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. 4, Issue 2, February 2016

CFKM: An Optimal Consensus Clustering Using Fuzzy Based Kernel Mapping Algorithm

P.Sathyasri, Prof. K.Jeyalakshmi

Research Scholar, PG & Research Department of Computer Science, Hindustan College of Arts and Science, Coimbatore, Tamilnadu, India

Associate Professor, PG & Research Department of Computer Science, Hindustan College of Arts and Science, Coimbatore, Tamilnadu, India

ABSTRACT: Clustering is the application of data mining techniques to discover patterns from the datasets. Thus, when multi-temporal images are considered, they allow us to detect many possible differences in HS images. This paper proposed a novel approach to measures the data dissimilarity data elements in high dimensional data clustering. Clustering becomes difficult due to the increasing sparsity of such data, as well as the increasing difficulty in distinguishing distances between data points.The algorithm called "Fuzzy neighboring consensus clustering based on kernel Fuzzification degree (FNCKF)", which takes as key measures of correspondence between pairs of data points. The proposed method is to establish a unified framework for on both supervised and unsupervised data sets. Also, we examine some important factors, such as the clustering quality and assortment of basic partitioning, which may affect the performances of Fuzzy framework.Experimental results obtained on synthetic and real datasets to demonstrate the effectiveness of the clustering quality.

KEYWORDS: Consensus Clustering, nearest neighbours, Fuzzy logic, kernel mapping..

I. INTRODUCTION

Consensus clustering (CC), also known as cluster ensemble or clustering aggregation, aims to find a single partitioning of data from multiple existing basic partitioning (BPs) [1]. 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 [2]. 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 [3].

Clustering is a division of data into groups of similar objects. Each group, called cluster, consists of objects that are similar between themselves and dissimilar to objects of other groups. Representing data by fewer clusters necessarily loses certain fine details (akin to lossy data compression), but achieves simplification. It represents many data objects by few clusters, and hence, it models data by its clusters. Data modelling puts clustering in a historical perspective rooted in mathematics, statistics, and numerical analysis. From a machine learning perspective clusters correspond to hidden patterns, the search for clusters is unsupervised learning, and the resulting system represents a data concept. Therefore, clustering is unsupervised learning of a hidden data concept and the fuzzy clustering is the most widely used technique for hidden data analysis.

Kernel Principal Component Analysis (KPCA) is a clustering method (as opposed to an algorithm according to the terminology that uses KPCA to do such a non-linear mapping. The strategy is to use the "kernel trick" to implicitly map the feature vectors to a higher dimensional space where hopefully the cluster structure of the data will be easily found by simple non-spectral clustering algorithms. KPCA can also be applied to both reconstruction and de-noising of data.

This clustering also can be done via density-based methods or distance-based methods. Distance-based methods have two weakness which leads to be not suitable for spatial data clustering, first they need a number of clusters as an

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. 4, Issue 2, February 2016

input and second they allocate all objects to the clusters and never identify noises. By the way there are some related works which firstly transform spatial or spatio-temporal data to same length multi-dimensional vectors and then apply a generic clustering algorithm like k-mean on the data.

II. RELATED WORK

In [1] authors proposed the problem of combining multiple partitionings of a set of objects into a single consolidated clustering without accessing the features or algorithms that determined these partitionings. We first identify several application scenarios for the resultant 'knowledge reuse' framework that we call cluster ensembles. The cluster ensemble problem is then formalized as a combinatorial optimization problem in terms of shared mutual information. In addition to a direct maximization approach, we propose three effective and efficient techniques for obtaining high-quality combiners (consensus functions). In [2] authors discussed the problem of combining multiple clustering's without access to the underlying features of the data. This process is known in the literature 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.In [3] authors introduceda 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 intra-class variance criterion using the generalized mutual information definition. Finally, to demonstrate the efficacy of combining partitions generated by weak clustering algorithms that use data projections and random data splits.In [4] authors illustrated a simple framework for extracting a consistent clustering, given the various partitions in a clustering ensemble. According to the 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 × n similarity matrix between the n patterns.In [5] authors discussed the data set can be clustered in many ways depending on the clustering algorithm employed, parameter settings used and other factors. Can multiple clustering's be combined so that the final partitioning of data provides better clustering? 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. In [6] authors 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.

III. PROPOSED ALGORITHM

The proposed architecture accepts the user parameters as input which contains the MATLABR2010a simulation where the Fuzzy based kernel mappings with adaptive Fuzzification algorithm is applied to the datasets. This architecture in figure 1 follows a path from the start to end state. The users initialize the number of *k*-value as cluster parameters in which the clustering process is to be evaluated.

A. *Data preprocessing:*

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. 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. *Correlation of Fuzzy based Data Clusters*

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 outliers.

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. 4, Issue 2, February 2016

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.

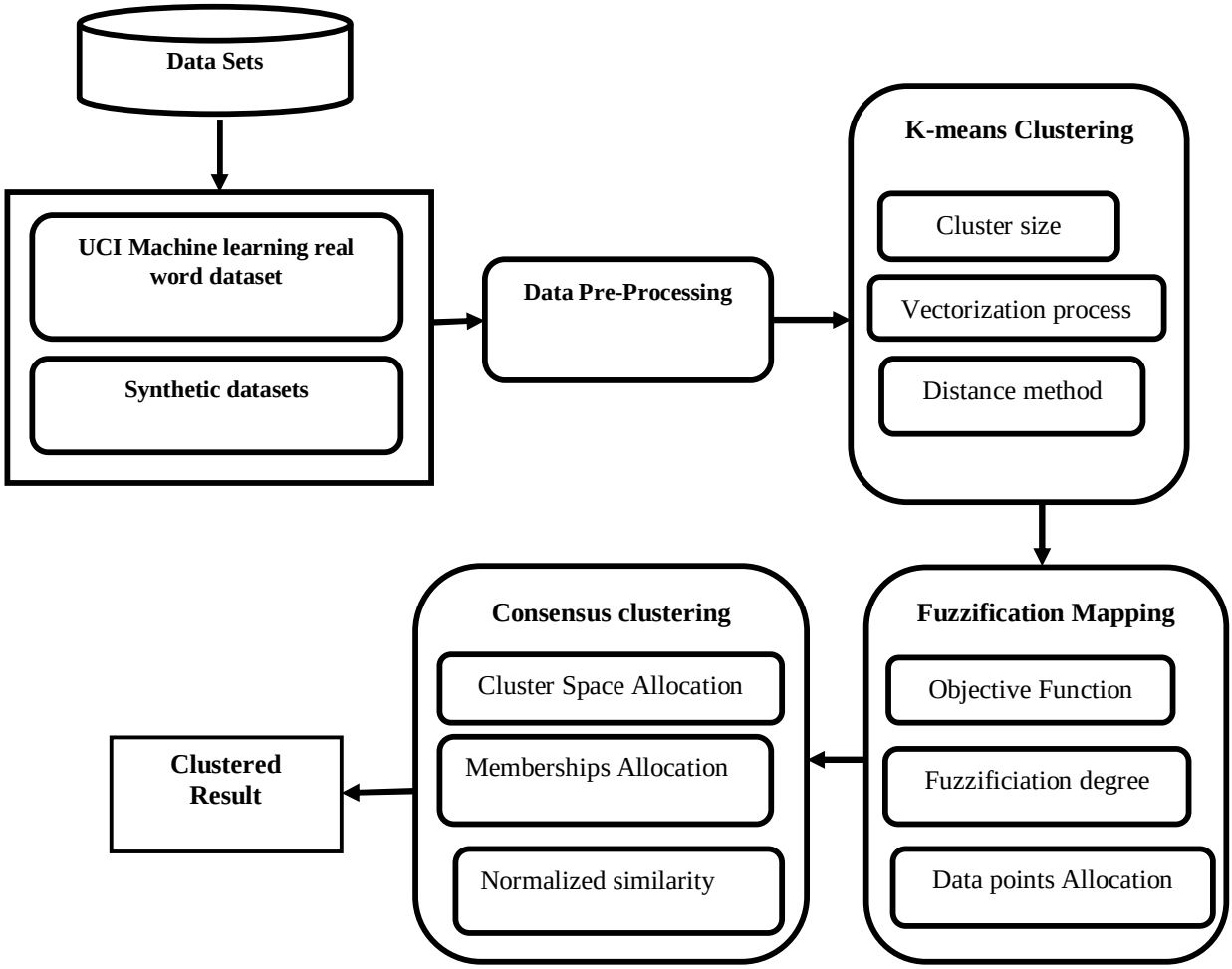

Figure 1 Architecture of Proposed System

C. *Kernel based Fuzzification Connectivity*

The degree of branching can be specified with a kernel k that is directly applied to the similarity matrix. It is shown that the generated clusters can still be monotonic depending on the used linkage measure even though the induced dissimilarity measures are no longer ultra metrics. Using the pair-wise merged clusters; an additional shrinking process is proposed to generate topic related groups with more than two cluster elements.

- The process of determining the degree to which a value belongs in a kernel set
- The value returned by a shared-Neighbor cluster
- Most variables in a hub-based system have multiple data points attached to them
- Kernel mapping that variable involves passing the crisp value through each neighbor attached to that value

Here dataset is an object matrix. Clusters are groups of similar data elements. Resemblance coefficient represents the degree of similarity and non similarity between the items. The main aim of clustering analysis is identify and quantification of these architecture elements. Identifying the membership and location center of the clusters is main

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. 4, Issue 2, February 2016

process in the cluster analysis. Some time data in the cluster is well packed. But due to the complex nature of the components the data may not be packed well in the clusters. Some of the elements lie outside the cluster region.

D. *Consensus-Neighbor clustering Algorithm*

The Consensus-neighbour clustering algorithm works message passing among data points. Each data points (hubs) receive the availability from others data points (from pattern) and send the responsibility message to others data points (to pattern). Sum of responsibilities and availabilities for data points identify the cluster patterns.
The high-dimensional data point availabilities *A (i, k)* are zero: *A (i, k)* = 0, *R (i, k)* is set to the input similarity between point *i* and point *k* as its pattern, minus the largest of the similarities between point *i* and other candidate patterns.

E. Allocation of Data Memberships and Cluster Space

This approach computes two kinds of messages exchanged between data points. The first one is called "responsibility" r(i, j): it is sent from data point i to candidate exemplar point j and it reflects the accumulated evidence for how well-suited point j is to serve as the exemplar for point i. The second message is called "availability" a(i, j) it is sent from candidate exemplar point j to point i and it reflects the accumulated evidence for how appropriate it would be for point i to choose point j as its exemplar. At the beginning, the availabilities are initialized to zero: a (i, j) = 0.

IV. EXPERIMENTAL RESULTS

For evaluating the proposed work, three types of consensus clustering methods, namely the K-means-based algorithm, the graph partitioning algorithm (GP), and the hierarchical algorithm (HCC), were employed for the comparison purpose. GP is actually a general concept of three benchmark algorithms: CSPA, HGPA and MCLA [1], which were coded in Matlab and provided by Streh. HCC is essentially an agglomerative hierarchical clustering algorithm based on the so-called co-association matrix. It was implemented by ourselves in MATLAB following the algorithmic description in [4]. We also implemented fuzzy based Consensus clustering in MATLAB, which includes ten utility functions, namely Utility Category (Uc) for particular group selection, Utility Shannon Entropy (UH) for performing predictable data points, Utility Cosine similarity (Ucos) for calculating similarity function, Utility kull-back Leibler Divergence (UL5) and UL8 for measure of the difference between probability distributions L5 and L8, and their corresponding normalized versions (denoted as NUx). The Rand index [7] or Rand measure in statistics, and in particular in data clustering, is a measure of the similarity between two data clusterings. A form of the Rand index may be defined that is adjusted for the chance grouping of elements; this is the adjusted Rand index.

$$R_n = \frac{a+b}{a+b+c+d} = \frac{a+b}{\binom{n}{2}}$$

Table 1: Clustering Quality on the UCI machine learning datasets

K	**2**	**4**	**6**	**8**	**10**	**12**	**14**	**16**
U_c	0.0556	0.506	0.111	0.1212	0.1488	0.3767	0.3122	0.0352
U_H	0.4296	0.0661	0.1476	0.4628	0.4039	0.5702	0.4743	0.4296
U_{cos}	0.111	0.4359	0.7352	0.5814	0.5322	0.0421	0.1448	0.3647
NU_H	0.5470	0.7069	0.0537	0.1336	0.4938	0.3619	0.4093	0.2412
FNCKF	0.5582	0.6894	0.5992	0.5863	0.6321	0.5769	0.4956	0.4723

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. 4, Issue 2, February 2016

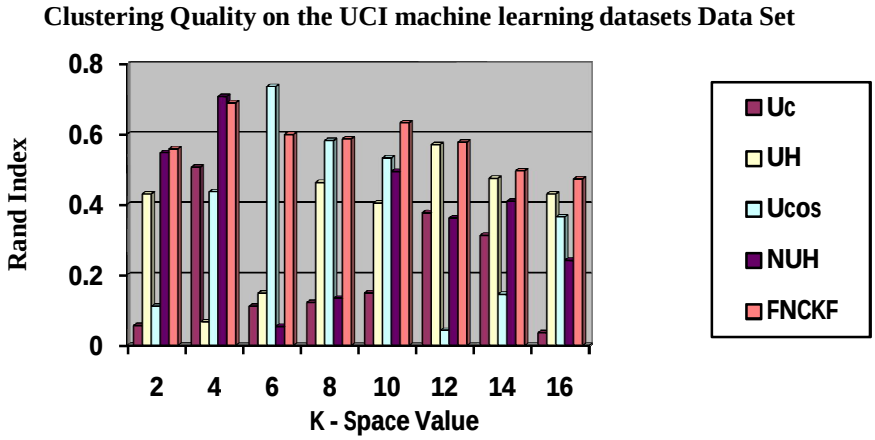

Fig 2:Performance Evaluation Chart

The above Fig 2 shows the performance evalution chart of the proposed algorithm(FNCKF).

V. CONCLUSION AND FUTURE WORK

In this paper, proposed a fuzzy based kernel Fuzzification to approximate local data centers is not only a feasible option, but also frequently leads to improvement over the centroid-based approach. The proposed the Fuzzy Neighbouring Consensus clustering based on kernel Fuzzification degree (FNCKF) algorithm for the consensus clustering 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 baseline on both synthetic and real-world data

In future work, we intend to enhance the Clustering algorithm to apply to the real data sets we need to refine the adjacency matrix by the hard-thresholding, say, and this area is worth pursuing as future research.

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 International Conference of 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 International Conference of 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. Z. Lu, Y. Peng, and J. Xiao, "From comparing clusterings to combining clusterings," in Proc. 23rd AAAI Conference Artif.Intell., 2008, pp. 361–37.