A Universal Similarity Model for Transactional Data Clustering

[1] Mr.G.Kamalraja [2] Mrs.K.Prathiba [3]Ms.C.M.Parameshwari
[1] Assistant Professor, Department Of IT, , Excel Engineering college, Namakkal.
[2] Assistant Professor, Department Of CSE, Shree Sathyam college of Engineering and Technology, Salem.
[3] Lecturer, Department Of CSE, Shree Sathyam college of Engineering and Technology, Salem.
[1] kamal.raja31@gmail.com, [2]kprathi60@gmail.com, [3]parameshalages@gmail.com

ABSTRACT-Data mining methods are used to extract hidden knowledge from large database. Data partitioning methods are used to group up the relevant data values. Similar data values are grouped under the same cluster. K-means and Partitioning Around Medoids (PAM) clustering algorithms are used to cluster numerical data. Distance measures are used to estimate the transaction similarity.

Data partitioning solutions are identified using the cluster ensemble models. The ensemble information matrix presents only cluster data point relations. Ensembles based clustering techniques produces final data partition based on incomplete information. Link-based approach improves the conventional matrix by discovering unknown entries through cluster similarity in an ensemble. Link-based algorithm is used for the underlying similarity assessment. Pairwise similarity and binary cluster association matrices summarize the underlying ensemble information. A weighted bipartite graph is formulated from the refined matrix. The graph partitioning technique is applied on the weighted bipartite graph.

The Particle Swarm Optimization (PSO) clustering algorithm is a optimization based clustering scheme. It is integrated with the cluster ensemble model. Binary, categorical and continuous data clustering is supported in the system. The attribute connectivity analysis is optimized for all attributes. Refined cluster-association matrix (RM) is updated with all attribute relationships.

*****

1. INTRODUCTION

Clustering is the process of organizing data objects into a set of disjoint classes called clusters. Clustering is an example of unsupervised classification. Classification refers to a procedure that assigns data objects to a set of classes. Unsupervised means that clustering does not depends on predefined classes and training examples while classifying the data objects. Cluster analysis seeks to partition a given data set into groups based on specified features so that the data points within a group are more similar to each other than the points in different groups. Therefore, a cluster is a collection of objects that are similar among themselves and dissimilar to the objects belonging to other clusters. Clustering is an crucial area of research, which finds applications in many fields including bioinformatics, pattern recognition, image processing, marketing, data mining, economics, etc. Cluster analysis is a one of the primary data analysis tool in the data mining. Clustering algorithms are mainly divided into two categories: Hierarchical algorithms and Partition algorithms. A hierarchical clustering algorithm divides the given data set into smaller subsets in hierarchical fashion. A partition clustering algorithm partition the data set into desired number of sets in a single step.

Particle swarm optimization (PSO) is an evolutionary computation technique motivated by the simulation of social behavior. PSO optimizes a problem by having a population of candidate solutions, here dubbed particles, and moving these particles around in the search-space according to simple mathematical formulae over the particle's position and velocity. Each particle's movement is influenced by its local best known position and is also guided toward the best known positions in the search-space, which are updated as better positions are found by other particles. This is expected to move the swarm toward the best solutions.

PSO simulates the behaviors of bird flocking. Group of birds are randomly searching food in an area. There is only one piece of food in the area being searched. All the birds do not know where the food is. But they know how far the food is in each iteration. So what's the best strategy to find the food? The effective one is to follow the bird which is nearest to the food. PSO learned from the scenario and used it to solve the optimization problems. In PSO, each single solution is a "bird" in the search space. We call it "particle". All of particles have fitness values which are evaluated by the fitness function to be optimized, and

have velocities which direct the flying of the particles. PSO is initialized with a group of random particles and then searches for optima by updating generations.

2. RELATED WORKS

The difficulty of categorical data analysis is characterized by the fact that there is no inherent distance between attribute values. The RM matrix that is generated within the LCE approach allows such measure between values of the same attribute to be systematically quantified. The concept of link analysis is uniquely applied to discover the similarity among attribute values, which are modeled as vertices in an undirected graph. In particular, two vertices are similar if the neighboring contexts in which they appear are similar. In other words, their similarity is justified upon values of other attributes with which they co-occur. While the LCE methodology is novel for the problem of cluster ensemble, the concept of defining similarity among attribute values has been analogously adopted by several categorical data clustering algorithms.

ROCK makes use of a link graph, in which nodes and links represent data points and their similarity, respectively. The graph models used by ROCK and LCE are dissimilar—the graph of data points and that of attribute values, respectively. CACTUS also relies on the co-occurrence among attribute values. In essence, two attribute values are strongly connected if their support exceeds a prespecified value. Unlike LCE, the underlying problem is not designed using a graph based concept. Besides these approaches, traditional categorical data analysis also utilizes the "market-basket" numerical representation of the nominal data matrix. This transformed matrix is similar to the BM, which has been refined to the RM counterpart by LCE. A similar attempt identifies the connection between "category utility" of the conceptual clustering and the classical objective function of k-means. Despite the fact that many clustering algorithms and LCE are developed with the capability of comparing attribute values in mind, they achieve the desired metric differently, using specific information models. LCE uniquely and explicitly models the underlying problem as the evaluation of link-based similarity among graph vertices, which stand for specific attribute values or generated clusters. The resulting system is more efficient and robust, as compared to other clustering techniques emphasized thus far. In addition to SPEC, many other classical clustering techniques, k-means and PAM among others, can be directly used to generate the final data partition from the proposed RM. The LCE framework is generic such that it can be adopted for analyzing other types of data.

The Link based Cluster Ensembles (LCE) model is integrated with K-means clustering technique to improve the cluster accuracy. The system is tuned to cluster binary, categorical and continuous data values. The attribute connectivity analysis is optimized for all attributes. Refined cluster-association matrix (RM) is updated with all attribute relationships. The link based similarity model is suitable for categorical data only. Continuous data values are converted into categorical data format before similarity analysis. The binary and categorical data are compared in the same way. Continuous data associations are integrated with the similarity model.

Homogeneous, random-k and heterogeneous models are used for the ensemble selection process. Random-k selection model is used for the categorical data clustering process. The ensemble selection is formulated with integrated similarity analysis model. Homogeneous ensemble selection model is used in the multi model data clustering process. Graph based clustering approach is used in categorical data clustering process. The accuracy level is limited in the graph based clustering process. The pairwise similarity based clustering model is used. The refined matrix with similarity information is used in the K-means clustering algorithm.

3. CLUSTER ENSEMBLE METHODOLOGY

3.1. Categorical Data Clustering

Many categorical data clustering algorithms have been introduced in recent years, with applications to interesting domains such as protein interaction data. The initial method was developed making use of Gower's similarity coefficient. Following that, the k-modes algorithm extended the conventional k-means with a simple matching dissimilarity measure and a frequency-based method to update centroids. As a single-pass algorithm, Squeezer [9] makes use of a prespecified similarity threshold to determine which of the existing clusters which a data point under examination is assigned. LIMBO is a hierarchical clustering algorithm that uses the Information Bottleneck (IB) framework to define a distance measure for categorical tuples. The concepts of evolutionary computing and genetic algorithm have also been adopted by a partitioning method for categorical data. Cobweb is a model-based method primarily exploited for categorical data sets. Different graph models have also been investigated by the STIRR, ROCK and CLICK techniques. In addition, several density-based algorithms have also been devised for such purpose, for instance, CACTUS, COOLCAT and CLOPE.

A large number of algorithms have been introduced for clustering categorical data, the No Free Lunch theorem suggests1 there is no single clustering algorithm that performs best for all data sets and can discover all types of cluster shapes and structures presented in data. Each algorithm has its own strengths and weaknesses. For a

particular data set, different algorithms, or even the same algorithm with different parameters, usually provide distinct solutions. Therefore, it is difficult for users to decide which algorithm would be the proper alternative for a given set of data. Recently, cluster ensembles have emerged as an effective solution that is able to overcome these limitations, and improve the robustness as well as the quality of clustering results. The main objective of cluster ensembles is to combine different clustering decisions in such a way as to achieve accuracy superior to that of any individual clustering.

3.2. Ensemble Generation Methods

It has been shown that ensembles are most effective when constructed from a set of predictors whose errors are dissimilar. To a great extent, diversity among ensemble members is introduced to enhance the result of an ensemble. Particularly for data clustering, the results obtained with any single algorithm over much iteration are usually very similar. In such a circumstance where all ensemble members agree on how a data set should be partitioned, aggregating the base clustering results will show no improvement over any of the constituent members. As a result, several heuristics have been proposed to introduce artificial instabilities in clustering algorithms, giving diversity within a cluster ensemble. The following ensemble generation methods yield different clusterings of the same data, by exploiting different cluster models and different data partitions.

- Homogeneous ensembles. Base clusterings are created using repeated runs of a single clustering algorithm, with several sets of parameter initializations, such cluster centers of the k-means clustering technique.
- Random-k. One of the most successful techniques is randomly selecting the number of clusters (k) for each ensemble member.
- Data subspace/sampling. A cluster ensemble can also be achieved by generating base clusterings from different subsets of initial data. It is intuitively assumed that each clustering algorithm will provide different levels of performance for different partitions of a data set [5]. Practically speaking, data partitions are obtained by projecting data onto different subspaces, choosing different subsets of features, or data sampling.
- Heterogeneous ensembles. A number of different clustering algorithms are used together to generate base clusterings.
- Mixed heuristics. In addition to using one of the aforementioned methods, any combination of them can be applied as well.

3.3. Cluster Ensembles of Categorical Data

While a large number of cluster ensemble techniques for numerical data have been put forward in the previous decade, there are only a few studies that apply such a methodology to categorical data clustering. The method introduced creates an ensemble by applying a conventional clustering algorithm to different data partitions, each of which is constituted by a unique subset of data attributes. Once an ensemble has been obtained, the graph-based consensus functions are utilized to generate the final clustering result.

Unlike the conventional approach, the technique developed acquires a cluster ensemble without actually implementing any base clustering on the examined data set. In fact, each attribute is considered as a base clustering that provides a unique data partition. In particular, a cluster in such attribute-specific partition contains data points that share a specific attribute value. Thus, the ensemble size is determined by the number of categorical labels, across all data attributes. The final clustering result is generated using the graph-based consensus techniques. Specific to this so-called “direct” ensemble generation method, a given categorical data set can be represented using a binary cluster-association matrix. Such an information matrix is analogous to the “market-basket” numerical representation of categorical data, which has been the focus of traditional categorical data analysis [8].

4. PROBLEM STATEMENT

Cluster ensembles are used to combine different clustering decisions. The ensemble information matrix presents only cluster data point relations. Ensembles based clustering techniques produces final data partition based on incomplete information. Link-based approach improves the conventional matrix by discovering unknown entries through cluster similarity in an ensemble. Link-based algorithm is used for the underlying similarity assessment. Pairwise similarity and binary cluster association matrices summarize the underlying ensemble information. A weighted bipartite graph is formulated from the refined matrix. The graph partitioning technique is applied on the weighted bipartite graph. The following drawbacks are identified from the system.

- Similarity analysis is tuned for categorical data only
- Limited cluster accuracy level
- The system supports graph based partitioning model only

- Cluster ensemble selection is not optimized

5. CATEGORICAL DATA CLUSTERING USING LINK SIMILARITY

Existing cluster ensemble methods to categorical data analysis rely on the typical pairwise-similarity and binary cluster-association matrices [1], which summarize the underlying ensemble information at a rather coarse level. Many matrix entries are left "unknown" and simply recorded as "0." Regardless of a consensus function, the quality of the final clustering result may be degraded. As a result, a link based method has been established with the ability to discover unknown values and, hence, improve the accuracy of the ultimate data partition. In spite of promising findings, this initial framework is based on the data pointdata point pairwise-similarity matrix, which is highly expensive to obtain. The link-based similarity technique, SimRank that is employed to estimate the similarity among data points is inapplicable to a large data set.

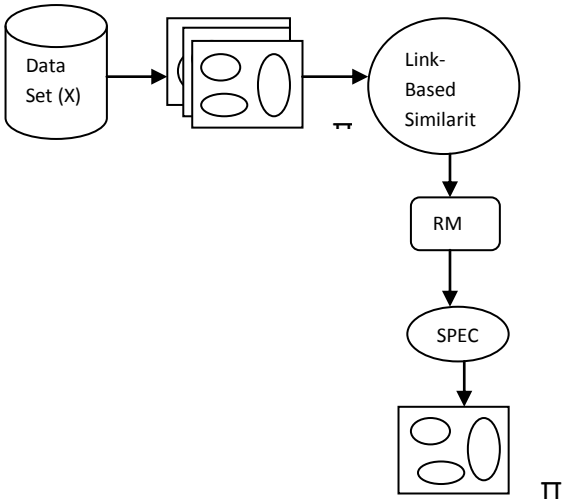

Fig.1. The link-based cluster ensemble framework

To overcome these problems, a new link-based cluster ensemble (LCE) approach is introduced herein. It is more efficient than the former model, where a BM-like matrix is used to represent the ensemble information. The focus has shifted from revealing the similarity among data points to estimating those between clusters [10]. A new link-based algorithm has been specifically proposed to generate such measures in an accurate, inexpensive manner. The LCE methodology is illustrated in Fig. 1. It includes three major 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^*) by exploiting the spectral graph partitioning technique as a consensus function.

5.1. Creating a Cluster Ensemble

The first type of cluster ensemble transforms the problem of categorical data clustering to cluster ensembles by considering each categorical attribute value as a cluster in an ensemble. Let X = $\{x_1, \ldots, x_N\}$ be a set of N data points, A =$\{a_1, \ldots, a_M\}$ be a set of categorical attributes, and Π = $\{\pi_1, \ldots . \pi_M\}$ be a set of M partitions. Each partition π_i is generated for a specific categorical attribute $a_i \in$ A. Clusters belonging to a partition $\pi_i = \{C_1^i, C_2^i, \ldots \ldots . C_{ki}^i\}$ correspond to different values of the attribute $a_i = \{a_1^i, , \ldots \ldots , a_{ki}^i\}$, where $U_{j=1}^{k_i} C_j^i = a_i$ and k_i is the number of values of attribute a_i. With this formalism, categorical data X can be directly transformed to a cluster ensemble Π, without actually implementing any base clustering. While single-attribute data partitions may not be as accurate as those obtained from the clustering of all data attributes, they can bring about great diversity within an ensemble. Besides its efficiency, this ensemble generation method has the potential to lead to a high-quality clustering result.

5.2. Generating a Refined Matrix

Several cluster ensemble methods, both for numerical [6] and categorical data [7], are based on the binary cluster-association matrix. Each entry in this matrix BM (x_i, cl) $\in$ {0, 1} represents a crisp association degree between data point $x_i \in$ X and cluster cl $\in$ Π. An example of cluster ensemble and the corresponding BM, a large number of entries in the BM are unknown, each presented with "0." Such condition occurs when relations between different clusters of a base clustering are originally assumed to be nil. In fact, each data point can possibly associate to several clusters of any particular clustering. These hidden or unknown associations can be estimated from the similarity among clusters, discovered from a network of clusters.

Based on this insight, the refined cluster-association matrix is put forward as the enhanced variation of the original BM. Its aim is to approximate the value of unknown associations ("0") from known ones ("1"), whose association degrees are preserved within the RM, i.e., BM(x_i, cl) = 1→ RM (x_i, cl) = 1. For each clustering π_t, t = 1 . . .M and their corresponding clusters $C_1^i, , \ldots \ldots C_{ki}^i$ (where k_t is the number of clusters in the clustering π_t), the

association degree RM (x_i, cl) $\in$ [0, 1] that data point $x_i \in X$ has with each cluster cl $\in \{C_1^i, \dots \dots C_{ki}^i\}$ is estimated as follows:

$$RM(x_i, cl) = \begin{cases} 1, & if cl = c_x^t(x_i), \\ sim(CL, c_x^t(x_i)), & otherwise, \end{cases} \quad (1)$$

where $c_*^t(x_i)$ is a cluster label to which data point x_i belongs. In addition, sim (C_x, C_y) $\in$ [0,1] denotes the similarity between any two clusters C_x, C_y, which can be discovered using the link-based algorithm. Note that, for any clustering $\pi_t \in \Pi$, $1 \leq \sum_{\forall C \in \pi t}$ RM(x_i, C) $\leq k_t$. Unlike the measure of fuzzy membership, the typical constraint of $\sum_{\forall C \in \pi t}$ RM(x_i, C) = 1 is not appropriate for rescaling associations within the RM. In fact, this local normalization will significantly distort the true semantics of known associations ("1"), such that their magnitudes become dissimilar, different from one clustering to another. According to the empirical investigation, this fuzzy-like enforcement decreases the quality of the RM, and hence, the performances of the resulting cluster ensemble method.

5.2.1. Weighted Triple-Quality (WTQ) Algorithm

Given a cluster ensemble Π of a set of data points X, a weighted graph G = (V,W) can be constructed, where V is the set of vertices each representing a cluster and W is a set of weighted edges between clusters. Formally, the weight assigned to the edge $w_{xy} \in W$, that connects clusters C_x, $C_y \in V$, is estimated by the proportion of their overlapping members.

$$W_{xy} = \frac{|L_x \cap L_y|}{|L_x \cup L_y|} \quad - \quad (2)$$

where $L_z \subset X$ denotes the set of data points belonging to cluster $C_z \in V$. Note that circle nodes represent clusters and edges exist only when the corresponding weights are nonzero.

Shared neighbors have been widely recognized as the basic evidence to justify the similarity among vertices in a link network [2]. Formally, a vertex $C_k \in V$ is a common neighbor of vertices C_x, $C_y \in V$, provided that w_{xk}, $w_{yk} \in W$. Many advanced methods extend this basis by taking into account common neighbors that may be many edges away from the two under examination: for instance, Connected-Path, SimRank and a variation of random walk algorithms [3]. Despite reported effectiveness, these techniques are computationally expensive, or even impractical for a large data set. Henceforth, the Weighted Triple-Quality algorithm is proposed, as part of the current research, for the efficient approximation of the similarity between clusters in a link network. Unlike the technique in [4] that simply counts the number of triples, WTQ aims to differentiate the significance of triples and hence their contributions toward the underlying similarity measure. WTQ is inspired by the initial measure in [11], which evaluates the association between home pages. The WTQ algorithm. 1 is summarized below:

The WTQ algorithm is summarized below:

Algorithm 1: WTQ(G, G_x,G_y)

G=(V,W), a weighted graph, where C_x, $C_y \in V$;

$N_k = \subset V$, a set of adjacent neighbors of $C_k \in V$;

$W_k = \sum_{\forall Z_C \in Z} w_{tk}$;

WTQ_{xy} , the WTQ measure of C_x {and} C_y;

1. WTQ_{xy} WT← 0
2. For each c∈ N_x
3. If c∈ N_y
4. $WTQ_{xy} \leftarrow WTQ_{xy} + \frac{1}{W_e}$
5. Return WTQ_{xy}

Following that, the similarity between clusters C_x and C_y can be estimated by

$$Sim(C_x, C_y) = \frac{WTQ_{xy}}{WTQ_{\max}} \times DC, \quad (3)$$

where WTQ_{max} is the maximum WTQ_{pq} value of any two clusters C_p, $C_q \in V$ and DC $\in$ [0,1] is a constant decay factor. With this link-based similarity metric, sim(C_x, C_y) $\in$ [0, 1] with sim(C_x, C_x) = 1, C_x, $C_y \in V$. It is also reflexive such that sim(C_x, C_y) is equivalent to sim(C_y, C_x).

5.3. Consensus Function to RM

Having obtained an RM, a graph-based partitioning method is exploited to obtain the final clustering. This consensus function requires the underlying matrix to be initially transformed into a weighted bipartite graph. Given an RM representing associations between N data points and P clusters in an ensemble Π, a weighted graph G = (V, W) can be constructed, where $V = V^X \cup V_C$ is a set of vertices representing both data points V^X and clusters V^C, and W denotes a set of weighted edges that can be defined as follows:

- $w_{ij} \in W$ when vertices v_i, $v_j \in V^X$.
- $w_{ij} \in W$ when vertices v_i, $v_j \in V^C$.

- Otherwise, $w_{ij} = RM(v_i, v_j)$ when vertices $v_i \in V^X$ and $v_j \in V^C$. Note that the graph G is bidirectional such that w_{ij} is equivalent to w_{ji}.

Given such a graph, a spectral graph partitioning method similar is applied to generate a final data partition. This is a powerful method for decomposing an undirected graph, with good performance being exhibited in many application areas, including protein modeling, information retrieval, and identification of densely connected online hyper textual regions. Principally, given a graph G = (V ,W), SPEC first finds the K largest eigenvectors $u_1, \ldots, u_K$ of W, which are used to formed another matrix U, whose rows are then normalized to have unit length. By considering the row of U as K-dimensional embedding of the graph vertices, SPEC applies k-means to these embedded points in order to acquire the final clustering result.

6. PARTICLE SWARM OPTIMIZATION

Particle swarm optimization is a stochastic population based optimization approach, first published by Kennedy and Eberhart in 1995.The PSO algorithm is in the step which calculates a new position for the particle based on three influences. The particle is moving in a search space and can measure the `fitness' of any position.

The influences - the components that lead to the updated position – are:

Current velocity: the particle's current velocity;

Personal Best: The particle remembers the fittest position it has yet encountered, called the personal best. A component of its updated velocity is the direction from its current position to the personal best.

Global Best: Every particle in the swarm is aware of the best position that any particle has yet discovered. The final component of velocity update, shared by all particles, is a vector in the direction from its current position to this globally best known position.

Particle swarm optimization (PSO) is a stochastic optimization approach which maintains a swarm of candidate solutions, referred to as particles. Particles are ‘‘flown’’ through hyper-dimensional search space, with each particle being attracted towards the best solution found by the particles neighborhood and the best solution found by the particle. The position, xi, of the ith particle is adjusted by a stochastic velocity vi which depends on the distance that the particle is from its own best solution and that of its neighborhood. For the original PSO,

$$V_{ij}(t+1) = v_{ij}(t) + \emptyset_{1j}(t)(y_{ij}(t) - x_{ij}(t)) + \emptyset_{2j}(t)(\hat{y}_{ij}(t) - x_{ij}(t))$$

$$X_{ij}(t+1) = x_{ij}(t) + v_{ij}(t+1)$$

For i = 1. . . s and j = 1. . . n, where

$\emptyset_{1j}(t) = c_1 r_{1j}(t)$ and $\emptyset_{2j}(t) = c_2 r_{2j}(t)$,

s is the total number of particles in the swarm, n is the dimension of the problem, i.e. the number of parameters of the function Being optimized, c_1 and c_2 are acceleration coefficients,

$r_{1j}(t)$, $r_2j \sim U(0, 1)$,

$x_i(t)$ is the position of particle i at time step t, $v_i(t)$ is the velocity of particle i at time step t, $y_i(t)$ is the personal best solution of particle i, at time step t, $\hat{y}_i(t)$ is the best position found by the neighborhood of particle i, at time step t.

7. DATA PARTITIONING USING TYPE INDEPENDENT SIMILARITY ANALYSIS

The system is designed to perform multi model data clustering with efficient similarity analysis model. The link based similarity model is tuned to analyze multi model data values. Ensemble based clustering approach is used in the system. The system consists of six modules. They are Data cleaning process, Ensemble selection, Similarity analysis categorical data, Similarity analysis for independent data, Graph partitioning, and clustering with PSO.

The data cleaning module is designed to update noise data values. The ensemble selection module is designed to identify the cluster initial ensembles. Categorical data relationship analysis is carried out under similarity for categorical data module. Independent data similarity module is designed to analyze the data with type independency. Cluster process is done with graph partitioning model. The Particle Swarm Optimization (PSO) technique is used for the clustering process.

7.1. Data Cleaning Process

The cancer diagnosis details are imported from textual data files. Redundant data values are removed from the database. Missing elements are assigned using aggregation based substitution mechanism. Cleaned data values are referred as optimal data sets.

7.2. Ensemble Selection

Cluster ensembles are selected from the transaction collections. Cluster count is collected from the user. Random based ensemble selection is used in the system. Binary cluster association matrix is constructed using similarity analysis.

7.3. Similarity Analysis Categorical Data

The continuous data values are converted into categorical data. Median values are used in the conversion process. Link similarity model is used in the relationship analysis. The similarity values are updated into the binary cluster association matrix.

7.4. Similarity Analysis for Independent Data

The similarity analysis is performed to estimate the transaction relationship. Independent data similarity is designed for binary, categorical and continuous data types. Link similarity model is tuned to find similarity for all type of data values. Vector and link models are integrated for relationship analysis.

7.5. Graph Partitioning

The binary cluster association matrix is composed with incomplete similarity details. The refined matrix is prepared from the binary cluster association matrix values. The refined matrix values are updated into graphs. The similarity intervals are used to partition the data values.

7.6. Clustering with PSO

The clustering process is divided into two types. They are categorical data clustering and mixed data clustering models. Categorical data clustering model converts all the data values into categorical data. Binary, categorical and continuous data values are used in mixed data clustering process. PSO based clustering algorithm is used with similarity weight values.

8. CONCLUSION

Cluster ensembles are used to cluster the categorical data values. Link-based cluster ensemble approach is used to categorical data clustering. The Link based cluster ensembles (LCE) technique is integrated with K-means clustering model. The LCE model is tuned to cluster all types of data values. The system performs the data partitioning based on the link similarities. The similarity estimation is optimized for all type of data values. The system improves the matrix refinement process. PSO based clustering is applied on the refined similarity matrix for data partitioning process.

REFERENCES

[1] Z. He, X. Xu, and S. Deng, "A Cluster Ensemble Method for Clustering Categorical Data," Information Fusion, vol. 6, no. 2, pp. 143-151, 2005.

[2] D. Liben-Nowell and J. Kleinberg, "The Link-Prediction Problem for Social Networks," J. Am. Soc. for Information Science and Technology, vol. 58, no. 7, pp. 1019-1031, 2007.

[3] E. Minkov, W.W. Cohen, and A.Y. Ng, "Contextual Search and Name Disambiguation in Email Using Graphs," Proc. Int'l Conf. Research and Development in IR, pp. 27-34, 2006.

[4] P. Reuther and B. Walter, "Survey on Test Collections and Techniques for Personal Name Matching," Int'l J. Metadata, Semantics and Ontologies, vol. 1, no. 2, pp. 89-99, 2006.

[5] 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.

[6] 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.

[7] M. Al-Razgan, C. Domeniconi, and D. Barbara, "Random Subspace Ensembles for Clustering Categorical Data," Supervised and Unsupervised Ensemble Methods and Their Applications, pp. 31-48, Springer, 2008.

[8] P.N. Tan, M. Steinbach, and V. Kumar, Introduction to Data Mining. Addison Wesley, 2005.

[9] Z. He, X. Xu, and S. Deng, "Squeezer: An Efficient Algorithm for Clustering Categorical Data," J. Computer Science and Technology, vol. 17, no. 5, pp. 611-624, 2002.

[10] Natthakan Iam-On, Tossapon Boongoen, Simon Garrett and Chris Price, "A Link-Based Cluster Ensemble Approach for Categorical Data Clustering" IEEE Transactions on Knowledge and Data Engineering, Vol. 24, no. 3, March 2012.

[11] L.A. Adamic and E. Adar, "Friends and Neighbors on the Web," Social Networks, vol. 25, no. 3, pp. 211-230, 2003.