International Journal of Enterprise Computing and Business Systems
ISSN (Online) : 2230-8849
Volume 6 Issue 1 January - June 2016
International Manuscript ID : ISSN22308849-V6I1M1-1112016

IMPROVED ALGORITHM ON DYNAMIC CLUSTERING USING METAHEURISTICS IN ADVANCE DATA MINING

Nanasaheb Mahadev Halgare
Asstt. Professor
M.S. Bidve Engineering College
WasWadi Latur, Maharashtra, India

Dr.Ali Akbar Bagwan
Research Supervisor
Kalinga University Chhattisgarh, India

ABSTRACT

Clustering or Aggregation of Items refers to the grouping of objects that belongs to the same class using some particular methodology. In other words, similar objects are grouped in one cluster and dissimilar objects are grouped in another cluster. It is one of the major tasks that is performed in Data Mining. Data mining (the analysis step of the "Knowledge Discovery in Databases" process, or KDD), an interdisciplinary subfield of computer science, is the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. Aside from the raw analysis step, it involves database and data management aspects, data pre-processing, model and inference considerations, interestingness metrics, complexity considerations, post-processing of discovered structures, visualization, and online updating. In this work, a unique metaheuristic approach Simulated Annealing is proposed to be used for the dynamic unsupervised clustering and it can produce the effective results

Keywords – Data Mining, Clustering, Simulated Annealing, Metaheuristic

INTRODUCTION

Clustering is the process of making a group of abstract objects into classes of similar objects.

- A cluster of data objects can be treated as one group.

- While doing cluster analysis, we first partition the set of data into groups based on data similarity and then assign the labels to the groups.
- The main advantage of clustering over classification is that, it is adaptable to changes and helps single out useful features that distinguish different groups.

APPLICATIONS

- Clustering analysis is broadly used in many applications such as market research, pattern recognition, data analysis, and image processing.
- Clustering can also help marketers discover distinct groups in their customer base. And they can characterize their customer groups based on the purchasing patterns.
- In the field of biology, it can be used to derive plant and animal taxonomies, categorize genes with similar functionalities and gain insight into structures inherent to populations.
- Clustering also helps in identification of areas of similar land use in an earth observation database. It also helps in the identification of groups of houses in a city according to house type, value, and geographic location.
- Clustering also helps in classifying documents on the web for information discovery.
- Clustering is also used in outlier detection applications such as detection of credit card fraud.
- As a data mining function, cluster analysis serves as a tool to gain insight into the distribution of data to observe characteristics of each cluster.

The following points throw light on why clustering is required in data mining −

- Scalability − We need highly scalable clustering algorithms to deal with large databases.
- Ability to deal with different kinds of attributes − Algorithms should be capable to be applied on any kind of data such as interval-based (numerical) data, categorical, and binary data.
- Discovery of clusters with attribute shape − The clustering algorithm should be capable of detecting clusters of arbitrary shape. They should not be bounded to only distance measures that tend to find spherical cluster of small sizes.
- High dimensionality − The clustering algorithm should not only be able to handle low-dimensional data but also the high dimensional space.
- Ability to deal with noisy data − Databases contain noisy, missing or erroneous data. Some algorithms are sensitive to such data and may lead to poor quality clusters.

- Interpretability − The clustering results should be interpretable, comprehensible, and usable.

CLUSTERING METHODS

Clustering methods can be classified into the following categories −

- Partitioning Method
- Hierarchical Method
- Density-based Method
- Grid-Based Method
- Model-Based Method
- Constraint-based Method

PARTITIONING METHOD

Suppose we are given a database of 'n' objects and the partitioning method constructs 'k' partition of data. Each partition will represent a cluster and $k \leq n$. It means that it will classify the data into k groups, which satisfy the following requirements −

- Each group contains at least one object.
- Each object must belong to exactly one group.
- For a given number of partitions (say k), the partitioning method will create an initial partitioning.
- Then it uses the iterative relocation technique to improve the partitioning by moving objects from one group to other.

HIERARCHICAL METHODS

This method creates a hierarchical decomposition of the given set of data objects. We can classify hierarchical methods on the basis of how the hierarchical decomposition is formed.

There are two approaches here −

- Agglomerative Approach
- Divisive Approach

AGGLOMERATIVE APPROACH

This approach is also known as the bottom-up approach. In this, we start with each object forming a separate group. It keeps on merging the objects or groups that are close to one another. It keep on doing so until all of the groups are merged into one or until the termination condition holds.

DIVISIVE APPROACH

This approach is also known as the top-down approach. In this, we start with all of the objects in the same cluster. In the continuous iteration, a cluster is split up into smaller clusters. It is down until each object in one cluster or the termination condition holds. This method is rigid, i.e., once a merging or splitting is done, it can never be undone.

APPROACHES TO IMPROVE QUALITY OF HIERARCHICAL CLUSTERING

Here are the two approaches that are used to improve the quality of hierarchical clustering −

- Perform careful analysis of object linkages at each hierarchical partitioning.

- Integrate hierarchical agglomeration by first using a hierarchical agglomerative algorithm to group objects into micro-clusters, and then performing macro-clustering on the micro-clusters.

DENSITY-BASED METHOD

This method is based on the notion of density. The basic idea is to continue growing the given cluster as long as the density in the neighborhood exceeds some threshold, i.e., for each data point within a given cluster, the radius of a given cluster has to contain at least a minimum number of points.

GRID-BASED METHOD

In this, the objects together form a grid. The object space is quantized into finite number of cells that form a grid structure.

ADVANTAGE

- The major advantage of this method is fast processing time.
- It is dependent only on the number of cells in each dimension in the quantized space.

MODEL-BASED METHODS

In this method, a model is hypothesized for each cluster to find the best fit of data for a given model. This method locates the clusters by clustering the density function. It reflects spatial distribution of the data points.

This method also provides a way to automatically determine the number of clusters based on standard statistics, taking outlier or noise into account. It therefore yields robust clustering methods.

CONSTRAINT-BASED METHOD

In this method, the clustering is performed by the incorporation of user or application-oriented constraints. A constraint refers to the user expectation or the properties of desired clustering results. Constraints provide us with an interactive way of communication with the clustering process. Constraints can be specified by the user or the application requirement.

DYNAMIC CLUSTERING

Dynamic clustering is a technique to find entries in your log similar to the current situation. Essentially, it is a K-nearest neighbor algorithm, and not actually clustering at all. Despite this misnomer, the term "Dynamic Clustering" has stuck with the Robocode community.

The idea is to record a "state" (or termed "situation") for each entry in your log. The state can contain any data that you deem valuable, such as lateral velocity, advancing velocity, or enemy distance. Save this along with your data. Then to use the data, you find a "distance" between current state and past states. Find some number of entries with the lowest distance, and use them for targeting, movement, or whatever you like.

The earliest method doing this was by iterating through the log and calculating the distance for each log entry. If you have a large log this is very slow.

International Journal of Enterprise Computing and Business Systems
ISSN (Online) : 2230-8849
Volume 6 Issue 1 January - June 2016
International Manuscript ID : ISSN22308849-V6I1M1-1112016

More recently kd-trees have been used. Corbos was the first one to mention them on the RoboWiki, which caught the interest of Chase-san and Simonton.

FUZZY CLUSTERING

In fuzzy clustering, every point has a degree of belonging to clusters, as in fuzzy logic, rather than belonging completely to just one cluster. Thus, points on the edge of a cluster, may be *in the cluster* to a lesser degree than points in the center of cluster. An overview and comparison of different fuzzy clustering algorithms is available.

Any point x has a set of coefficients giving the degree of being in the kth cluster $w_k(x)$. With fuzzy c-means, the centroid of a cluster is the mean of all points, weighted by their degree of belonging to the cluster:

$$c_k = \frac{\sum_x w_k(x)^m x}{\sum_x w_k(x)^m}.$$

The degree of belonging, $w_k(x)$, is related inversely to the distance from x to the cluster center as calculated on the previous pass. It also depends on a parameter m that controls how much weight is given to the closest center. The fuzzy c-means algorithm is very similar to the k-means algorithm:

- Choose a number of clusters.
- Assign randomly to each point coefficients for being in the clusters.
- Repeat until the algorithm has converged (that is, the coefficients' change between two iterations is no more than ε, the given sensitivity threshold) :
  - Compute the centroid for each cluster, using the formula above.
  - For each point, compute its coefficients of being in the clusters, using the formula above.

The algorithm minimizes intra-cluster variance as well, but has the same problems as k-means; the minimum is a local minimum, and the results depend on the initial choice of weights.

Using a mixture of Gaussians along with the expectation-maximization algorithm is a more statistically formalized method which includes some of these ideas: partial membership in classes. Another algorithm closely related to Fuzzy C-Means is Soft K-means.

Fuzzy c-means has been a very important tool for image processing in clustering objects in an image. In the 70's, mathematicians introduced the spatial term into the FCM algorithm to improve the accuracy of clustering under noise.

Metaheuristics are used to solve Combinatorial Optimization Problems, like Bin Packing, Network Routing, Network Design, Assignment Problem, Scheduling, or Time-Tabling Problems, Continuous Parameter Optimization Problems, or Optimization of Non-Linear Structures like Neural Networks or Tree Structures as they often appear in Computational Intelligence.

Metaheuristics are generally applied to problems for which there is no satisfactory problem-specific algorithm or heuristic; or when it is not practical to

implement such a method. Most commonly used Metaheuristics are focused to combinatorial optimization problems, but obviously can handle any problem that can be recast in that form, such as solving Boolean equations.

HEURISTICS AND METAHEURISTICS

Heuristic refers to "discover". A Heuristic is used when

1. Exact method are not on any help, due to execution time
2. There are errors in input data or is unreliable
3. Improvement in the performance of exact methods is required
4. There is need of a solution after a limited period of time.
5. We have to choose between addressing a more realistic model and provide an approximate solution instead of a simpler, unrealistic model that we can prove that can solve to optimality.
6. There is need of good starting points for an exact method.

DISADVANTAGES OF USING HEURISTICS

1. In many cases, convergence is generally guaranteed
2. Optimality may be achieved but it is not proved
3. In many cases, they may not be able to generate a feasible solution.

Metaheuristics are said to be high level procedures which coordinate simple heuristics such as local search, to find solutions that are of better quality than those found by simple heuristics done.

COMMONLY USED METAHEURISTICS

- Tabu search [Glover, 89 et 90]
- Simulated Annealing [Kirckpatrick, 83]
- Threshold accepting [Deuck, Scheuer, 90]
- Variable neighborhood [Hansen, Mladenovi´c, 98]
- Iterated local search [Loren¸co et al, 2000]
- Genetic Algorithm, Holland 1975 – Goldberg 1989
- Memetic Algorithm, Moscatto 1989
- Ant Colony Optimization, Dorigo 1991
- Scatter search, Laguna, Glover, Marty 2000

Countless variants and hybrids of these techniques have been proposed, and many more applications of Metaheuristics to specific problems have been reported. This is one of the active fields of research, with a considerable literature, a large community of researchers and users, and a wide range of applications.

Traditional methods of search and optimization are too slow in finding a solution in a very complex search space, even implemented in supercomputers. Metaheuristics consist of number of methods and theories having robust search method requiring little information to search effectively in a large or poorly-understood search space. There exists an extensive range of problems which can be formulated as obtaining the values for a vector of variables subject

to some restrictions. The elements of this vector are denominated decision-variables, and their nature determines a classification of this kind of problems. Specifically, if decision-variables are required to be discrete, the problem is said to be combinatorial. The process of finding optimal solutions (maximizing or minimizing an objective function) for such a problem is called combinatorial optimization.

Combinatorial optimization problems have been traditionally approached using exact techniques such as Branch and Bound (Lawler and Wood, 1966). Finding the optimal solution is ensured with these techniques but, unfortunately, they are seriously limited in their application due to the so-called combinatorial explosion. As an example, consider the Traveling Salesman Problem (TSP). This problem (obtaining a minimal Hamiltonian tour through a complete graph of n nodes) is a classical example of NP-complexity: the work-area to be explored grows exponentially according with the number of nodes in the graph, and so does the complexity of every know algorithm to solve this problem.

It is not only a good example of a combinatorial optimization problem, but also an approach to real problems like VLSI-design or X-ray Crystallography.

SIMULATED ANNEALING – A PROMINENT METAHEURISTIC APPROACH

The name Simulated Annealing (SA) is taken from annealing in metallurgy, a well known technique involving heating and controlled cooling of a material to increase the size of its crystals and reduce their defects. The heat makes the atoms become unstuck from their initial positions (a local minimum of the internal energy) and stroll randomly through states of elevated energy; the slow cooling gives more chances of finding configurations with lower internal energy than the initial one.

Each step in the SA algorithm replaces the current solution by an arbitrary "nearby" solution, chosen with a probability which depends on the difference between the corresponding function values and on a global parameter T (called the temperature), that is gradually decreased during the process. The dependency is such that the current solution changes almost randomly when T is large, but increasingly "downhill" as T goes to zero.

The method was independently described by Scott Kirkpatrick, C. Daniel Gelatt and Mario P. Vecchi in 1983, and by Vlado Černý in 1985. The method is an adaptation of the Metropolis-Hastings algorithm, a Monte Carlo method to generate sample states of a thermodynamic system, invented by N. Metropolis et al. in 1953.

The table below shows the mapping of physical annealing to Simulated Annealing.

Thermodynamic Simulation	**Combinatorial Optimization**
System States	Feasible Solutions
Energy	Cost
Change of State	Neighboring Solutions
Temperature	Control Parameter

International Journal of Enterprise Computing and Business Systems
ISSN (Online) : 2230-8849
Volume 6 Issue 1 January - June 2016
International Manuscript ID : ISSN22308849-V6I1M1-1112016

Frozen State	Heuristic Solution

Table 1: Relationship between physical annealing and Simulated Annealing

Using these mappings, any combinatorial optimization problem can be converted into an annealing algorithm.

The major advantage of SA over other methods is an ability to evade becoming trapped at local minima. This algorithm employs a random search, which not only accepts changes that decrease objective function, f, but also some changes that increase it. The latter are accepted with a probability

$$p = \exp (- \delta f/T)$$

where δf is the increase in f and T is a control parameter.

The algorithm starts by generating an initial solution and by initializing the temperature parameter T. Then the following is repeated until the termination condition is satisfied: A solution s' from the neighborhood N(s) of the solution s is randomly sampled and it is accepted as new current solution depending on f(s), f(s') and T. s' replaces s if f(s') < f(s) or, in case f(s') >= f(s), with a probability which is a function of T and f(s') - f(s). The probability is generally computed following the Boltzmann distribution exp(-(f(s') - f(s))/T).

The temperature T is decreased during the search process, thus at the beginning of the search the probability of accepting uphill moves is high and it gradually decreases, converging to a simple iterative improvement algorithm. This process is analogous to the annealing process of metals and glass, which assume a low energy configuration when cooled with an appropriate cooling schedule. Regarding the search process, this means that the algorithm is the result of two combined strategies: random walk and iterative improvement. In the first phase of the search, the bias toward improvements is low and it permits the exploration of the search space; this erratic component is slowly decreased thus leading the search to converge to a (local) minimum. The probability of accepting uphill moves is controlled by two factors: the difference of the objective functions and the temperature.

On the one hand, at fixed temperature, the higher the difference f(s')- f(s), the lower the probability to accept a move from s to s'. On the other hand, the higher T, the higher the probability of uphill moves.

SIMULATED ANNEALING BASED CLUSTERING

- Solution space of initial sample data items
- Cost function for cluster formation and outlier analysis
  - Determines how "good" a particular solution is
- Perturbation rules (Acceptance or Rejection of the Solution)
  (Transforming a solution to a new one)
- Simulated Annealing engine
  - A variable T, analogous to temperature
  - An initial temperature T_0 (T_0 =

40,000)

- A freezing temperature Tempfreezing (Tempfreezing = 0.1)
- A cooling schedule (T = 0.95 * T)

- Analysis of the Cluster and its acceptance value
- Generation of Final Outliers and Clusters

Another variant of Simulated Annealing also exists with the name Adaptive simulated annealing (ASA), in which the algorithm parameters that control temperature schedule and random step selection are automatically adjusted with the advancement of algorithm.

It makes the algorithm more efficient and less sensitive to user defined parameters than canonical Simulated Annealing.

CONCLUSION

Simulated Annealing is commonly said to be the oldest among the metaheuristics and surely one of the first algorithms that had an explicit strategy to avoid local minima. The fundamental idea is to allow moves resulting in solutions of worse quality than the current solution (uphill moves) in order to escape from local minima. The probability of doing such a move is decreased during the search. Using SA, the effective clustering based on unsupervised approach can be implemented and results can be evaluated good as compared to the classical approach.

REFERENCES

[1] Nock, R. and Nielsen, F. (2006) "On Weighting Clustering", IEEE Trans. on Pattern Analysis and Machine Intelligence, 28 (8), 1–13

[2] Bezdek, James C. (1981). Pattern Recognition with Fuzzy Objective Function Algorithms. ISBN 0-306-40671-3.

[3] Ahmed, Mohamed N.; Yamany, Sameh M.; Mohamed, Nevin; Farag, Aly A.; Moriarty, Thomas (2002). "A Modified Fuzzy C-Means Algorithm for Bias Field Estimation and Segmentation of MRI Data" (PDF). IEEE Transactions on Medical Imaging 21 (3): 193–199. doi:10.1109/42.996338. PMID 11989844.

[4] Improving Simulated Annealing-based FPGA placement with directed moves, K. Vorwerk, A. Kennings, J. W. Greene IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, Volume 28, Issue 2, pp. 179-192, February 2009.

[5] A technique for minimizing power during FPGA placement, K. Vorwerk, M. Raman, J. Dunoyer, Y.-C. Hsu, A. Kundu, A. Kennings, International Conference on Field Programmable Logic and Applications, Heidelberg, Germany, September 8-10, 2008.

[6] Osman, I.H. (1993), “Metastrategy Simulated Annealing and Tabu Search Algorithms for the Vehicle Routing Problem”, Annals of Operations Research 41, 421-451.

[7] Osman, I.H. and J.P. Kelly (eds.) (1996), Meta-Heuristics: Theory and Applications, Kluwer Academic Publishers, Norwell, MA.

[8] Metaheuristics: Progress as Real Problem

Solvers
Series: Operations Research/Computer Science Interfaces Series, Vol. 32 Ibaraki, Toshihide; Nonobe, Koji; Yagiura, Mutsunori (Eds.) 2005, XII, 414 p. 106 illus., ISBN: 978-0-387-25382-4

[9] On the Relations Between Search and Evolutionary Algorithms (1996), D. Ernst, Xiaowei Shen

[10] Performance evaluation of metaheuristic search techniques in the optimum design of real size pin jointed structures, O. Hasançebi, S. Çarbaş, E. Doğan, F. Erdal and M.P. Saka 2009

[11] http://cisnet.mit.edu/pageview/4i7o9/6ab9o/9as3u/168 Last Accessed : 15 May 2009

[12] http://www.doc.ic.ac.uk/~nd/surprise_96/journal/vol4/tcw2/report.html Last Accessed : 09 January 2009.

[13] Improving simulated annealing-based FPGA placement with directed moves, K. Vorwerk, A. Kennings, J. W. Greene IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, Volume 28, Issue 2, pp. 179-192, February 2009.

[14] A technique for minimizing power during FPGA placement, K. Vorwerk, M. Raman, J. Dunoyer, Y.-C. Hsu, A. Kundu, A. Kennings, International Conference on Field Programmable Logic and Applications, Heidelberg, Germany, September 8-10, 2008.

[15] VLSI floorplan repair using dynamic whitespace management,constraint graphs and linear programming, K. Vorwerk, A. Kennings, M. Anjos, Engineering Optimization, Volume 40, Number 6, pp. 559-577, 2008.

[16] Osman, I.H. (1993), "Metastrategy Simulated Annealing and Tabu Search Algorithms for the Vehicle Routing Problem", Annals of Operations Research 41, 421-451.

[17] Osman, I.H. and J.P. Kelly (eds.) (1996), Meta-Heuristics: Theory and Applications, Kluwer Academic Publishers, Norwell, MA.