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 6, June 2015

Improving Co-Clustering Efficiency for Hetrogenous Fusion in Multimedia Data

Aparna A P, Neethu Susan Jacob

Pursuing M.Tech, Dept. of CSE, Caarmel Engineering College, MG University, Kerala, India

Assistant Professor, Dept of CSE, Caarmel Engineering College, MG University, Kerala, India

ABSTRACT: In order to retrieve information or web document, it is required to manage or scale very large repositories in order to get the information of demand. So clustering techniques are employed in order to enhance searching.. An efficient co-clustering algorithm GHF-ART is used here (Generalized form of Heterogeneous Fusion Adaptive resonance Theory). The main advantages of this technique are that it uses a multiple channel and each channel is feed with different data pattern so that efficiency increases. The main advantages of this algorithm are that it have strong noise resistance, channel weighting factors are adaptive, low computational overhead and clustering technique are incremental. Since it has a lot of advantage, this mechanism yet has certain drawbacks. A novel algorithm known as sparse graph based discriminant analysis block-structured similarity matrix can be used for feature selection which effectively selects the feature from a large set of dataset. That means this algorithm can be used for dimensionality reduction.

KEYWORDS: Clustering, ART, heterogeneous fusion, Semi-supervised learning.

I. INTRODUCTION

Clustering, which aims to efficiently organize the data set, is an old problem in machine learning and data mining community. The existing algorithms like K-means , FuzzyC means algorithm only shows desired result for clustering homogeneous data, i.e. the data points are of a single kind. For example it can be either documents or images. However the dataset, for any real world application which can be seen today contains heterogeneous data. The best example of such category is flicker dataset. The major challenge faced is that the different types of data points are dependant of other. Usually there exist close relationships between different data types, for the traditional algorithm to make use of these relationship is a challenging function. They will not provide most precise result. Co-clustering algorithm mainly utilizes the relationship present between heterogeneous type of data. The heterogeneous data co-clustering technique can be applied to both text and image domain. It accepts a technique which combines the equitable activity of every feature vectors, so that it reduces the global cost. For co-clustering the web documents data, existing algorithms have three challenges, they are: meta-information that are available is usually precise so that the extracted tags or the categories cannot be adequatively weighted by conventional data mining routine such as term frequency-inverse document frequency (tf-idf). Next the feature vectors weight in the function depends on various experimental settings lead to unsounded conclusion. Finally, there is high computational overhead for ensuring the concurrence which requires a repeating process.

So it is clear that the existing methods cannot be applied to datasets, which contains thousand of documents . The main disadvantage with the existing techniques is that it cannot used to scale very large dataset and for yielding results it takes more time. The problem statement is that it defines the theme discovery of web multimedia data as a heterogeneous data co-clustering problem, which identifies the semantic categories of data patterns through the fusion and recognition of multiple types of features. So in order to address the above mentioned problem GHF-ART can be employed. Adaptive Resonance Theory is a neural network theory for cognitive information processing. ART performs unsupervised learning in which it models the cluster as memory prototypes and it encodes every input pattern through a two-way similarity measure[7]. This is similar to how our brains capture, recognize and memorize various details and information of objects and events[8]. The input pattern can be intended to be a member of a cluster when the difference

DOI: 10.15680/ijircce.2015.0306156

between input pattern and the winner does not exceed a threshold value known as vigilance parameter. ART has the advantages of fast and robust learning. ART shows very high noise resistance.

II. PROPOSED SYSTEM

Generalized Heterogeneous Fusion ART is an extension of HF-ART(Heterogeneous Fusion Adaptive Resonance Theory). HF-ART encompass two channels only whereas GHF-ART have multiple channels and each channel is capable of receiving different data patterns, so that it can be used for clustering many kinds of data. So that it is able to assimilate different features across various channels. GHF-ART can be considered as a self-establishing neural network which performs the co-clustering of the heterogeneous data. In the GHF-ART clustering step, it partitions the category space into various cluster regions which are done by incrementally learning the cluster templates from the input patterns and thus identifying the key features in the document taken.

The summary of the GHF-ART algorithm is given below.

- Based on the user preferences first of all have to generate a predefined cluster for the initializing the network. If there are no any knowledge available from the user then have to create an uncommitted cluster having all weight vectors containing 1's.
- Feature extraction and feature selection.
- Calculate choice function for every category in the category field
- Then it is required to calculate the winner cluster having the largest value for the choice function.
- Calculate the match function
- A resonance occurs if the winner clusters meet the expectation of the vigilance criteria, leading to the learning steps. If not, a new cluster will be selected
- If the winner cluster appeases the criteria then the weight vectors are updated.
- Algorithm gets terminated if there is no input pattern

In the below paragraph we can detailise the algorithm

Data Pre-processing

For doing this algorithm it is needed to pre processes the data. Data can be of text or images. In this work we are contemplating or dealing with images only. It is needed to pre-process the images for boosting or augmenting it. Here it is needed to reconstruct the image from one colour space to another. The image has been first converted to standard continuous sRGB space. Then by using standard sRGB-to-XYZ conversion, it is able to convert the pixel values to XYZ

Extracting Features

First of all we have to extract the feature that means feature extraction have to be done. Here we are considering the images features only. There are many feature for the images. From those images we are extracting the dominant five features. That is here we are considering color histogram, edge direction, edge histogram, gabor wavelet and GLCM features. For extracting those features we are using different filters. For making the input values of the interval [0, 1] min-max normalization is used.

Feature Selection

After extracting the features then the next step we are doing feature selection for dimensionality reduction. Here it can be done with the help of algorithm that is sparse graph based discriminant analysis block-structured similarity matrix(BSGD). Since there are many images in a data set we are selecting the best features with the help of this algorithm.

Cluster Assignment

In this step we are finding out on which cluster the particular input pattern belongs and also we are evaluating the matching between the input pattern and the template pattern of the cluster. For doing this there are two steps they are choice function and match function[7]. In the category choice step, a choice function is used for understanding the similarity between the input pattern which we have given and the template pattern of each cluster. For finding out the choice function it is necessary to perform AND operation between input vector and the weight vectors. The cluster with the highest choice function will be selected as the winner cluster. After identifying the winner cluster a match function is used to found out the similarity between the input pattern and the winner. The match function for the entire channel is calculated. For the entire feature channel it has to meet the expectation of the vigilance criteria. Then only the input pattern is categorized as winner cluster. If it is not satisfying the vigilance criteria then, a new cluster will be selected from remaining clusters. The main principle of clustering is that it should maintain high intra cluster similarity and low inter cluster similarity. Pattern with the small distance be selected as the winners. After obtaining the particular category on which the input pattern belongs, labels will be assignments according to the user preferences.

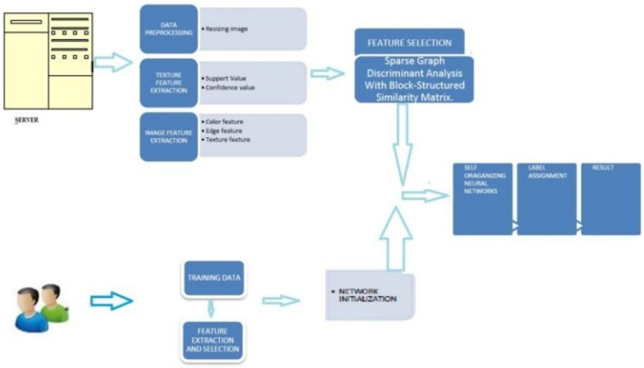
Adjusting parameters

The vigilance parameter and contribution parameter have a greater effect on clustering results. When it cannot identify a matching cluster or the winner cluster, the vigilance parameter ρ is updated with a small value of epsilon. The value of epsilon is greater than zero**.** During the category choice function, weighting factor for the each feature channel are specified by the contribution parameter[]. Higher weight is provided for those channels which are robust in differentiating the class of patterns. So it is necessary to scale those feature channels by learning from the input pattern. The difference between each feature vector within an intra cluster is calculated, they are then be used to find out the overall difference of feature vector. Thus the robustness measure can be expressed in terms of the difference between the feature vectors in a cluster. If the robustness measure becomes one then it is clear that the feature is belonging to the same class. By this way the contribution parameter can be tuned during the clustering process.

System Architecture

The below diagram shows the system architecture**.** User preferences are needed for the network initialization. When a new data comes they can be either a text or images, the features are extracted from them . The main advantages is that it use a multiple channel and each channel can receive different data pattern. Keyword extracted are from the text having support value. For images extraction it uses different filters. Here texture features , edge histogram, edge direction features, colour features and GLCM features are extracted out. The texture features can be extracted with the help of wavelet texture feature. Shape and edge features are extracted with the help of canny or sobel edge detectors. For extracting the colour features RGB histogram can be used.

DOI: 10.15680/ijircce.2015.0306156



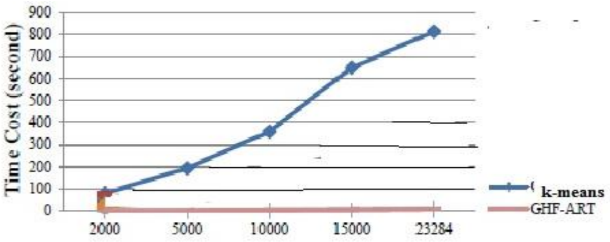


Figure 1:System Architecture

After extracting the features they are placed in various categories. Since it is a neural network, user previous knowledge are used for initializing the network. When an incoming data comes, the self organizing neural network find the appropriate cluster, thus preserving the maximum intra cluster similarity. Then the label will be assigned to the winner cluster.

III. RESULTS

For evaluating the performance of the proposed system, the NUS-WIDE dataset has been taken. This is implemented on MATLAB. The figure below shows the clustering result of the image taken from the NUS-WIDE data set. The distance having lower value will be selected as the winner. Here we can see first ten clustered results. In that the first image has the lower distance value and that can be regarded as the best matching image.

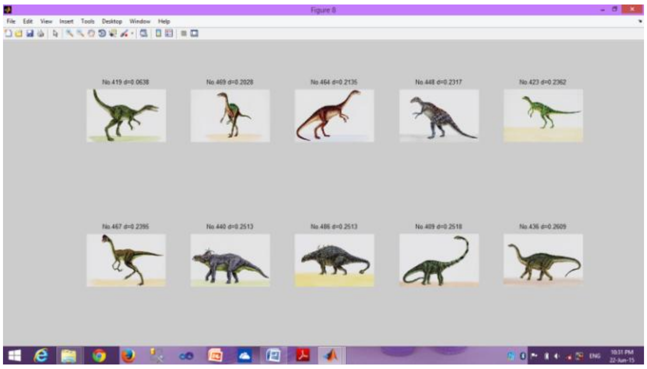


Figure 2 clustered results

Parameters values are taken as 0.01, 0.6 and 0.1 for choice parameter, learning parameter and vigilance parameter respectively. The below figure shows the cost of two algorithm that is K-means and our proposed algorithm on NUS-WIDE data. The GHF-ART with the dimensionality reduction reduces the time and cost of clustering.

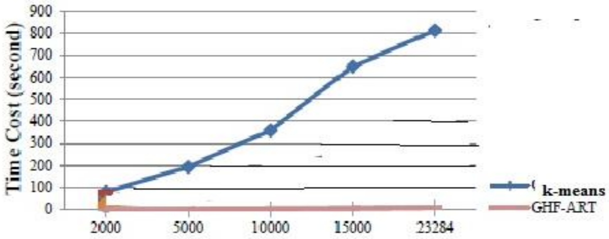


Figure 3. Performance evaluation

IV. CONCLUSION

For the fast and robust clustering of multimedia documents, a novel heterogeneous data co-clustering algorithm known as Generalized Heterogeneous Fusion ART (GHF-ART) have been developed. Here in GHF-ART there are multiple channels and each channel can receive different type of data patterns. Dimensionality reduction based on BSGD algorithm is employed for feature selection for improving the co-clustering efficiency. GHF-ART has many advantages from the existing techniques. It shows very high noise resistance and also it adjusts its parameters adaptively. The proposed algorithm saves time and cost of clustering.

REFERENCES

[1] T. Jiang and A.-H. Tan, “Learning Image-Text Associations,” IEEE Trans. Knowledge and Data Eng., vol. 21, no. 2, pp. 161-177, Feb. 2009.

[2] T. Jiang and A.-H. Tan, “Discovering Image-Text Associations for Cross-Media Web Information Fusion,” Proc. 10th European Conf. Principle and Practice of Knowledge Discovery in Databases (PKDD), pp. 561-568, 2006.

[3] X. Hu, N. Sun, C. Zhang, and T.-S. Chua, “Exploiting Internal and External Semantics for the Clustering of Short Texts Using World Knowledge,” Proc. ACM Conf. Information and Knowledge Management, pp. 919-928, 2009.

[4] L. Meng and A.-H. Tan, L. Nguyen, K. Woon, and A.-H. Tan, “A Self-Organizing Neural Model for Multimedia Information Fusion,” Proc. Int’l Conf. Information Fusion, pp. 1-7, 2008.

[5] S. Harabagiu and F. Lacatusu, “Using Topic Themes for Multi- Document Summarization,” ACM Trans. Information Systems, vol. 28, no. 3, pp. 1-47, 2010

[6] D. Cai, X. He, Z. Li, W. Ma, and J. Wen, “Hierarchical Clustering of WWW Image Search Results Using Visual, Textual and Link Information,” Proc. 12th Ann. ACM Int’l Conf. Multimedia, pp. 952-959,

[7] L. Meng and A.-H. Tan, Dong Xu, “Semi-Supervised Heterogeneous Fusion for Multi media Data Co-Clustering,” Trans. Knowledge and Data Eng., vol. 26, no. 9, Sept 2014

[8] A.-H. Tan, “Adaptive Resonance Associative Map,” Neural Networks,vol. 8, pp. 437-446, 1995

BIOGRAPHY

Aparna A P received her Bachelor of Engineering in Computer Science and Engineering from Anna University, Chennai, 2013. At present, she is pursuing M.Tech in Computer Science and Engineering at Caarmel Engineering College, Kerala, Affiliated to MG University. Her research interests include Data Mining, Big Data processing, Cloud Computing.

Neethu Susan Jacob received her M.Tech in Computer Science and Engineering from Viswajyothi College of Engineering & Technology, Muvattupuzha, Affiliated to MG University and B.Tech in Computer Science and Engineering from Caarmel Engineering College,Ranni-Perunad, Affiliated to MG University. Currently, she is working as the Assistant Professor in Department of Computer Science and Engineering, Caarmel Engineering College. Her research interests are Network Security, Data Mining.

DOI: 10.15680/ijircce.2015.0306156