International Journal of Scientific Engineering and Technology (ISSN : 2277-1581)
Volume No.3 Issue No.7, pp : 845-847 1 July 2014

A Survey on Optimization Approaches to K-Means Clustering using Simulated Annealing

Abha Kaushik[1], Subhajit Ghosh[2] & Sunita kumari
School of Computing Sc, Galgotias University.
abhakaushikmtech@gmail.com[1], subhajit.ghosh@galgotiasuniversity.edu.in[2],
sunita.rao86@gmail.com[3]

Abstract : ***Clustering is one of the fastest growing research areas because of availability of huge amount of data. It models data into the clusters. Data modelling puts clustering in a historical perspective rooted in statistics, mathematics, and numerical analysis. From a machine learning perception clusters correspond to hidden patterns, the exploration for clusters is unsupervised learning, the resultant system represents a data model. There are many techniques for clustering of data based on similarity. K-Means is one of the simplest unsupervised learning methods among all partitioning based clustering methods. It classifies a set of data objects in clusters. All the data objects are placed in a cluster having centroid nearest to that data object. After processing the data objects centroids are recalculated, and the whole process is repeated. This paper presents a brief estimation of the existing body of work that employs simulated annealing approach to improve upon the k-means clustering process.***

Keywords: **optimization, k-means, clustering, simulated annealing**

Introduction

Data clustering is used regularly in many applications such as data mining, vector quantization, pattern recognition, and fault detection & speaker recognition. The most well-known, widely used and fast methods for clustering is K-means clustering developed by Mac Queen in 1967. The simplicity of K-means clustering made this algorithm used in various fields. K-means clustering is a partitioning clustering method that separates data into k mutually groups. Through such the iterative partitioning, K-means clustering minimizes the sum of distance from each data to its clusters. (5)

K-means is simple and can be easily used for clustering of data practice and the time complexity is O(nkt), n is the number of objects, k is the number of clusters, tis the number of iterations, so it is generally regarded as very fast. But there are some drawbacks in k-means algorithm.

• The result of K-means algorithm depends on initial clustering centres; different seed-points can cause different clusters, and can even lead to no resolution.

• The algorithm ends in local minima value. [1]

A new algorithm for optimization of K-means clustering is proposed in this paper. The new approach proposed based on the simulating annealing. Simulated annealing was proposed by Kirk- patrick et al. as a method for solving combinatorial optimization problems where a function of many variables is minimized or maximized. The idea was

derived from the algorithm pro- Simulating annealing show the effectiveness to improve the clustering results of K-means clustering. It can be appreciated that as the temperature of the system decreases the probability of accepting a worse change is reduced. This is the similar as slowly moving to a frozen state in physical annealing. Also note, that if the temperature is zero then only better moves will be accepted which effectively makes simulated annealing act like hill climbing.

1.K-Means

K-Means is one of the simplest unsupervised learning methods among all partitioning based clustering methods. [2] It categorises a given set of *n* data objects in *k* clusters, where *k* represents the number of desired clusters and it is required in advance. A centroid is defined for each cluster. All the data sets are placed in a cluster having centroid nearest (or most similar) to that data object. After processing entire data objects, k-means (centroids) are recalculated, and the whole process is repeated. The whole data objects are bound to the clusters depend on the new centroids. In each repetition centroids change their location in a predefined way. In other words, Centroids move in each iteration. This process is continued until no centroid is left to move. As a result, *k* clusters are found representing a set of *n* data objects. The idea behind clustering is to split data into clusters based on cluster center and assign each point to nearest center.

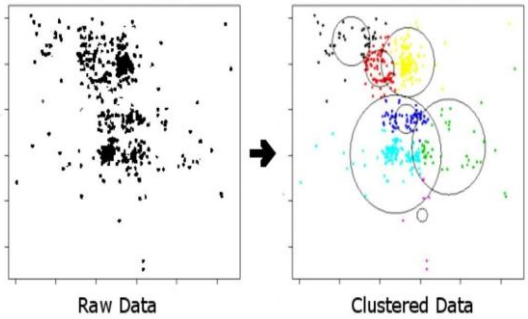

Figure: k-mean clustering

Raw data in the given figure has been grouped together to its neighboring points to form clusters. Let's see an example to understand how clustering is done using K-Means.

The objective function j

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2$$

International Journal of Scientific Engineering and Technology (ISSN : 2277-1581)
Volume No.3 Issue No.7, pp : 845-847 1 July 2014

Where $\left\| x_i^{(j)} - c_j \right\|^2$ is a chosen distance measure between the data point $x_i^{(j)}$ and the cluster centre c_j, is an indicator of the distance of the *n* data points from their centre of clusters.[11]

The algorithm for k-mean clustering is composed of the following steps:

Input: '*k*', is the number of clusters to be partitioned; '*n*', the number of objects.
Output: A set of '*k*' clusters based on given similarity function.
Steps: (i) Arbitrarily choose '*k*' objects as the initial cluster centers. No specific pattern is required while choose these.
ii) Repeat,
a. (Re) assign each object to the cluster to which the object is the most similar; based on the given similarity function;
b. Update the centroid, i.e., calculate the mean value of the objects for each Cluster. iii) These steps needs to be repeated until no change occurs in the mean value.

K-Means Algorithm Properties [8]:

- There are always K clusters.
- There is always at least one item in each cluster.
- Clusters do not intersect because they are in non-hierarchical order.
- Every member of a cluster is nearer to its cluster than any other cluster because closeness does not always involve the 'centre' of clusters

K-means error estimation:

- By comparing two dissimilar runs of the K-Means algorithm we have to be able to evaluate its quality.
- There is no ground truth (we don't know the clusters/labels beforehand instead of this sum of squared errors (SSE) is used: **SSE= K ∑ i=1 ∑ x∈ ci d(x, ci)2**

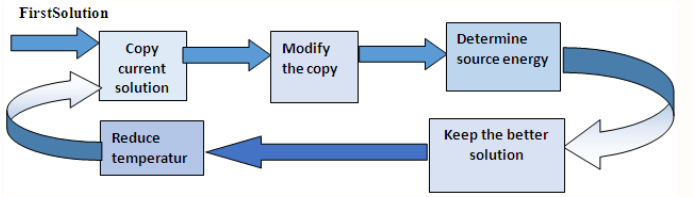

Fig: Process cycle of Simulated Annealing

3.Simulated annealing

Clustering Using Annealing Evolution:

Simulated annealing is a stochastic searching optimization algorithm based on Mente Carl iteration strategy. [1]The searching capabilities of simulated annealing and programming have been used for the purpose of appropriately determining a fixed number of cluster centers in thereby suitably clustering the set of unlabeled points. The clustering metric that has been adopted, is the sum of the Euclidian distances of the points from their respective cluster centers. [2]

The name and inspiration come from annealing in metallurgy, a technique including controlled cooling and heating of a material to increase the size of its crystals and reduce their defects. The heat effects the atoms to become unstuck from their initial positions, a local minimum of internal energy ,and wander unsystematically through states of higher energy; the slow cooling gives them more chances of finding configurations with lower internal energy than the initial one. By analogy through this physical process, each step of the simulating annealing algorithm replaces the current solution by a random "closed" solution, chosen with a probability that depends on the difference between the corresponding function values and on a global parameter T (represents the temperature), that is slowly decreased during the process. The dependency is such that the existing solution changes almost randomly when T is large, but increasingly "downhill" as T goes to zero. The allowance for "uphill" moves keeps the method from becoming stuck at local minima, which are the bane of greedier techniques. SA can be applied in a lot of applications and can be realized easily. But the compute time of SA is long, so the efficiency is low, and it is difficult to reach the strict convergence condition in the process of use.

The simulating annealing process has shown in the figure as follows:

The basic algorithm for simulating annealing based on the following figure is as follows:

1. Create the first solution, and get its energy value.
2. While temperature > minimum temperature...
   (i) Make a copy of the solution.
   (ii) Modify the copy.
   (iii) Get the copy's energy value.
   (iv) Keep the better of these two solutions.
   (v) Reduce the temperature.
3. Repeat

The algorithm ends when temperature touches a pre-set minimum value.

The better solution based on:

P=exp(-delta / Temp)

Where delta shows the difference between the copy's energy and the earlier solution's energy.

As the temperature slowly decreases, the difference in energy values has a deep effect on acceptance. At lower temperatures, when the energy difference is high, the algorithm is very discriminating. When the energy difference is low at higher temperatures, the algorithm will agree most anything.

Conclusion: K-mean algorithm has biggest advantage of clustering large data sets and its performance increases as number of clusters increases. This survey has presented the research work done on data clustering based on optimization techniques. The survey starts with a brief introduction about clustering in data mining, and explored various research papers (8, 2, 4, 5, 8, 18) related to techniques of clustering. More research works have to be carried out based on K-means clustering and simulating annealing.

References:

i. *Jinxin Dong &Minyong Qi,"K-means Optimization Algorithm for Solving Clustering Problem",IEEE,2009.*

ii. *L.Sun, F.Xu, Y.Liang, Y.Xie &R.Yu,"cluster analysis by the k-means algorithm and simulated annealing,Elsevier science B.V,1994.*

iii. *C. D. Gelatt, S. Kirkpatrick, M. P. Vecchi,"Optimization by Simulated Annealing",13 May, 1983, New Series, Vol. 220, No. 4598.*

iv. *Z.volkvovich and D. kitai,"Self-learning k-means clustering:- A global optimization approach," june, 2012.*

v. *Ali Ridho Barakbah," A New Algorithm for Optimization of K-Means Clustering with Determining Maximum Distance Between Centroids" IES 2006 – Politeknik Elektronika Negeri Surabaya - ITS*

vi. *D.Bertsimas and J.Tesitsiklis,"Simulated Annealing", 1993, vol.8, no.1, statistical science*

vii. *A.M.Bagirov,J.yearwood,"A new nonsmooth optimization algorithm for minimum sum-of-squares clustering problems, European journal of operational research,june 2004.*

viii. *A.Joshi, R.kaur, "A Review: Comparatives Study of Various Clustering Techniques in Data Mining",Volume3, International Journal of Advanced Research in Computer Science and Software Engineering, March 2013*

ix. *Ronald S. King, An Introduction to Cluster Analysis for Data Mining*

x. *http://minethedata.blogspot.in/2012/08/bisecting-k-means.html*

xi. *http://home.deib.polimi.it/matteucc/Clustering/tutorial_html/index.html*

xii. *AristidisLikasa,NikosVlassisb,JakobJ.Verbeekb,"The global k-means clustering algorithm," Elsevier Science Ltd,march, 2002.*

xiii. *https://www.google.co.in/#q=Machine+Learning+for+Data+Mining+Week+6%3A+Clustering*

xiv. *http://lion.disi.unitn.it/reactive-search/thebook/node17.html*

xv. *Jiawei Han and Micheline Kamber,"Data Mining Concepts and Techniques", University of Illinois at Urbana-Champaign, Second Edition,*

xvi. *http://mnemstudio.org/simulated-annealing-introduction.htm*

A. *Shafeeq B M, H. K S," Dynamic Clustering of Data with Modified K-Means Algorithm" 2012 International Conference on Information and Computer Networks (ICICN 2012*

xvii. *N. Zhong and y. Li and Sheng-Teng Wu,"Effective Pattern Discovery in Text Mining", Transactions on knowledge and data Engeneering, IEEE, January 2012.*