Scientific Journal Impact Factor (SJIF) : 1.711
International Journal of Modern Trends in Engineering and Research
ISSN (ONLINE) : 2349 - 9745 ISSN (PRINT) : 2393 - 8161

LINK RELATIONSHIP BASED DATA PARTITIONING ON HEART DISEASE DIAGNOSIS DATA

Ms. K. Uma [1], Mrs. R. Sasiregha [2]
[1] M.Sc., MPhil., Research Scholar, [2] M.Sc., MPhil., Ph. D., Assistant Professor
Department of Computer Science,
SSM College of Arts & Science, Komarapaliyam, Tamilnadu, India

Abstract - Hidden knowledge discovery is carried out using data mining techniques in a variety of applications. Data partitioning methods are adapted to perform relevant record grouping process. Similarity measures are employed to estimate the similarity measures. Cosine and Euclidean distance measures are .Record link details are also applied to estimate the relationship levels.

Link based similarity estimation mechanism is adapted in the data partitioning process. Record links are identified using K Nearest Neighbor (KNN) model. Link relationship is referred as hubness relationship. Link relationship based model is applied for the data partitioning process on high dimensional data environment. Hubness measures are estimated with KNN query results. K Hubs algorithm is applied to define the clusters. Data partitioning operations are carried out using the Hubness-proportional K-means (HPKM) algorithm.

Link relationship based data partitioning methods are employed to perform heart disease analysis. Heart patient diagnosis data values are grouped using the hubness or link based data partitioning methods. Kernal Mapping Clustering (KMC) scheme and Shared Neighbor Clustering (SNC) scheme are enhanced with link based similarity scheme. The data partitioning process is also improved with automatic cluster count estimation mechanism.

Index Terms: Data clusters, Hubness measure, Heart disease diagnosis, Shared neighbor clustering and Nearest Neighbor search

I. INTRODUCTION

Data clustering is one of the fundamental tools for understanding the structure of a data set. It plays a crucial, foundational role in machine learning, data mining, information retrieval, and pattern recognition. Clustering aims to categorize data into groups or clusters such that the data in the same cluster are more similar to each other than to those in different clusters. Many well-established clustering algorithms, such as k-means and PAM, have been designed for numerical data, whose inherent properties can be naturally employed to measure a distance (e.g., euclidean) between feature vectors. However, these cannot be directly applied for clustering of categorical data, where domain values are discrete and have no ordering defined.

Semi-Supervised clustering aims to improve clustering performance with the help of user-provided side information. One of the most studied types of side information is pairwise constraints, which include mustlink and cannot-link constraints specifying that two points must or must not belong to the same cluster. A number of previous studies have demonstrated that, in general, such constraints

can lead to improved clustering performance. However, if the constraints are selected improperly, they may also degrade the clustering performance. Moreover, obtaining pairwise constraints typically requires a user to manually inspect the data points in question, which can be time consuming and costly. For example, for document clustering, obtaining a must-link or cannot-link constraint requires a user to potentially scan through the documents in question and determine their relationship, which is feasible but costly in time. For those reasons,The system optimizes the selection of the constraints for semi-supervised clustering, which is the topic of active learning.

II. RELATED WORK

While Huang's method is developed specifically for document clustering, one could potentially apply the underlying active learning approach to handle other types of data by assuming appropriate probabilistic models. The system highlights a key distinction between Huang's method and Huang's method makes the selection choice based on pairwise uncertainty, The system is focused on the uncertainty of a point in terms of which neighborhood it belongs to. This difference is subtle, but important. Pairwise uncertainty captures only the relationship between the two points in the pair. Depending on the outcome of the query, the system may need to go through a sequence of additional queries. Huang's method only considers the pairwise uncertainty of the first query, and fails t measure the benefit of the ensuing queries. Uncertainty measures the total amount of information gained by the full sequence of queries as a whole. Furthermore, the method also takes into account the expected number of queries to resolve the uncertainty of a point, which has not been considered previously. This is different from the focus of this system, only request pairwise must-link and cannotlink constraints, and do not require the user to provide specific distance values.

III. DATA CLUSTERING WITH HUBNESS RELATIONSHIP

Clustering in general is an unsupervised process of grouping elements together, so that elements assigned to the same cluster are more similar to each other than to the remaining data points. This goal is often difficult to achieve in practice. Over the years, various clustering algorithms have been proposed, which can be roughly divided into four groups: partitional, hierarchical, density based and subspace algorithms. Algorithms from the fourth group search for clusters in some lower dimensional projection of the original data and have been generally preferred when dealing with data that are high dimensional. The motivation for this preference lies in the observation that having more dimensions usually leads to the so-called curse of dimensionality, where the performance of many standard machine-learning algorithms becomes impaired. This is mostly due to two pervasive effects: the empty space phenomenon and concentration of distances. The former refers to the fact that all high-dimensional data sets tend to be sparse, because the number of points required to represent any distribution grows exponentially with the number of dimensions. This leads to bad density estimates for high-dimensional data, causing difficulties for density-based approaches. The latter is a somewhat counterintuitive property of high-dimensional data representations, where all distances between data points tend to become harder to distinguish as dimensionality increases, which can cause problems with distance-based algorithms.

Hubness is a good measure of point centrality within a high-dimensional data cluster and that major hubs can be used effectively as cluster prototypes. The algorithms frequently offer improvements in cluster quality and homogeneity. The comparison with kernel K means reveals that kernel-based extensions of the initial approaches should also be considered in the future. The current focus was mostly on properly selecting cluster prototypes, with the proposed methods tailored for detecting approximately hyper spherical clusters.

The hubness relationship is used in the data clustering process with predefined cluster count values. The hubness model is integrated with the K means clustering scheme to partition the health care data values. The following problems are discovered from the hubness based clustering approach. The system performs the clustering using predefined cluster count. The system fetches hyper spherical clusters only. Inter cluster distance is not optimized and Clustering accuracy is low.

IV. HUBNESS INTEGRATED SHARED NEIGHBOR CLUSTERING SCHEME

Hub objects are data points which have a small distance to many other data points in high dimensional spaces which are related to the phenomenon of concentration of distances. This behavior has a negative impact on many machine learning tasks including classification, nearest neighbor based recommendation outlier detection and clustering. Shared Nearest Neighbors (SNN) is an algorithm that re-scales distance spaces to so-called secondary distances. Without referring to the problem of hubness, it has been discussed as a way to "defeat the curse of dimensionality". Local scaling and mutual proximity, two approaches inspired by the general idea behind SNN, have already been shown to decisively reduce hubness and the concentration of distances. SNN itself has been applied successfully to high dimensional image recognition data also reducing hubness to a certain degree. The main contributions of this study are (i) an evaluation of the ability of SNN to reduce hubness on a larger set of high dimensional real world data sets from other domains and (ii) a comparison of SNN to local scaling and mutual proximity.

1. Construct the similarity matrix.
2. Sparsify the similarity matrix using knn sparsification.
3. Construct the shared nearest neighbor graph from knn sparsified similarity matrix.
4. For every point in the graph, calculate the total strength of links coming out of the point. (Steps 1-4 are identical to the Jarvis – Patrick scheme.)
5. Identify representative points by choosing the points that have high total link strength.
6. Identify noise points by choosing the points that have low total link strength and remove them.
7. Remove all links that have weight smaller than a threshold.
8. Take connected components of points to form clusters, where every point in a cluster is either a representative point or is connected to a representative point.

Algorithm 1: Shared Neighbor Clustering

The number of clusters is not given to the algorithm as a parameter. Depending on the nature of the data, the algorithm finds "natural" clusters. Also note that not all the points are clustered using out algorithm. Depending on the application, the system might actually discard many of the points. By using noise removal and the representative points, the system obtain the two clusters. The points that do not belong to any of the two clusters can be brought in by assigning them to the cluster that has the closest core point.

V. SYSTEM IMPLEMENTATION

The data clustering operations are carried out with the similarity or distance measures. Different distance measures are used for the high dimensional and low dimensional data values. Cluster accuracy is estimated with reference to the distance measures. The transactional link information is used in the hubness model. The hubness property refers the relationship between the transactions. The hubness relationship can be used for the deterministic and probabilistic data values.

The hubness model is constructed with the nearest neighbor relationships. The K-Nearest Neighbor (KNN) search algorithm is applied to fetch the nearest neighbor transactions. The K-Hubs algorithm is used to estimate the hubness score values. The Hubness Proportional K-Means (HPKM)

clustering algorithm is used to partition the data values. The system uses the user defined cluster count for the partitioning process.

The hubness based clustering scheme is improved with hubness score interval analysis mechanism. Kernel mapping scheme is enhanced with hubness relationship analysis. Shared neighbor clustering is integrated with the hubness mechanism. Hub based automatic cluster count estimation mechanism is integrated with the system.

The system uses three different clustering algorithms Hubness Proportional K-Means (HPKM) clustering algorithm, Shared Neighbor Clustering (SNC) algorithm and Kernal Map Clustering (KMC) algorithm. Each algorithm is implemented for user defined cluster count model and optimal cluster count based model. The user defined cluster count model collects the cluster count from the user. The optimal cluster count estimation procedure is applied to estimate the cluster count automatically. The hubness score is used in the optimal cluster count estimation process. Hubness score and its interval levels are used in the cluster count estimation process.

The hubness based data clustering scheme is designed to partition the healthcare data values. The patient diagnosis information is collected by the doctor. The diagnosis information is updated into the diagnosis database. The data preprocessing techniques are applied to eliminate the noises. The K-Nearest Neighbor (KNN) search process is initiated on the diagnosis data values. The hubness score is estimated using the KNN query results. The system performs the clustering process in two ways. They are predefined cluster count model and optimal cluster model. The clustering process is carried out using Hubness Proportional K-Means Clustering (HPKM) algorithm, Kernal Map Clustering (KMC) and Shared Neighbor Clustering (SNC) algorithm. The hubness based similarity estimation procedure is used in the Kernal Map Clustering (KMC) algorithm and the Shared Neighbor Clustering (SNC) algorithms. The cluster results are verified with different cluster quality estimation measures.

VI. EXPERIMENTAL ANALYSIS

The hubness relationship based clustering scheme is implemented to group up the heart patient diagnosis data values. The Hubness Proportinate K-Means Clustering (HPKMC) algorithm, Kernal Mapping Cluster (KMC) algorithm and Shared Neighbor Cluster (SNC) algorithm are used in the system. All the clustering schemes are implemented and tested with predefined cluster count and optimal cluster count model. In the predefined cluster count model the cluster count value is collected from the user. The optimal cluster count model automatically estimates the feasible cluster count for the partitioning process. The system is tested with two performance parameters to measure the cluster quality levels. They are purity and separation index levels. The system is tested with different data intervals.

6.1. Datasets

The clustering system is analyzed using the Heart patient data sets collected from the University of California Irwin (UCI) machine learning repository. The dataset is downloaded from http://archive.ics.uci.edu/ml/datasets.html. The diagnosis details are collected from patients from different countries. The dataset contains 1000 transactions with 15 attributes. The attribute name, description and type details. Missing values are replaced in preprocess. Aggregation based data substitution mechanism is used for the data preprocess. Redundant transactions are removed from the datasets during the preprocess.

6.2. Purity

The purity of a cluster represents the fraction of the cluster corresponding to the largest class of documents assigned to that cluster; thus, the purity of the cluster j is defined as

$$Purity(j) = \frac{1}{n_j} \max_i (n_{ij}) \quad (1)$$

The overall purity of the clustering result is a weighted sum of the purity values of the clusters as follows:

$$Purity = \sum_j \frac{n_j}{n} Purity(j) \quad (2)$$

In general, the larger the purity value is, the better the clustering result is (2).

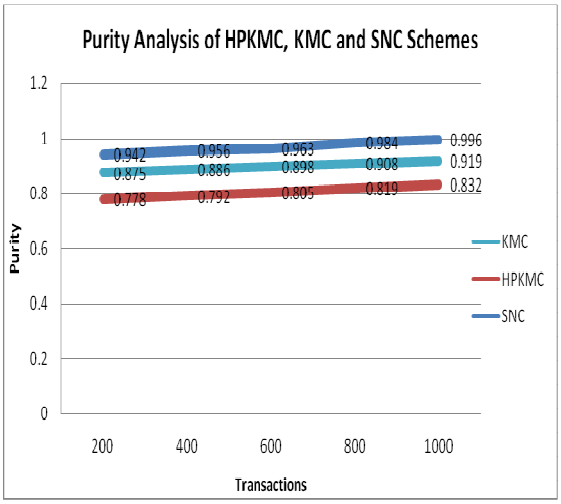

Figure No: 6.1. Purity analysis of HPKMC, KMC and SNC Schemes

The Purity analysis between the Hubness Proportionate K-Means Clustering (HPKMC) algorithm, Kernal Mapping Cluster (KMC) algorithm and Shared Neighbor Clustering (SNC) algorithm are shown in figure 6.1. The analysis result shows that the Kernal Mapping scheme improves the clustering accuracy 10% than the Hubness Proportionate K-Means Clustering (HPKMC) algorithm. The Shared Neighbor Clustering (SNC) scheme increases the clustering accuracy level 20% than the Hubness Proportionate K-Means Clustering (HPKMC) scheme.

6.3. Separation Index

Separation Index (SI) is another cluster validity measure that utilizes cluster centroids to measure the distance between clusters, as well as between points in a cluster to their respective cluster centroid.

. It is defined as the ratio of average within-cluster variance to the square of the minimum pairwise distance between clusters:

$$SI = \frac{\sum_{i=1}^{NC} \sum_{xj \in ci} dist(x_j, m_i)^2}{N_D \min_{1 \le r,s \le N_c} dist(m_r, m_s)\}^2}$$

$$= \frac{\sum_{i=1}^{N_c} \sum_{xj \in ci} dist(x_j, m_i)^2}{N_D \cdot dist_{\min}^2}$$

Where m_i is the centroid of cluster c_i, and distmin is the minimum pairwise distance between cluster centroids. Clustering solutions with more compact clusters and larger separation have lower Separation Index, thus higher values indicate better solutions. This index is more computationally efficient than other validity indices, such as Dunn's index, which is also used to validate clusters that are compact and well separated. In addition, it is less sensitive to noisy data.

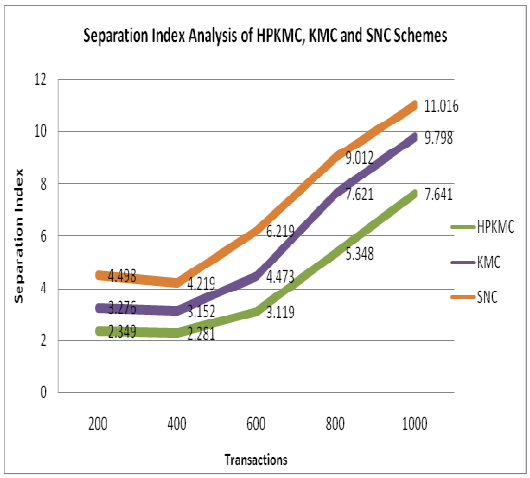

Figure No: 6.2. Separation Index Analysis of HPKMC, KMC and SNC Schemes

The Separation analysis between the Hubness Proportionate K-Means Clustering (HPKMC) algorithm, Kernal Mapping Cluster (KMC) algorithm and Shared Neighbor Clustering (SNC) algorithm are shown in figure 6.2. The analysis result shows that the Kernal Mapping scheme improves the clustering accuracy 10% than the Hubness Proportionate K-Means Clustering (HPKMC) algorithm. The Shared Neighbor Clustering (SNC) scheme increases the clustering accuracy level 30% than the Hubness Proportionate K-Means Clustering (HPKMC) scheme.

VII. CONCLUSION AND FUTURE WORK

The data clustering methods are applied to group up the relevant transactions. Distance measures or similarity measures are used to estimate the transaction relationship levels. Euclidean distance measure and cosine distance measures are used for the similarity analysis process. The hubness relationship is applied to estimate the similarity on high dimensional data environment. The K-Nearest Neighbor (KNN) search scheme is used for the similar transaction identification process. The hubness score is used for the clustering process. The Hubness Proportional K-Means (HPKM) clustering algorithm is implemented with user defined cluster count and automatic cluster count models. The hubness model is also adapted for the Kernal Map Clustering (KMC) and Shared Neighbor Clustering (SNC) algorithms. The automatic cluster count estimation mechanism is integrated in all the clustering techniques. The system can be enhanced with the following features.

- The data partitioning process is implemented under stand alone database environment. The clustering scheme can be improved to support clustering under distributed database environment.
- The clustering model can be adapted to perform clustering on data stream based data source model.
- The system can be integrated with classification schemes to assign labels for the transactions.
- The system can be enhanced to support privacy preserved data clustering process.
- The system can be adapted to support hierarchical clustering process.
- The clustering scheme can be adapted to cluster text documents using hubness relationships.

REFERENCES

1. A.Kaban, “Non-Parametric Detection of Meaningless Distances in High Dimensional Data,” Statistics and Computing, vol. 22, no. 2, 2012.
2. Ali Shahbazi and James Miller, “Extended Sub tree-A New Similarity Function for Tree Structured Data” IEEE Transactions On Knowledge And Data Engineering, Vol. 26, No. 4, April 2014
3. D. Corne and F. Glover, New Ideas in Optimization. McGraw-Hill, 1999.

4. Hoda Mashayekhi, Jafar Habibi, Tania Khalafbeigi, Spyros Voulgaris and Maarten van Steen, "GDCluster: A General Decentralized Clustering Algorithm", IEEE Transactions On Knowledge And Data Engineering, Vol. 27, No. 7, July 2015.
5. Jianbin Huang, Heli Sun and Qinbao Song, "Revealing Density-Based Clustering Structure from the Core-Connected Tree of a Network", IEEE Transactions On Knowledge And Data Engineering, August 2012
6. L. Dee Miller and Leen-Kiat Soh, "Cluster-Based Boosting", IEEE Transactions on Knowledge And Data Engineering, Vol. 27, No. 6, June 2015.
7. M. Radovanovic, A. Nanopoulos and M. Ivanovic, "Hubs in Space: Popular Nearest Neighbors in High-Dimensional Data," J. Machine Learning Research, vol. 11, pp. 2487-2531, 2010.
8. N. Tomasev, D. Mladenic and M. Ivanovic, "A Probabilistic Approach to Nearest-Neighbor Classification: Naïve Hubness Bayesian KNN," Proc. 20th ACM Int'l Conf. Information and Knowledge Management, 2011.
9. Natthakan Iam-On and Chris Price, "A Link-Based Cluster Ensemble Approach for Categorical Data Clustering", IEEE Transactions On Knowledge And Data Engineering, March 2012
10. Nenad Tomasev, Milos Radovanovic and Mirjana Ivanovic "The Role of Hubness in Clustering High-Dimensional Data", IEEE Transactions On Knowledge And Data Engineering, Vol. 26, No. 3, March 2014
11. Qinbao Song, Jingjie Ni and Guangtao Wang, "A Fast Clustering-Based Feature Subset Selection Algorithm for High-Dimensional Data" IEEE Transactions On Knowledge And Data Engineering, January, 2013
12. Qinpei Zhao and Pasi Franti, "Centroid Ratio for a Pairwise Random Swap Clustering Algorithm" IEEE Transactions On Knowledge And Data Engineering, Vol. 26, No. 5, May 2014
13. Shuo Shang, Kai Zheng, Christian S. Jensen, Bin Yang, Panos Kalnis, Guohe Li and Ji-Rong Wen, "Discovery of Path Nearby Clusters in Spatial Networks", IEEE Transactions On Knowledge And Data Engineering, Vol. 27, No. 6, June 2015.
14. Sicheng Xiong, Javad Azimi and Xiaoli Z. Fern, "Active Learning of Constraints for Semi-Supervised Clustering" IEEE Transactions On Knowledge And Data Engineering, Vol. 26, No. 1, January 2014
15. V. Satuluri and S. Parthasarathy, "Bayesian Locality Sensitive Hashing for Fast Similarity Search," Proc. VLDB Endowment, vol. 5, no. 5, 2012.

