ISSN No:-2456 – 2165

Omnipresence of Cluster Analysis for Optimal Solutions

Sykam V Narendra Kumar
Department of Computer Science and Engineering
Lovely Professional University,
Phagwara, India
sykam.narendra@gmail.com

Ravishanker
Department of Computer Science and Engineering
Lovely Professional University,
Phagwara, India
ravishanker20@gmail.com

Abstract—Clustering can be observed as an omnipresent technique that spread its scope in different areas such as pattern recognition, machine learning, energy optimization, cognitive radio networks, etc. With the enormous increase in the data being produced by huge number of applications daily which needs to be analyzed for efficient results, it should be organized and computed using different data mining techniques. Clustering technique plays a significant role in contributing its usage to obtain a global solution. Handling of data in dynamic environment is to be done properly so as to ensure effective communication. Data correlation is a common issue in a sensor network which needs to be resolved. This paper explores significant clustering techniques that were applied in various domains.

Keywords—Clustering; Dynamic-Environment; Time-Evolving; Social Network; Biological Network; Communication

I. INTRODUCTION

Clustering in a network domain is for generating a stable network backbone which aims at efficient resource utilization and load balancing where a large network is divided into various connected subnets comprising of cluster heads (CHs) and cluster members [1]. CHs play a major role in managing the network topology, routing messages form one cluster to the other. Centralized clustering algorithms uses central network utilities where as a distributed clustering algorithms does not.

Clustering deals with massive, high dimensional data that allows determining groups of similar data. Cluster analysis has a large scope in game analytics [2] to identify the patterns in behavior of the customers. In Smart Grid applications [3], time-series clustering is used on sensor data to analyze the daily data patterns. Clustering aims at partitioning a number of unlabeled data into various clusters depending upon the similarity.

Incremental clustering techniques [4] use the firstly available data to form the first cluster and consequently deciding over the new data to form new clusters by adding it to the existing based on similarity estimation.

This paper presented few clustering techniques that were applied in some domains which were succeed by achieving the two requirements i.e., efficiency and accuracy of the system without compromising the system's actual behaviour.

II. APPLICATIONS OF DIFFERENT CLUSTERING TECHNIQUES

A. MSF Clustering for Handling Dynamically Changing Text Information Stream

Introduction: Clustering aims at determining an inherent structure in the data and introduces this structure as set of groups. The entities within each group exhibits high degree of similarity. Typically the document clustering algorithms can be categorized into two significant clustering techniques: partitioning and hierarchical, in which partitioning technique (e.g., K-means algorithm) is preferable due to its low computational requirements. Disadvantage in K-means algorithms is that it requires a prior knowledge of the probable number of clusters for document collection. This problem then resolved by many biologically inspired approaches such as Genetic Algorithm (GA), algorithms based on Swarm Intelligence (SI), Self-Organizing Maps (SOM), etc.

Information retrieving and real-time analyzing of dynamically changing text information stream is not an easy task for a single computer to perform. There exist many clustering techniques which could only solve the problem of analyzing the static document collection. But the real problem of such techniques lies in the computation resource and the more computation time for producing the accurate result. Document clustering is a basic operation used for unsupervised document organization, automatic topic extraction as well as information retrieval by providing a structure for organizing a huge amount of text for effective

browsing and searching. Document clustering can be made easy with partitioning clustering techniques such as K-means algorithm, Ant clustering algorithm, etc, where the latter is preferable to the former. But due to lack of self mobility of clustered data objects in Ant clustering algorithm which consumes more computation time for the job to be done by less number of ant agents that responds very slow to the dynamic changes, the author proposed a dynamic reactive flocking clustering algorithm called MSFC algorithm [5] which extends the Flocking model, is a distributed multi-agent flocking approach where each document is considered as a bird agent, reacts quickly to the changes of documents contents to resolve the issues of load balancing and state synchronization thus proved to be scalable and pro-active to the dynamic information change. MSF model added a new rule called "feature resemblance" to the Flocking model of three rules. This feature enables each entity (represented as 'boid') in the system to sense a boid within the limited sensing range and then moves in the computing space according to the four rules. MSFC algorithm assumes each document vector as a boid in the distributed virtual computing space. The speciality of MSFC algorithm is that there is no centralized controller thereby increasing the clustering speed of distributed MSFC algorithm.

There are two basic communication schemes that each boid agent follows to exchange and update the state information with and on various processors while moving in the distributed virtual space. One is broadcast scheme in which every agent broadcast its state information to all other agents of same node as well as different nodes thus identifies the neighbor boids and calculates next moving velocity. The other is the location proxy agent scheme where every agent will send its state information only to the location proxy agent of the same node but not of different nodes. These schemes enable the agents of each node to have a global view of the complete system. The limitation of the broadcast communication scheme is that it could not serve properly for a large number of boid agents present in the system. Because higher the number of boid agents, higher will be the execution time. The communication complexity of broadcast scheme is O((n-1)!) whereas O(2n) is to location proxy agent communication scheme. This limitation of broadcast scheme might be resolved by making an efficient node as a cluster head which has information of all its member nodes. Then broadcasting of state information could be done by the cluster heads first to reduce the traffic overhead as well.

In the distributed environment, processing nodes are responsible for load balancing to ensure that each node is assigned the same workload for which MAB implementation is used for balancing the distributed workload, and synchronization of status information is to ensure the regularity and maintenance of casuality among the system, and finally reducing the communication overhead among the nodes.

Unlike Ant clustering algorithm, where small number of ant agents carries isolated data objects does not communicate with each other and hence cannot know the target node to drop the data object consumes more computation time in moving randomly and searching the required target node, MSFC algorithm's bird agents follows the four rules to generate a global behavior of complete flock and subsequently the agents those carrying similar documents will be merged as one document cluster separated from those which have different behavior. The heart of the MSFC algorithm is the searching mechanism that help bird agents, which moves continuously in the distributed virtual space, to quickly regenerate clustering result and reactive to the dynamic changes of any single document that are being fed into the system continuously.

B. Affinity Propagation Clustering in WSN

Time critical nature of emerging applications of WSNs demands QoS requirements for an effective discovery process. Depending up on the need of information, a sensor can use either proactive or reactive strategy for information discovery to achieve maximum network lifetime as well as minimum delay in query processing. Data interference or correlation is a common problem in a WSN which can be resolved by affinity propagation clustering technique [6] so as to achieve high information dissemination. This technique aims at obtaining a highly possible affinity among existing objects as same group and the varying objects as different group. "Exemplars" play key roles which are used as data fusion points which help in achieving the goal of optimizing the energy of individual sensor node.

In such dynamic environment, as energy consumption is a major concern, two types of messages that are exchanged between the sensors are used for the purpose. The responsibility message that a sensor sends to the data fusion point, indicates the suitability of a sensor node for serving as data fusion point by taking into consideration of remaining potential data fusion points in the system. The other message is the availability message that a sensor receive from the data fusion point ensuring that it could be chosen as its data fusion point by ignoring others because of their non-support from other sensor nodes. The self responsibility r(j,j) is defined as sensor's residual energy. The self availability a(j,j) indicates the energy level of the sensor node j as well as collected evidence that it is a local aggregation or data fusion point, based on positive responsibilities collected from other sensors.

Network operation involves reduction of energy level of every local aggregation point which leads to decrease in the network lifetime. The affinity propagation algorithm helps in solving this with its property of message-passing. Whenever there is a reduction in energy level of data fusion points, the algorithm identifies this dynamic change and thereby it makes out the corresponding data fusion points based on energy levels as well as sensor affinity. The three metrics

information discovery cost, query resolution delay and query discovery ratio are considered for providing the information on effectiveness and completeness of the presented approach in obtaining energy efficiency and in enhancing QoS parameter i.e., latency.

C. Graph Clustering in Biological and Social Networks

Complex frameworks have been generally examined to portray their basic structural behaviors from a topological point of view. High modularity is one of the intermittent elements of a real-world complex frameworks. Different graph clustering algorithms have been applied to distinguishing groups or communities in social networks or modules in biological networks. In any case, their pertinence to real-world frameworks has been restricted as a result of the gigantic scale and complex availability of the systems. In the observed work, the authors exploit a novel informational-theoretic model for graph clustering [7]. The entropy-based clustering technique determines locally optimal groups by growing an arbitrary seed in a way that limits graph entropy. Relegating need in seed-choice and seed-development is well pertinent to the sans scale systems portrayed by the center arranged structure. Registering seed-development in parallel streams likewise breaks down a to a great degree extensive system productively. The experimental outcomes with genuine social and biological networks demonstrate that the entropy-based approach has preferable performance over contending techniques as far as exactness and effectiveness is considered.

D. Differential Evolution for Evolutionary Clustering Problem

The traditional clustering techniques assume that all the samples are collected from the same probability distribution that are static in nature as time pass by. There exists time-evolving data in areas like social web analysis which cannot be dealt with by traditional techniques. Such data has a tendency of changing its structure. Hence this kind of issue is to be handled by a feasible clustering technique i.e., evolutionary clustering technique, which able to deal with processing of time-evolving data so as to produce sequences of clusterings. Time-evolving data is related with the change in environment where stale data gets disappeared and replaced by newly arriving data and evolutionary algorithm is capable of responding to such change and can compute the global optima at any given time. Comparing with traditional techniques, evolutionary clustering based on DE (deEC) [8] performs a robust as well as adaptive global searching operation in the virtual solution space. The proposed deEC provides a solution by producing a sequence of clustering.

The two milestone criteria are to be achieved by the clustering result: one is to accurately reflect the present data achieved by slowly adjusting the movements of solutions during further stages of searching operation; and the other is not to deviate rapidly from one time stamp to the other. These criteria needs two objective functions or fitness functions: snapshot cost for measuring the cluster quality at a given time stamp that has inverse relationship, and history cost for measuring the temporal smoothness which can be expressed as adaptability of every single genotype in the stale environment that also maintains inverse relationship too.

In deEC, a genotype or a chromosome is denoted as a vector of real numbers. Given any probable chromosome in two dimension space, nearest prototype rule can be used to restore the cluster solution. For distance calculation, Euclidean distance was used to find the distance between the data vectors from the cluster centers. Local search operation was carried out using one step k-means algorithm so that fine-tuned partitions could be found by deEC otherwise global searching takes more time in iterating the DE operators. This type of local exploitation of searching the solution space sometime prevents from complex computation thus global optima can be achieved quickly.

E. Incremental Clustering

Due to web based learning nature, incremental clustering algorithms can deal with a continuous data stream. Specifically, different incremental clustering techniques in view of Adaptive Resonance Theory (ART) have been appeared to have low computational unpredictability in versatile learning and are less delicate to boisterous data. Nonetheless, parameter regularization in existing ART techniques is applied either on various features or on various clusters solely. In this paper, the authors presented a Interest-Focused Clustering based upon Adaptive Resonance Theory (IFC-ART) [4], which self-controls the vigilance parameter related with every feature and every cluster. The procedure uses domain knowledge into IFC-ART to concentrate on specific inclinations amid the self-managed clustering process. For performance assessment, we utilize a genuine informational collection, named American Time Use Survey (ATUS), which records almost 160,000 phone interviews directed with U.S. occupants from 2003 to 2014. In particular, contextual analysis is directed to investigate three sorts of fascinating relationships, concentrating on the age, wage, and arrangement of elderly care. Exploratory outcomes demonstrate that the execution of IFC-ART is profoundly aggressive and stable when contrasted and two settled clustering techniques and three ART models.

F. Clustering based on Flower Pollination Algorithm (FPA)

As clustering is for grouping of homogeneous data denotes the issue of modern data mining. It can be drawn closer through assortment of strategies based on statistical inference or heuristic methods. As of recent techniques utilizing novel meta-heuristics are of significant interest – as they can adequately handle the issue NP-hard. The paper concentrates on the utilization of bio-inspired Flower

Pollination Algorithm (FPA) [9] for grouping with internal measure of Calinski-Harabasz index being utilized as advancement standard. Alongside calculation's portrayal its execution is being assessed over an arrangement of benchmark occurrences and contrasted and the one of understood K-implies methodology. It is concluded that the use of presented procedure brings extremely encouraging results.

FPA tries to imitate an arrangement of complex systems pivotal to the accomplishment of plants reproductive procedures in the optimization domain. An individual flower or pollen gamete constitutes an solution to the optimization problem, with the entire flower populace being utilized. Their steadiness will be comprehended as solution fitness. Pollen will be moved over the span of two operations utilized interchangeably i.e., global and local pollination. The first utilizes pollinators to convey pollen to long distances towards individual described by higher fitness. Local pollination happens inside constrained scope of individual flower because of pollination mediators like wind or water.

G. Clustering using Firefly Algorithm

Sensor nodes have the consistent radio pattern (i.e., a similar detection and communication ranges). The network is at first connected (i.e., from any node x we can be reaching some other node y). They have heterogeneous mobility pattern. Every node knows their positional data which makes the tracking and repairing connection operation conceivable continuously.

Firefly algorithm is the replacement of the existing DBSCAN algorithm due to the higher reliability provided by using the fitness function F.

Light intensity is inversely proportional with the distance from the source of light. This relationship concludes with a proof that light intensity will be increased if the distance from the source of light will be decreases and vice versa.

Clustering of nodes in Wireless Sensor Networks is an issue of concern towards numerous researches. The real challenge is to implement an algorithm which can optimizes the estimation of different performance parameters like PDR and network lifetime of a node in the network. This paper demonstrates the importance of firefly algorithm [10] to perform clustering in remote sensor networks. The system lifetime of a node is enhanced as energy consumption is decreased while transmitting data in a network.

Phase 1: Calculating the Strength of Device Synergy (SDS) metric for distance calculation between two nodes. Firefly algorithm constructs clusters depending up on the distance between two nodes. in this scenario, the distance is determined by using the strength of each node which is in association with the nearest neighbor node. Strength of each node is based on the random movements and directions of nodes.

If a node moves slowly from within the communication range to its out of range, the SDS value of a node will be gradually decreased and that node is designated as ‘non-resident node (NR)’.

If a node bypasses another node in the communication range, then the strength of that by-passing node will be set to negative value (typically ‘-1’) and that node is designated as ‘foreign node (FN)’.

If a node moves slowly from out of communication range to within the communication range, then the strength of that node is set to a positive value (typically ‘+1’) and that node changes its designation from ‘non-resident’ to ‘resident node (RN)’.

Edge nodes (ENs) are the nodes that act as bridge nodes between the disjoint clusters.

Phase 2: Firefly algorithm performs clustering and a fitness function will be derived that is based up on value of residual energy as well as distance between the nodes. Finally clustering will be done using the optimum value of the following fitness function F.

Phase 3: This phase is meant for connectivity reestablishment among the nodes of clusters using ENs for the reestablishment of node connectivity. The SDS value for EN will be a lowest value. Hence a reliable connectivity will be guaranteed among the network.

III. CONCLUSION

Clustering is a heart task in data analysis. This paper reviews the various major enhancements and extensions of variety of algorithms. Every algorithm has its own importance in corresponding field. Efficiency and accuracy are the two important goals of any clustering algorithm wherever it is applied. When to apply a specified algorithm with a consideration of restricted parameters is to be taken care of. By introducing different types of metrics for CH selection so as to have a strong binding with the cluster members, the clustering technique as a whole contributes itself to the problem resolution.

REFERENCES

[1]. J. Zhang, H. Zhao, L. Cao, and Y. Chen, “Robust Clustering for Cognitive Radio Ad Hoc Networks with Group Mobility,” pp. 6–11, 2015.
[2]. C. Bauckhage, A. Drachen, and R. Sifa, “Clustering Game Behavior Data,” vol. 7, no. 3, pp. 266–278, 2015.
[3]. A. Maurya, “Time-Series Clustering for Data Analysis in Smart Grid,” 2016.
[4]. D. Wang and A. Tan, “Self-Regulated Incremental Clustering with Focused Preferences,” pp. 1297– 1304,

2016.
[5]. X. Cui and T. E. Potok, "A Distributed Flocking Approach for Information Stream Clustering Analysis," 2006.
[6]. R. Doss and G. Li, "Exploiting Affinity Propagation for Energy-Efficient Information Discovery in Sensor Networks," pp. 1–6, 2008.
[7]. E. C. Kenley and A. G. Entropy, "Entropy-Based Graph Clustering : Application to Biological and Social Networks," 2011.
[8]. G. Chen, "Evolutionary Clustering with Differential Evolution," 2014.
[9]. S. Łukasik, P. A. Kowalski, M. Charytanowicz, and P. Kulczycki, "Clustering using Flower Pollination Algorithm and Calinski-Harabasz Index," no. 1, pp. 2724–2728, 2016.
[10]. M. S. Manshahia and M. Dave, "Firefly Algorithm Based Clustering Technique for Wireless Sensor Networks," pp. 1273–1276, 2016.