ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 8, August 2015

Fuzzy Logic Based Consent Neighbor Clustering In High Dimensional Data

M.Mohanapriya[1], Dr. Antony Selvadoss Thanamani[2]

M.Phil. Research Scholar, Dept. of Computer Science, NGM College, Pollachi, Tamil Nadu, India[1]

Assistant Professor and Head, Dept. of Computer Science, NGM College, Pollachi, Tamil Nadu, India[2]

ABSTRACT: High dimensional data clustering arises naturally in a lot of domains, and have regularly presented a great deal with for usual data mining techniques. In this paper, presents an optimal perspective on the problem of Consensus Clustering in high-dimensional data. The proposed method called "*Fuzzy based and kernel mappings with Consensus Neighboring clustering* (*FKCNC*)", which takes as key measures of correspondence between pairs of data points. The proposed method is to establish a unified framework for *FKCNC* on both supervised and supervised data sets. Also, we examine some important factors, such as the clustering quality and assortment of basic partitioning, which may affect the performances of *FKCNC*. Experimental results on various synthetic and real world data sets demonstrate that *FKCNC* is highly efficient and is equivalent to the state-of-the-art methods in terms of clustering index quality. In addition, *FKCNC* shows high robustness to incomplete basic partitioning with many anomaly values.

KEYWORDS: Fuzzy logic, Consensus Clustering, High dimensionality, kernel mapping, Distance function

I. INTRODUCTION

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: partition, hierarchical, density based, and subspace algorithms.

Clustering often is a first step in data analysis. Many different clustering methods have been developed such as hierarchical agglomerative clustering, mixture densities, graph partitioning, and spectral clustering. Most clustering methods focus on finding a single optimal or near-optimal clustering according to some specific clustering criterion. Consensus clustering can provide benefits beyond what a single clustering algorithm can achieve. Consensus clustering algorithms often: generate better clustering's. Find a combined clustering unattainable by any single clustering algorithm; are less sensitive to noise, outliers or sample variations; and are able to integrate solutions from multiple distributed sources of data or attributes

Consensus clustering (CC), also known as cluster ensemble or clustering aggregation, aims to find a single partitioning of data from multiple existing basic partitioning. It has been widely recognized that consensus clustering can help to generate robust clustering results, find bizarre clusters, handle noise, outliers and sample variations and integrate solutions from multiple distributed sources of data or attributes. Consensus clustering is a combinatorial optimization problem in essence, and in some special cases, e.g., using median partition with the Mirkin distance, it has been proven to be NP-complete.

There are two main contributions of this paper. First, in experiments on synthetic data we show that mapping connectivity is a good measure of point centrality within a high-dimensional data cluster and that major data points can be used effectively as cluster prototypes. In addition, we propose three new fuzzy based clustering algorithms and evaluate their performance in various high-dimensional clustering tasks.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 8, August 2015

II. RELATED WORK

In addressed the problem of combining multiple clustering's without access to the underlying features of the data. This process is known in the text as clustering ensembles, clustering aggregation, or consensus clustering.

Consensus clustering yields a stable and robust final clustering that is in agreement with multiple clustering's. The authors find that an iterative EM-like method is remarkably effective for this problem. To present an iterative algorithm and its variations for finding clustering consensus. In authors introduced a unified representation for multiple clustering's and formulates the corresponding categorical clustering problem. Second a probabilistic model of consensus using a finite mixture of multinomial distributions in a space of clustering's.

A combined partition is found as a solution to the corresponding maximum-likelihood problem using the EM algorithm. Third, we define a new consensus function that is related to the classical interclass variance criterion using the generalized mutual information definition. Finally, we demonstrate the efficacy of combining partitions generated by weak clustering algorithms that use data projections and random data splits. In authors illustrated about the framework for extracting a consistent clustering, given the various partitions in a clustering ensemble. According to the Evidence Accumulation (EAC) concept, each partition is viewed as an independent evidence of data organization, individual data partitions being combined, based on a voting mechanism, to generate a new $n \times n$ similarity matrix between the n patterns. In authors discussed the answer depends on the quality of clustering's to be combined as well as the properties of the fusion method. First, we introduce a unified representation for multiple clustering's and formulate the corresponding categorical clustering problem.

As a result, we show that the consensus function is related to the classical intra-class variance criterion using the generalized mutual information definition. Second, its show the efficacy of combining partitions generated by weak clustering algorithms that use data projections and random data splits. A simple explanatory model is offered for the behavior of combinations of such weak clustering components. In Authors had proposed a novel grouping method in this paper, which stresses connectedness of image elements via mediating elements rather than favoring high mutual similarity.

This grouping principle yields superior clustering results when objects are distributed on low-dimensional extended manifolds in a feature space, and not as local point clouds. In addition to extracting connected structures, objects are singled out as outliers when they are too far away from any cluster structure. The objective function for this perceptual organization principle is optimized by a fast agglomerative algorithm. In authors considered the Consensus clustering and semi-supervised clustering are important extensions of the standard clustering paradigm. Consensus clustering (also known as aggregation of clustering) can improve clustering robustness, deal with distributed and heterogeneous data sources and make use of multiple clustering criteria.

Semi-supervised clustering can integrate various forms of background knowledge into clustering. The authors have to show how consensus and semi-supervised clustering can be formulated within the framework of nonnegative matrix factorization (NMF). Its show that this framework yields NMF-based algorithms that are: extremely simple to implement; provably correct and provably convergent. It conducts a wide range of comparative experiments that demonstrate the effectiveness of this NMF-based approach.

III. PROPOSED ALGORITHM

A. *Pre-processing*

The data pre-processing is incomplete the lacking attribute values, lacking certain attributes of interest, or containing only aggregate data. The pre-processing method follows the data conversion approach that facilitates of data clustering. Our approach, called optimal association link, strives to extract the underlying structure or sub-concepts of each raw attribute automatically, and uses the orthogonal combination of these sub-concepts to define a new, semantically richer, space.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 8, August 2015

The supporting labels of each point in the original space determine the position of that point in the transformed space. The labels are prone to uncertainty inherent in the original data and in the initial extraction process, a combination of labelling schemes that are based on different measures of uncertainty will be presented.

B. Fuzzy Logic Algorithm

The objective function of Fuzzy logic is to discover the data points as cluster centroid has to the optimal membership Link for estimating the centroids, and typicality is used for improving the disagreeable effect of anomalies. The function is composed of two expressions:

- The first is the fuzzy logic function and uses a Euclidean distance exponent,
- The second is fuzzificaion weighting function exponent; but the two coefficients in the objective function are only used as exhibitor of membership link and typicality.

The fuzzy aggregation assigns data points to c partitions by using optimal memberships. Let $X = \{x_1, x_2, x_3\ldots x_n\}$ denote a set of data points to be portioned into c clusters, where x_i ($i = 1, 2, 3 \ldots n$) is the data points. The objective function is to discover nonlinear relationships among the data, kernel (root) methods use embedding linking's that connectivity features of data to new feature spaces. The proposed technique Fuzzy based kernel mapping (FKM) algorithm is an iterative clustering technique that minimizes the objective function.

Given an dataset, $X = \{x_1\ldots x_n\}\subset Rp$, the original *KFCNC* algorithm partitions X into c fuzzy partitions by minimizing the following objective function as,

$$J(w,U,V) = \sum_{i=1}^{c}\sum_{k=1}^{n} u_{ik}^{m} \| x_k - v_i \|^2 \tag{1}$$

Where c is the number of clusters and selected as a specified value, n the number of data points, u_{ik} the membership link of x_k in class i, satisfying the, $\sum_{i=1}^{c} u_{ik} = 1$ m the quantity scheming clustering fuzzification, and V the set of cluster canter's or prototypes ($v_i \in Rp$).

C .Consensus Neighbouring clustering Algorithm

The Consensus kernel Neighbouring clustering algorithm works link mapping passing among the each data points. The Consensus kernel Neighbouring Clustering (CKNC) method involves sub-sampling from a set of data items, such as incomplete dataset, and determines clustering's of specified cluster average counts (k). Then, pair wise (Euclidean Distance) values, the proportion that two items taken the same cluster out of the number of times they occurred in the same samples are intended and stored in an impartial consensus matrix for each k. The consensus template (CKNC) is reviewed in several graphical displays that enable a user to make a decision upon a logical cluster number and membership link.

Our proposed algorithms can be applied to any group of clustering's, including clustering's in which each character clustering may have different numbers of clusters. In addition, our new algorithms conform to two constraints (must link and cannot link) which most other consensus clustering algorithms follow.

IV CONCLUSION AND FUTURE WORK

This paper proposed the Fuzzy based and kernel mappings with Neighbor clustering (FKCNC) algorithm for the consensus clustering problem. Our proposed algorithm is in core variations of fuzzy based Consensus Neighboring clustering algorithm using different weight measures applied to the vector of base-level clustering's. In future work, we intend to enhance the same algorithm by relaxing the two constraints of the allocating membership space linking and resource Allocation. First, we want to use consensus to determine the best number of clusters in the consensus clustering. Second, we will develop the experimental methods for bi-level optimization to control the overlapping features of the result data.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 8, August 2015

REFERENCES

1. A. Strehl and J. Ghosh, "Cluster ensembles—A knowledge reuse framework for combining partitions," J. Mach. Learn. Res., vol. 3, pp. 583–617, 2002.
2. N. Nguyen and R. Caruana, "Consensus clusterings," in Proc. IEEE Int. Conf. Data Mining, 2007, pp. 607–612.
3. A. Topchy, A. Jain, and W. Punch, "Clustering ensembles: Models of consensus and weak partitions," IEEE Trans. Pattern Anal. Mach. Intell., vol. 27, no. 12, pp. 1866–1881, Dec. 2005.
4. A. Fred and A. Jain, "Combining multiple clusterings using evidence accumulation," IEEE Trans. Pattern Anal. Mach. Intell., vol. 27, no. 6, pp. 835–850, Jun. 2005.
5. A. Topchy, A. Jain, and W. Punch, "Combining multiple weak clusterings," in Proc. 3rd IEEE Int. Conf. Data Mining, 2003, pp. 331–338.
6. R. Fischer and J. Buhmann, "Path-based clustering for grouping of smooth curves and texture segmentation," IEEE Trans. Pattern Anal. Mach. Intell., vol. 25, no. 4, pp. 513–518, Apr. 2003.
7. T. Li, D. Chris, and I. J. Michael, "Solving consensus and semi supervised clustering problems using nonnegative matrix factorization," in Proc. 7th IEEE Int. Conf. Data Mining, 2007, pp. 577–582.