jaccard coefficient and Cosine similarity can even define similarity seeing whether two attribute values occur together for any data point or not. If the attributes are not naturally ordered like in categorical data, similarity between data points cannot be measured through the existing measures.

- ***High Dimensionality***
  The categorical datasets have high dimensionality. The traditional clustering approaches fail to work on high dimensional data because of the curse of dimensionality phenomenon .

- ***Existence of subspace clusters***
  Categorical data, being high dimensional, fail to cluster data in all dimensions and are limited to a certain number of dimensions.

- ***Conversion of categorical to numeric data***
  The only possible approach initially for clustering categorical data was to convert it into equivalent numeric form. However, the converted values are arbitrary and seem of no use beyond using it as a convenient label of a particular value. The reason behind the same is that each value in a categorical attribute represents a logical separate concept and therefore can neither be meaningfully ordered nor can be manipulated the way numbers could be.

3. Classic Categorical Clustering Algorithms

This section lists the basic categorical clustering algorithms on which almost all the related research works are dependent. Dependence is in terms of the basic ideology or the methodology of these algorithms applicable even now in clustering categorical data. The recent developments however are either extensions to the traditional algorithms or are aimed to work on improving their efficiency. These include PAM [1], CLARA [1], BIRCH [2], K-modes [3], STIRR [4], CACTUS [5] and ROCK [6] arranged chronologically.

3.1 PAM (1990)

PAM (Partition Around Medoids) by [1], first finds medoids, with medoids being a sequence of objects centrally located in objects. These medoids are then places in a set of selected objects, say S. Taking set S and a set O containing all the objects, difference of objects of set S from the objects of set O gives us the set U, a set of unselected objects. The objective behind the proposal was to minimize the average dissimilarity of objects to their closest selected object. Likewise, the sum of dissimilarities between object and their closest selected object is minimized. After finding the set of selected and unselected objects among the set of all objects in a dataset, the second phase swaps the selected objects with unselected objects for improving the quality of the selected objects and hence, clustering.

3.2 CLARA (1990)

CLARA (Clustering LARge Applications) by Kauffman and Rousseeuw in 1990 [1] is based on the PAM clustering algorithm. The procedure of the algorithm starts with taking multiple samples of the input dataset and applying the PAM clustering on each sample for finding medoids. If the taken samples are representative, then the medoids of the sample can be used to approximate the medoids of the entire dataset. The only advantage of the algorithm over PAM is that it is able to handle large datasets unlike PAM. However, the efficiency of the algorithm is dependent on the sample size. Another limitation is the fact that though breaking a dataset into samples gives improved results, the same cannot be the case when applying the same clustering approach to whole large datasets provided the samples taken are biased. The accuracy of the clusters formed is measured through average dissimilarity of all the objects in the entire dataset.

3.3 BIRCH (1996)

Clustering in a multi dimensional dataset with minimized I/O costs were the two problems addressed by Zhang et al who then proposed the BIRCH algorithm (Balanced Iterative Reducing and Clustering)[2]. The proposed algorithm is an incremental and dynamical approach that takes as input multi-dimensional data points and produces good quality clusters with the available resources as output. It typically requires a single scan to cluster the given data points and a few additional scans for improving the quality of clustering. Another advantage of the proposed algorithm is that it is effective in handling noise or outliers, i.e. the points not of the underlying pattern and is the first algorithm in this direction. The I/O cost of the algorithm is found linear to the size of the dataset. The ability of the algorithm to handle large databases is by the use of a compact summary and the concentration only on the densely occupied patterns. The similarity between data points is found by using measurements which are stored and updated incrementally in a height balanced tree. BIRCH's efficiency to work with any given memory of input data and its linear time complexity contribute in the prevailing success of the algorithm.

3.4 K-Modes (1998)

The K-means clustering algorithm [7] has been one of the most popular clustering algorithms proposed till date. Its simplicity, easy implementation and scalability sums up the reasons for the popularity of this algorithm even after years of its proposal. However, it worked on numeric data and failed to cluster high dimensional categorical data. A modification to the k-means algorithm for clustering categorical data, K-Modes algorithm [3] was then proposed. The three major modifications in the algorithm included changes in the distance function, representation of the centroids and the iterative clustering process. The Euclidean distance metric was replaces by a simple matching dissimilarity measure, the calculation of mean for representing centroids was replaced by modes because the categorical data can neither be calculated through mean or medians. The iterative process was modified by considering only the most frequent categorical values and updating their modes in each iteration of the clustering process. The k-modes algorithm was found effective for categorical data clustering while retaining the efficiency of the k-means algorithm and guarantees the convergence of the clustering process to a local minimal result.

3.5 STIRR (1999)

STIRR (Sieving Through Iterated Relational Reinforcement) [4] is an iterative algorithm for clustering categorical data by the use of dynamical systems. The algorithm initiates by taking as input a database represented as a graph with each distinct value in the domain of each attribute represented through a weighted node. Therefore, for N attributes having the domain size of i^{th} attribute as d_i, then the number of nodes in the graph is $\sum_i d_i$ .An edge in each tuple represents a set of nodes participating in that tuple. Thus, representation of a tuple is done as a collection of nodes one from each attribute type. The proposed structure is a set of weights of all the nodes. The initial weights of nodes are assigned either uniformly or randomly. The algorithm runs iteratively and with each iteration updates weight of any node by a combiner function. The role of this combiner function is to combine the weights of other nodes participating in any tuple with the given node for which the weight is to be updated. The algorithm continues in the same way till a stable point, Basin is reached. The convergence of the algorithm depends on the combiner function.

3.6 ROCK (1999)

Guha et al proposed a new notion of links to cluster a group of objects in the ROCK [6] algorithm. It is a bottom up hierarchical clustering approach for categorical and Boolean attributes. Similarity between two objects is deduced through any distance metric or non metric similarity function. Two objects are considered neighbors if the similarity between them reaches a certain set threshold. Links can then be calculated by calculating the number of common neighbors for the points. The number of links between two data points hence corresponds to the number of common neighbors between the same points. After the links are computed for points and taking each point as a single cluster, the algorithm merges clusters using a goodness measure. The algorithm computes the same way till the required number of clusters is formed or till no links remain. The other different approach in the algorithm is that it does not involve a complete dataset for clustering, rather a randomly drawn sample is taken and clustering is performed. Links are employed to the sampled data points. Finally, the clusters involving only the sampled points are used to assign the remaining data points on disk to the appropriate clusters.

4. Recent Developments

Khandelwal and Sharma [8] proposed a fast categorical clustering algorithm based on statistical distance calculations for similarity as done in the numeric clustering algorithms. For this, the categorical values are converted into equivalent numeric values as a pre-processing step. The proposed pre-processing step indicates that two data points are observed more similar if their categorical values hold equal prominence. The next step is the clustering process with the basic idea of separately capturing the variation of each dimension of the dataset. A dimension summary can then be used to assign a cluster label to any data on the summarized dimension. The algorithm takes firstly only one dimension and initializes cluster accordingly. The next step involves taking two dimensions and similar initialization. Similarly, the next step takes three dimensions. The distance calculations are therefore different in each step with updation of the cluster labels accordingly. The proposed clustering approach is therefore light, scalable and fast.

Sharma and Gaud [9] proposed modifications in the classic K-modes algorithm [3]. Instead of taking the attributes on frequency basis as done in k-modes, the selection of attributes in the proposed algorithm is done on the basis of information gain with better accuracy results. The process starts by reading the dataset and defining the required number of clusters. Information gain is calculated for every attribute of the dataset. The attributes having higher information gain are grouped into a common attribute. Agglomerative hierarchical clustering is then applied to the filtered dataset and initial centroids are calculated. K-modes clustering approach is then applied to the initially calculated centroids and labels of each attribute are calculated both by the k-modes and the proposed algorithm. Comparison of the labels proves the accuracy of the proposed algorithm.

The K-modes algorithm [3], though efficient, results in non repeatable clustering because of its criterion of randomly choosing the initial cluster centers for invoking every new execution may lead to non-repeatable clustering results. Ahmad and Khan [10] addressed this initialization problem of K-modes algorithm by proposing a cluster center initialization algorithm. The proposed algorithm performs multiple clustering of the data based on attribute values in different attributes and the calculated deterministic modes are then used as initial cluster centres.

A Categorical dataset has multiple attribute values. The similarity between data points can be calculated through calculating common data points, common attribute values or as an association between the two. However, the entire dataset and attributes are subjected to uncertainties which make almost all categorical clustering algorithms ineffective in their purpose of clustering categorical data when uncertainty arises thereby increasing complexity. The next common problem is stability problems related to multiple executions of the algorithms. Owing to these limitations, Hassanein and Elmelegy [11]proposed two algorithms based on the rough set theory taking into account the significance of attributes in information systems and dependence of attributes in the dataset namely Standard Deviation of Standard Deviation Significance (SSDS) and Standard Deviation of Standard Deviation Dependence (SSDD). SSDD was found to have achieved the highest purity among the predecessor algorithms based on the rough set theory. Also the proposed techniques worked fine on large datasets.

Cheung and Jia [12] observed that there exists an awkward gap between the similarity metrics used for numeric data clustering and categorical data clustering. The proposed work bridges this gap by a unified distance metric that can be effectively used for numeric, categorical or mixed data. This metric has a uniform criterion for the object-cluster similarity for numeric and categorical attributes and eliminates the need of transformation and adjustment of parameters. Authors have also proposed a clustering framework based on the concept of object-cluster similarity. Based on the proposed framework, an iterative algorithm is also proposed. A common problem in almost all clustering approaches is the determination of the required number of clusters. The authors address this problem by embedding competition and penalization mechanisms in the proposed framework that eliminate the redundant clusters and determine the number of clusters effectively.

Goswami and Mohanta [13] have proposed a clustering approach based on the work by San et al [14] and using the distance metric used in Dutta and Mohanta’s work for categorical data clustering [15]. San et al proposed an extension to the k-means algorithm[7] for clustering categorical data taking cluster representatives as sets with the elements of each set representing a category or value of an attribute. Goswami and Mohanta modify the work of San et al by first representing categorical data into numeric data for easy handling. San et al’ scheme [14] was applicable to cluster representatives and not individual data points. The proposed scheme provides a uniform representation for both cluster representatives and data points. The similarity measure of [15] is then used for calculating the similarity between a cluster representative and a data point or between two data points. The similarity measure is based on the notion of fuzzy sets and is a normalized one taking values of 0 and 1 inclusive.

Seman et al in 2012 proposed a k-Approximate Model Haplotype (k-AMH) for clustering a different type of categorical data known as Y-Short Tandem Repeat (Y-STR) categorical data [16]. Such data is composed of similar and almost similar objects in inter or intra classes and is unlike the data found in categorical datasets Mushroom, Voting etc. There were two problems associated with clustering such data. The first problem was that the obtained centroids were not found unique thereby resulting in empty clusters. The second problem is the inability of the obtained centroids in representing clusters further leading to a local minima problem. The mode based approach as in K-modes and the centroid based approach in the k-means algorithm failed to efficiently cluster this data, a solution to which was the medoid-based method proposed by Sanet et al in 2013 [17] used with the k-AMH algorithm of Sanet et al (2012)[16]. The proposed clustering approach was found to have outperformed results when compared against the previous k-based clustering approaches.

Ibrahim and Harbi proposed a Modified PAM (M-PAM) clustering algorithm [18] for mobile network planning. The problem encountered in effective mobile network planning is the decision regarding the optimal placement of base stations (BS) for achieving best services provided the incurred cost in the operation is low. This decision increases the complexity of the task and requires vast computational resources. The authors therefore introduce spatial clustering as a solution to the mentioned problem and modify the PAM clustering algorithm [1]. The original algorithm starts by specifying the number of clusters, k, followed by searching for the best locations of BS. The modified algorithm determines k through the radio network planning. Calculation of capacity and coverage is done and checked whether both satisfy the mobile requirements, otherwise k is gradually increased and both the algorithms are reapplied.

5. Conclusion

The focus of research in clustering data has moved from numeric data to categorical data because almost all real data is categorical. Clustering categorical data is a bit difficult than clustering numeric data because of the absence of any natural order, high dimensionality and existence of subspace clustering. One approach for easy handling of data is by converting it into an equivalent numeric form but that have their own limitations. Over the years, various classic clustering algorithms have been proposed and are still used. The new developments in this direction are either improvements or extensions of the old algorithms. This survey discusses the limitations in the categorical data clustering. Based on the uniqueness and difference of the categorical data from the numeric data, the survey is then partitioned into the classic categorical clustering approaches proposed years ago but used till date and the recent developments in this direction.

References

[1] L. Kaufman and P. Rousseeuw, “Finding Groups in Data: An Introduction to Cluster Analysis”, John Wiley and Sons, New York, NY, 1990.

[2] T. Zhang, R. Ramakrishnan, and M. Livny, “ BIRCH: an efficient data clustering method for very large databases”, Proceedings of the ACM SIGMOD Conference, 103-114, Montreal, Canada, 1996.

[3] Z. Huang, “Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values”, Data Mining and Knowledge Discovery, Vol. 2, No. 3, pp. 283 – 304, 1998.

[4] D. Gibson, J. Kleinberg and P. Raghavan, “Clustering categorical data: An approach based on dynamic systems”, Proceedings of the 24th International Conference on Very Large Databases, 311-323, New York, NY, 1998.

[5] V. Ganti, J. Gehrke and R. Ramakrishnan, “CACTUS-Clustering Categorical Data Using Summaries”, Proceedings of the 5th ACM SIGKDD, 73-83, San Diego, CA, 1999.

[6] S. Guha, R. Rastogi and K. Shim, “ROCK: A robust clustering algorithm for categorical attributes” Proceedings of the 15th ICDE, 512-521, Sydney, Australia, 1999.

[7] E.W.Forgy, “Cluster analysis of multivariate data: efficiency v/s interpretability of classifications”, Biometrics, 21, 768–769, 1965

[8] G. Khandelwal and R. Sharma, “A Simple Yet Fast Clustering Approach for Categorical Data”, International Journal of Computer Applications (0975 – 8887), Volume 120 – No.17, pp. 25-30, June 2015

[9] N. Sharma and N. Gaud, “K-modes Clustering Algorithm for Categorical Data”, International Journal

of Computer Applications (0975 – 8887), Volume 127 – No.17, pp. 1-6, October 2015

[10] S. S. Khan and A. Ahmad, "Cluster center initialization algorithm for K-modes clustering", Expert Systems with Applications, Vol. 40 pp. 7444–7456, 2013.

[11] W. A. Hassanein and A. A. Elmelegy, "Clustering algorithms for Categorical data using concepts of Significance and dependence of Attributes", European Scientific Journal, vol.10, No.3, pp 381-400, January 2014.

[12] Y.M. Cheung and H. Jia, "A Unified Metric for Categorical and Numerical Attributes in Data Clustering", Advances in Knowledge Discovery and Data Mining, Volume 7819 of the series Lecture Notes in Computer Science, Proceedings of the 17th Pacific-Asia Conference, PAKDD 2013, Gold Coast, Australia, April 14-17, 2013, Part II pp 135-146.

[13] O. M. San, V. N. Huynh, Y. Nakamori, "An Alternative Extension of the K-Means Algorithm For Clustering Categorical Data", International Journal of Applied Mathematics and Computer. Science, Vol. 14, No.2, 241-247, 2004.

[14] M. Dutta, A. K. Mahanta, "A Fast Summary Based Algorithm for Clustering Large Categorical Databases", Proceedings of ICWES12, Ottawa, Canada, 2012..

[15] J. P. Goswami and A. K. Mahanta, "Categorical Data Clustering based on an Alternative Data Representation Technique", International Journal of Computer Applications (0975 – 8887), Volume 72– No.5, May 2013.

[16] A. Seman, Z.A. Bakar and M.N. Isa, "An efficient clustering algorithm for partitioning Y-Short Tandem Repeats data", BMC Research Notes, Volume 5, 2012.

[17] A. Seman, Z. A. Bakar, A. M. Sapawi and I. R. Othman, "A Medoid-based Method for Clustering Categorical Data", Journal of Artificial Intelligence, Volume 6, Issue 4, pp. 257-265, 2013.

[18] L. F. Ibrahim, M. H. A. Harbi, "Using Modified Partitioning Around Medoids Clustering Technique in Mobile Network Planning", International Journal of Computer Science Issues Volume 9, Issue 6, pp. 1-10, November 2012.