An Accurate Revelation of the Similarity between Clusters

A.Veera Mahendra
M.Tech (II CSE), Dept of CSE, Madanapalle
Institute Of Technology & Science,
Madanapalle, A.P, India

S.M.Farooq
Assistant Professor, Dept of CSE, Madanapalle
Institute Of Technology & Science,
Madanapalle, A.P, India

ABSTRACT

The structure of the data set playing a vital role in datamining. In concept of datamining information recovery and pattern identification nothing but data clustering. There are multiple clustering algorithms have been commenced to clustering categorical data. Unfortunately these algorithms created an incomplete information. In recent times cluster ensembles have come out as an essential solution to overcome these limitations and to get the excellence results for clustering. A Link-Based similarity measure is proposed to guess unknown values from a link network of clusters and bridges the gap among the task of data clustering and that link examination. It also improves the ability of ensemble methodology for categorical data. A new Link-Based cluster ensemble approach is commenced which is well-organized than the previous model, where a binary cluster association matrix, like matrix is used to create the cluster ensembles. These cluster ensembles have impurity information, to overcome these problem Link-Based similarity algorithm is used to generate an accurate pure clusters.

Keywords

Data clustering, Cluster ensemble, Link-based similarity measure, Data sets.

1. INTRODUCTION

The elementary means for considering the structure of a data set playing a vital foundational function in mining of data, information recovery, and pattern identification is data clustering [1]. Many entrenched clustering algorithms have been planned for numerical data, whose intrinsic properties can be obviously engaged to calculate a distance connecting feature vectors. Classifying data into groups or clusters such that the data in the similar cluster are more comparable to each other than to those in different clusters is the main intention of clustering shown in fig1[3]. In recent years, by means of applications to interesting domains such as protein interaction data many categorical data clustering algorithms have been introduced. By making use of Gower's similarity coefficient the initial method was developed. The conventional k-means with a simple matching dissimilarity measure and a frequency-based method to update centroids was extended by the k-modes algorithm. By making use of a pre specified similarity threshold to determine which of the existing clusters to which a data point under examination is assigned through a single-pass algorithm [2]. A hierarchical clustering algorithm that makes usage of the Information Bottleneck framework to define a distance measure for categorical tuples was proposed.

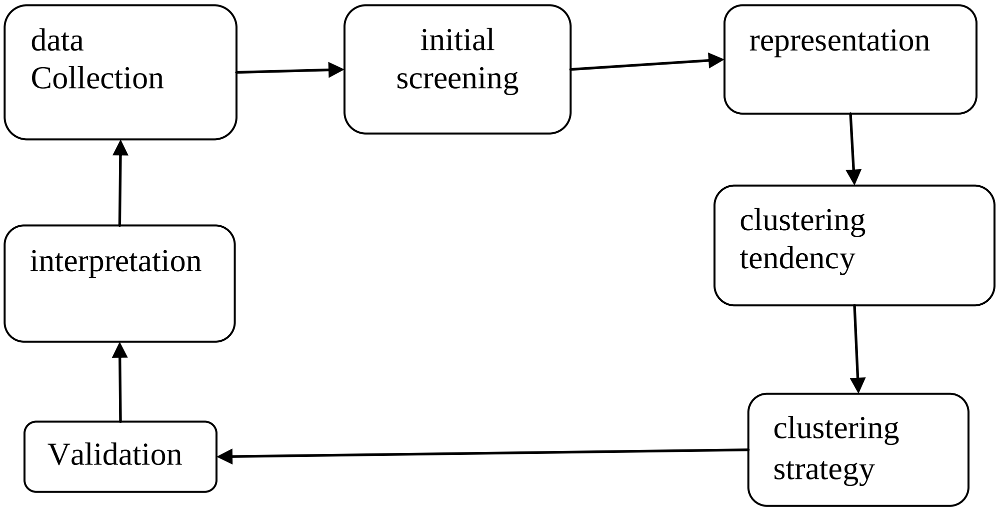

Fig1: An overview of clustering methodology.

By means of a partitioning method for categorical data, the concepts of evolutionary computing and genetic algorithm have also been adopted. Even though, a large number of algorithms have been commenced for clustering categorical data, the No Free Lunch theorem recommends there is no single clustering algorithm that carry out most excellent for all data sets and can find out all types of cluster shapes and structures obtainable in data. Every algorithm has its own potency and weaknesses. For particular data set, different algorithms have different parameters, usually provide distinct solutions. So it is very difficult for users to decide which is the best algorithm will be the best alternative for a given data set. To overcome these problems cluster ensembles have emerged and improve the robustness as well as the quality of clustering results. The main concept of cluster ensemble is to combine different clusters to achieve a final cluster result nothing but pure cluster.

2. GETTING PURE CLUSTER THROUGH LINK-BASED SIMILARITY ALGORITHM

There are multiple algorithms have been introduced to clustering the categorical data,but these algorithms didn't form the pure clusters. In recent times cluster ensemble have come out as best solution to get the pure cluster[4]. Cluster ensemble can form the group clusters into single site. These cluster ensembles have impurity information, so clusters are very close to each other as shown in figure 2. By using Link-based similarity algorithm, we can form the link between two similar cluster. By applying a spectral graph partitioning method on these clusters we can form the pure accurate clusters. The process of getting pure clusters is explained by following steps.

2.1 Choose database:

In this module we deploy the dataset (categorical data) items into our application and displayed in given format. In this application we can uploading the data base, this is one way to Conway my data base to this application. In the second part of this application. we can get the data set from the data base also this is the second way Conway to my application .

2.2 Creation of Cluster Ensemble:

The initial type of cluster ensemble transforms the difficulty of categorical data clustering to cluster ensembles by considering every categorical attribute value as a cluster in an ensemble. Consider S = {a1;...;aN}be a set of N data points, P = {q1;...;qN} be a set of categorical attributes, and π = {π1;...;πM} be a set of M partitions. Each partition πi is generated for a specific categorical attribute qi belongs to P. Categorical data S can be directly transformed to a cluster ensemble II, devoid of actually implementing any base clustering. Though single-attribute data partitions may not be as precise as those obtained from the clustering of all data attributes, they can convey about immense assortment within an ensemble. In addition to its competence, this ensemble generation method has the potential to lead to a high-quality clustering consequence.

2.3 Similarity between clusters:

In this module the dataset items are processed by using link-based algorithm and also we can see how much similarity between clusters based on percentages given a cluster ensemble _ of a set of data points X, where x is G =(V ;W) I can be constructed, where V is the set of vertices each representing a cluster and W is a set of weighted edges between clusters. Shared neighbours have been widely recognized as the basic evidence to justify the similarity among vertices in a link.

2.4. View refined matrix:

In this module generating a refined cluster-association matrix (RM) using a link-based similarity algorithm. Several cluster ensemble methods, both for numerical and categorical data are based on the binary cluster-association matrix . Each entry in this matrix BM(xi,cl) € {0,1} represents a crisp association degree between data point xi €X and cluster cl where c2 represents Cluster. It is efficient than the preceding representation where a binary cluster-association matrix-like matrix is used to match up the ensemble information. The spotlight has shifted from enlightening the resemblance between data points to estimating those among clusters. A new link-based algorithm has been specifically planned to generate such measures in an accurate, reasonably priced manner[5]. The link-based cluster ensemble methodology includes three major steps of: creating base clustering to form a cluster ensemble producing a refined cluster-association matrix by means of a link-based similarity algorithm, and producing the concluding data partition by making use of the spectral graph partitioning method as a consensus utility.

The graphical structure of getting pure cluster is as follows.

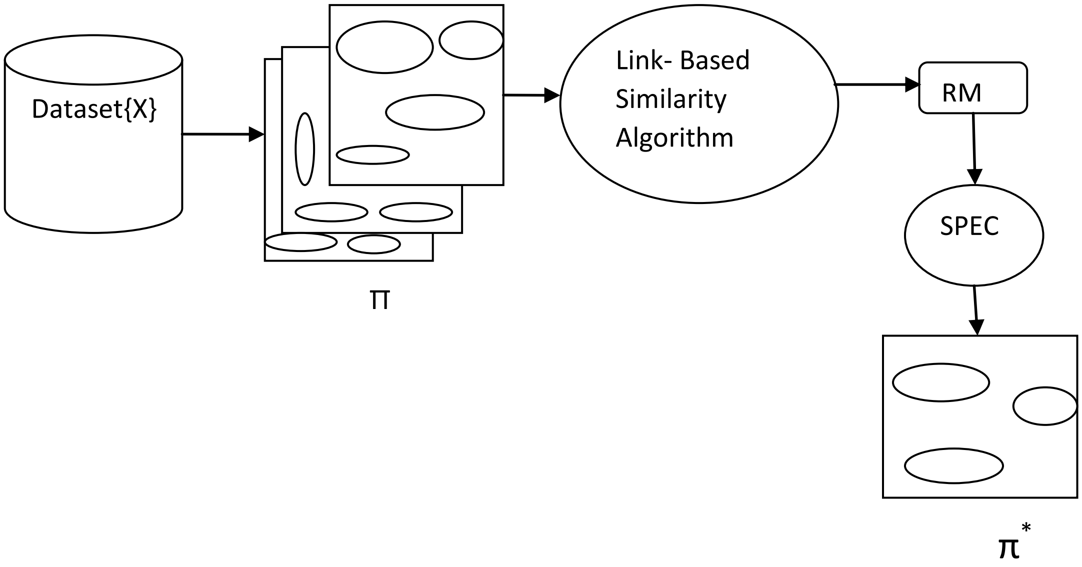

Fig: 2 The Link-based cluster ensemble Framework

Drawbacks:

Here there is no security for data. The datacan be erased while we carrying the data through the pen drives or compact disks. So we can't provide a security for those activities. There is a chance for intruder attacks. Intruder can erase the data keep in your laptops or other electronic devices.

3. PROVIDING SECURITY FOR PURE CLUSTER THROUGH CLOUD

To overcome these problems we proposed a cloud server to protect the data. This cloud server can provide a security to the data. Here we develop a web site as cloud server. Through this server user can upload the data. There is no chance for intruder attacks and data loss because we provide a username and password to the user. By using these username and password user can upload the data into our web site. From this cloud we can access the data set and we can form the pure cluster according to given data set.

The graphical structure of proposing system is as follows.

Fig3 explains about the how the user upload the data into cloud, and how the data accessed from cloud will be converted into pure cluster will be explained as follows. According to link-based similarity technique [6], it is only applicable to estimate the similarity among the data points is not applicable to large given data set. So to overcome these problem, a new link-based cluster ensemble(LCE) approach is introduced. This LCE includes three steps of : 1) creating base clusterings to form a cluster ensemble(π) 2) generating a refined cluster-association matrix (RM) using a link-based similarity algorithm, and 3) producing the final data partition (π*).

3.1 Cluster ensemble methodology:

There are three types of ways are available to form the cluster ensembles 1.Direct ensemble 2.Full-space ensemble 3.Subspace ensemble.

1.1Direct ensemble:

In this direct ensemble method categorical data X can be directly converted into cluster ensemble π. Let X ={x1,…..,xN} be a set of N data points, A={a1,…,aM} be a set of categorical attributes, and ∏ ={ π1,…, π M} be a set of M partitions. Each partition πi is created for a specific categorical attribute ai belongs to A.

1.2.Full-space ensemble:

In this full-space ensemble method k-modes technique[7] is used to generate base clusterings. These base clusterings are generated from the original data with all data attributes.

1.3.Subspace ensemble:

In this Subspace ensemble method, ensemble is created through the number of different data subsets. Here K-modes is applied to generate cluster ensemble from the set of subspace attributes, using both fixed-k and random-k schemes for selecting the number of clusters.

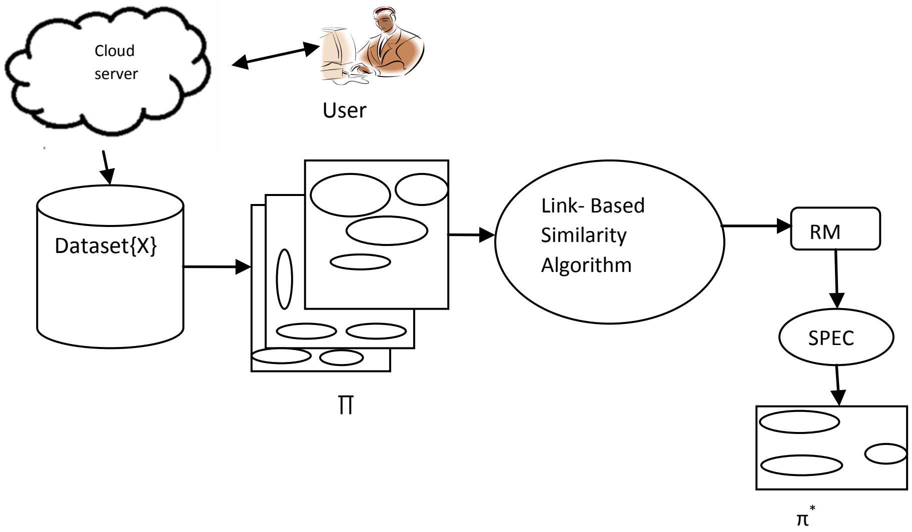

Fig:3 Categorical data clustering over cloud

3.2 Creation of refined cluster-association matrix using link based similarity:

From the cluster ensemble, a variety of consensus function is generate to create a refined cluster-association matrix. This consensus function uses the specific form of information matrix, which summarizes the base clustering results. There are three types of information matrix has been constructed 1)lable-assignment matrix, 2) Pairwise-similarity matrix, 3) Binary cluster association matrix.

3.2.1. Lable-assignment matrix:

Lable-assignment matrix of size N*M represent cluster lables that are assigned to each data point by different base clusterings.

3.2.2 Pairwise-similarity matrix:

The Pairwise-similarity matrix of size N*M, summarizes co-occurrence statistics among data points.

3.2.3. Binary cluster association matrix:

The binary cluster-association matrix (BM) provides acluster-specific view of the original label-assignment matrix.The association degree that a data point belonging to a specific cluster is either 1 or 0.

Consensus function utilizes the information of above three matrix and form the refined cluster association matrix.

3.2.4. Final cluster Result:

The final cluster result will be formed based on graph based approach [8,910]. In this approach it makes use of the graph representation to solve the cluster ensemble problem. A graph representing the similarity among the data points is created from a pair wise matrix. By using METIS [11] we can achieve the final cluster result. By using Spectral graph partitioning(spec)[12] method on refined matrix we can form the final cluster result.

This final cluster result is nothing but pure cluster. In this final cluster result, the clusters does not have the similar properties with each other. The dissimilarity between clusters is increased through cloud algorithm. The pure cluster will be reloaded into the cloud. Now the user can access the pure cluster from cloud with out data loss and impurity cluster

ADVANTAGES:

There is no chance for data loss. No one can copy the data from the cloud server. There is no chance for intruder attacks. User can easily access the data from anywhere and anytime.

4. RESULTS

Link-based cluster ensemble approaches frequently carry out better than the inspected gathering of cluster ensemble techniques and clustering algorithms for definite data. Link-based cluster ensembles also improve the performance of k-modes, which is used as base clustering. With the process of Link-based cluster ensemble models being mostly superior to those of the corresponding baseline counterparts, the excellence of the refined cluster-association matrix appears to be considerably better than that of the original, binary variation. Link-based cluster ensemble approach is more precise than other cluster ensemble methods.

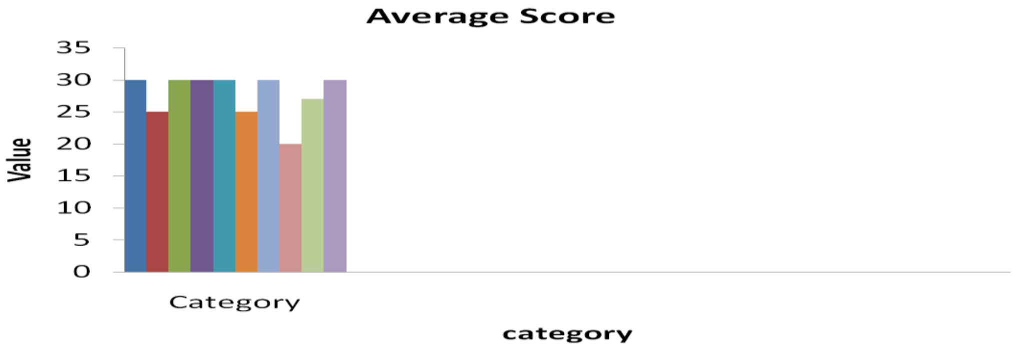

FIG: 4 Average score of Clusters

This graph is based on refine matrix, that means we are shown ensemble the categorical data. This section presents the evaluation of the proposed link based method using a variety of validity indices and real data sets. In this section we representing in the form graph chart that show the performance of the application.

5. CONCLUSION

Many entrenched clustering algorithms have been planned for numerical data, whose intrinsic properties can be obviously engaged to calculate a distance connecting feature vectors. In recent years, by means of applications to interesting domains such as protein interaction data many categorical data clustering algorithms have been introduced. A new link-based cluster ensemble approach is commenced which is well-organized than the previous model, where a binary cluster-association matrix-like matrix is used to correspond the ensemble information. The spotlight has shifted from enlightening the resemblance between data points to estimating those among clusters. A new link-based algorithm has been specifically planned to generate such measures in an accurate, reasonably priced manner. The link-based cluster ensemble methodology includes three major steps of: creating base clustering to form a cluster ensemble producing a refined cluster-association matrix by means of a link-based similarity algorithm, and producing the concluding data partition by making use of the spectral graph partitioning method as a consensus utility. The major future works consist of a broad study concerning the behaviour of other link-based similarity process within this trouble circumstance. In addition, the novel method will be applied to precise domains, together with tourism and medical data sets.

6. REFERENCES

[1] A Link-Based Cluster Ensemble Approach for Categorical Data Clustering Natthakan Iam-On, Tossapon Boongoen, Simon Garrett, and Chris Price

[2] L. Kaufman and P.J. Rousseeuw, Finding Groups in Data: An Introduction to Cluster Analysis. Wiley Publishers, 1990.

[3] A.K. Jain and R.C. Dubes, Algorithms for Clustering. Prentice-Hall, 1998.

[4] P. Zhang, X. Wang, and P.X. Song, "Clustering Categorical Data Based on Distance Vectors," The J. Am. Statistical Assoc., vol. 101, no. 473, pp. 355-367, 2006.

[5] S. Guha, R. Rastogi, and K. Shim, "ROCK: A Robust Clustering Algorithm for Categorical Attributes," Information Systems, vol. 25, no. 5, pp. 345-366, 2000.

[6] G. Jeh and J. Widom, "Simrank: A Measure of Structural-Context Similarity," Proc. ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD), pp. 538-543, 2002.

[7] Z. Huang, "Extensions to the K-Means Algorithm for Clustering Large Data Sets with Categorical Values," Data Mining and Knowledge Discovery, vol. 2, pp. 283-304, 1998.

[8] C. Domeniconi and M. Al-Razgan, "Weighted Cluster Ensembles:Methods and Analysis," ACM Trans. Knowledge Discovery from Data, vol. 2, no. 4, pp. 1-40, 2009.

[9] X.Z. Fern and C.E. Brodley, "Solving Cluster Ensemble Problems by Bipartite Graph Partitioning," Proc. Int'l Conf. Machine Learning (ICML), pp. 36-43, 2004.

[10] A. Strehl and J. Ghosh, "Cluster Ensembles: A Knowledge Reuse Framework for Combining Multiple Partitions," J. Machine Learning Research, vol. 3, pp. 583-617, 2002.

[11] G. Karypis and V. Kumar, "Multilevel K-Way Partitioning Scheme for Irregular Graphs," J. Parallel Distributed Computing, vol. 48, no. 1, pp. 96-129, 1998.

[12] A. Ng, M. Jordan, and Y. Weiss, "On Spectral Clustering: Analysis and an Algorithm," Advances in Neural Information Processing Systems, vol. 14, pp. 849-856, 2001.