TRUSTWORTHY APPROACH TO FIND OPTIMAL CENTROID FOR CLUSTERING MULTI-VARIABLE DATA

Dr. K. Reddy Madhavi[1], Dr. J. Avanija[2], Dr. L. Venkateswara Reddy[3]
[1,2]Associate Professor, CSE, [3]Professor, IT
Sree Vidyanikethan Engineering, College, A. Rangampet, Tirupati

Abstract

Identification of useful clusters in large datasets has attracted considerable interest in clustering process. In the process of generation of clusters, selecting initial cluster center (centroid) is the key factor that has high impact on accuracy of clusters. In existing algorithms like K-Means and K-Modes, this centroid selection was performed randomly. Because different random initializations of cluster centroids leads to different clusters, current work concentrates on generation of initial cluster centers by analyzing properties of data and formulating into function. Our previous work was concentrated on selecting optimal cluster centroid for single variable and two variables. This paper proposes a trustworthy approach to select centroid for multi-variable data. This method is independent of clustering algorithms; it can be applied to any type of clustering where random selection is made. By finding optimal centroid, it minimizes the impact of random selection of initial centers in

Keywords: Cluster centroid; trustworthy, and multi-variables.

I. INTRODUCTION

Data mining is the process of extracting or mining knowledge from bulky volumes of data [9]. Among its number of functionalities, Clustering is the major one in which there is still a room for research. Clustering is an unsupervised learning and generates clusters satisfying the principle of less similarity between other clusters and more similarity within the cluster. Since data has been clustered around centroids, the proposed work has been concentrated on identifying optimal centroid based on number of variables.

Current observation shows that, existing clustering algorithms selects initial cluster center randomly. selects any data point as initial cluster center and performs the clustering process. But different random initializations of cluster centers can lead to different clusters. Therefore choosing the suitable initial cluster center is a problem worthy of studying.

The remainder of the paper is systematized as follows: partition 2 focus on relevant study, partition 3 provides trust worthy approach to find centroid for unconstrained multivariable functions, partition 4 presents illustration and discussions and partition 5 concludes the summary with its future work.

II. RELEVANT STUDY

Data may be numerical, categorical and mixed. Various clustering algorithms support clustering of flowing numerical data. Little number of clustering algorithms focus on flowing categorical data [3]. New data keep on enters into the database, the importance of which must be added to the existing data clusters. Most popular K-means and K-modes algorithms are applied to handle streaming data [15] and at the same time suffers from problems of random initializations. A validity function treated as the objective function [11] for evaluating the effectiveness of the clustering model where incoming data is flowing always. An iterative optimization algorithm is used for solving an objective function. Proposed method for selection of initial cluster center needs preprocessing for analyzing the properties of data. Depending on these properties the objective function is formulated [14]. Our previous work was for selecting initial cluster center for single variable function, two variable functions and unconstrained multi variable and so on [16,17,18,19,20,21]. Now it is concentrated on identifying optimal centroid for multi-variable data with gradient information. Survey was done for this purpose and observed the following concepts.

For all the methods that employ gradient information, $f(y), \nabla f(y)$ and $\nabla^2 f(y)$ exist and are continuous where f(y) is

value of function at the given point y, $\nabla f(y)$ is the first order partial derivative of function or gradient $\nabla^2 f(y)$ and second order partial derivative of function or Hessian matrix. Cauchy's method takes stationary point in the design space and will determine the most local descent direction by considering the Taylor expansion by ignoring terms higher than degree 2. The method of steepest descent is to solve linear equations by using eq.―(1)‖

$$y^{(i+1)} = y^{(i)} - \theta^{(i)} \nabla f(y^{(i)}) \qquad (1)$$

where $y^{(i)}$ is the current estimate value. θ is the fixed positive step length parameter. The suitable value of θ has to be made and its value keeps on decreases near the minimum. θ has to be adjusted at each iteration. $\theta^{(i)}$ was determined such that $f(y^{(i+1)})$ is minimum. Because of line search using eq.―(1)‖ this method is more reliable than simple gradient method, but still the convergence rate will be insufferably slow for most of practical problems. Though this method is not having immense practical value, it reveals the procedure of most of gradient-based methods. The method will come close to the minimum slowly. It is assured that $f(y^{(i+1)}) \leq f(y^{(i)})$. The method requires only objective and gradient values at each iteration. Newton's method extends this strategy by using higher order information by using eq.―(2)‖ namely second order derivatives, for constructing more global strategy [13] for which it obtain Taylor expansion by ignoring the terms higher than order 3 and forms quadratic approximation function.

$$y^{(i+1)}=y^{(i)} - \nabla^2 f(y^{(i)})^{-1} \nabla f(y^{(i)}) \qquad (2)$$

This method minimize a quadratic function in exactly one step. It is unreliable for nonquadratic functions. This method is modified to eq.―(3)‖ by adding a line search. Now it is reliable and efficient.

$$y^{(i+1)}=y^{(i)} - \theta^{(i)} \nabla^2 f(y^{(i)})^{-1} \nabla f(y^{(i)}) \qquad (3)$$

Hestenes and Stiefel [10] put out conjugate gradient method which is iterative for solving linear equations. Next, FR [6] proved quadratic convergence and extended to nonquadratic functions. FM [7] expressed the utility of the method to linear sets that result from finite element discretizations with sparse coefficient matrix. MC [12] gave an extension to FR method as eq.―(4)‖

$$y^{(i+1)}=y^{(i)}+\theta^{(i)}\{-\nabla f(y^{(i)})\} + \delta^{(i)}\, d(y^{(i-1)}) \qquad (4)$$

where $\theta^{(i)}$ and $\delta^{(i)}$ are required at each iteration and d is search direction. It is efficient w.r.to number of iterations but more function and gradient evaluations are needed. To locate the minimum for quadratic functions N line searches and N-1 conjugate search directions are needed in the absence of round-off error. Here N is problem dimension. For non quadratic functions, more directions and line searches are necessary. FR method assumes both a quadratic objective and exact line searches. Some computational experience by a number of investigators suggest it is prudent to restart i.e., set d(y) = $-\nabla f(y^{(i)})$ every N or N - 1 steps. Few works stipulates exact line searches by assuming more general objective model that produces $y^{(i+1)}$ by using eq.―(5)‖.

$$y^{(i+1)} = y^{(i)} + \theta^{(i)} d(y^{(i)}) \qquad (5)$$

$$\text{where } d(y^{(i)}) = -\nabla f(y^{(i)}) + \delta^{(i)}\, d(y^{(i-1)})$$

It was proved that this method is superior to the FR method for general functions. Furnishes linear rate of convergence without restart. Less sensitive to inexact line searches. Methods in [1,4] assume exact line searches. But employ more general objective function than quadratic function. The methods intended to imitate positive characteristics of Newton's method (Quasi-Newton Method) generate directions using eq.―(6)‖ with only first-order information

$$d(y^{(i)}) = - M^{(i)} \nabla f(y^{(i)}) \qquad (6)$$

where $M^{(i)}$ is N x N matrix. The Davidon–Fletcher–Powell (DFP) method is robust and does well for extensive practical problems. The major disadvantage is storing of $N \times N$ matrix M. Another method proposed almost simultaneously by BFGS [2,5,8,22] has received wide approval because of its advantages, such as, reduced need to restart, less dependent upon exact line searching using eq. ―(7)‖

$$y^{(i+1)} = [I - U]\, M^{(i)}\, [I - U]^{T} + \frac{\Delta y^{(i)} \Delta y^{(i)T}}{(\Delta y^{(i)})^{T} \Delta(\nabla f(y^{(i)}))}$$

Where $U = \frac{\Delta y^{(i)} \Delta(\nabla f(y^{(i)}))^{T}}{(\Delta y^{(i)})^{T} \Delta(\nabla f(y^{(i)}))}$ and I is identity matrix (7)

III. PROPOSED METHOD

All these methods discussed in relevant study differs in the appropriate search direction from current iteration to next iteration. But all these methods follow line searching strategy that consumes most of the time. Hence we propose the trustworthy gradient based method that does not use line search strategy. Functions of multi-variables are constrained or unconstrained. Proposed work concentrates on finding centroid (minimum point) for unconstrained objective function with multi-variables based on gradient information. Trust region methods use a slightly different strategy than all the methods focused on relevant study. This method forms a trustworthy approximation of $f(y)$ and computes minimum. The trust region is defined as ―a neighborhood around the current iterate $y^{(t)}$ in which the current approximation of $f(y)$ produces the descent‖. Now the importance shifts from producing descent directions to outlining safe approximations to $f(y)$. The method challenges to take advantage of ―both the steepest descent and Newton methods‖ by modifying the diagonal elements of the Hessian matrix H. The procedure includes following steps.

Step 1: Start by defining $y^{(0)}$ which is an initial estimate of y*, K is maximum number of iterations allowed, Ɛ is convergence criterion and set i=0, constants $\beta^{(0)} = 10^4$, $C_1(0<C_1<1)$ and $C_2(C_2>1)$ Step 2: Compute gradient of function$\nabla f(y^{(i)})$.

Step 3: Test for optimality. if $\|\nabla f(y^{(i)})\| <$ Ɛ then print the results and stop the process. else go to step 5.

Step 4: if i≥K, then print the results and stop the process. else go to step 5.

Step 5: Compute the new vector

$$y^{(i+1)} = y^{(i)} + d(y^{(i)})$$

Step 6: Search direction $d(y^{(i)}) = -[H^{(i)} + \beta^{(i)} I]^{-1} \nabla f(y^{(i)})$ Step 7: Compare the values of $f(y^{(i+1)})$ and $f(y^{(i)})$. if $f(y^{(i+1)}) < f(y^{(i)})$ then set $\beta^{(i+1)} = C_1\, \beta^{(i)}$ and i=i+1. Go to step 2. else set $\beta^{(i)} = C_2\, \beta^{(i)}$. Go to step 6.

This method is simple, the descent property, the outstanding convergence rate near y*and the absence of a line search. This is the motivation to introduce this method for multi-variables based on gradient information.

IV. ILLUSTRATION AND DISCUSSIONS

Among the methods discussed in our relevant study that deals with unconstrained gradient objective function with multi-variables, a study was performed by Sargent including the BFGS, DFP, and FR algorithms by testing the effect of line search termination criteria. Their results indicate the superiority of the BFGS method of quasi-Newton methods for general functions. The tests performed on the problems for relative computational efficiency using the methods such as Powell's direct-search method, DFP and Newton's method for structural applications. Shown that DFP and Newton's methods are superior to Powell's method. DFP superior to Newton's method for large problems. But it is fair to say that the tests suggest a superiority of the BFGS method over other available methods. All these gradient based methods uses common gradient based algorithm that computes $\alpha^{(i)}$ such that

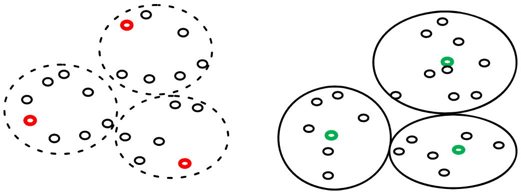
$$f(y^{(i)} + \alpha^{(i)} d(y^{(i)})) \text{ is minimum using } \theta \qquad (8)$$

where θ is the line search convergence criteria. Different methods were produced by defining appropriate search direction $d(y^{(i)})$ in each method that are discussed in relevant study of this paper. An efficient implementation would certainly contain additional tests for ―(8)‖ that are suggested by DFP. Certainly, exact line searches should be avoided whenever possible. Tests indicate that line searching is most time consuming. Hence proposed trustworthy method to find minimum point for unconstrained nonlinear objective function with multi-variables based on gradient information is better.

This minimum point acts as optimal centroid.

A. Comparison with k-means method

Process of traditional K-Means algorithm [9] is shown in ―Fig.1‖.Assume set of objects located in a particular space would like to be partitioned into three clusters. Then kmeans algorithm first randomly chooses three objects as the three initial cluster centers represented with red color. Each object is allocated to a cluster based on the nearest cluster center, encircled by dotted curves, as shown in ―1(a)‖. Next, the cluster centers are revised by recalculating mean of each cluster with current objects of the cluster. Then the objects are redistributed to the clusters forming new centers. This process iterates, leading to final clusters with their objects as shown in―1(b)‖. The new final cluster centers are represented with green color.



1(a) 1(b)

Figure 1. Clustering of objects using k-means algorithm In Fig 1(a) randomly selected any one of object as initial cluster center. In this randomization any object may be selected as initial cluster center. But different random initializations lead to different results which has a lot of affection on cluster accuracy. This paper focuses on identifying minimum that can act as initial cluster center for unconstrained nonlinear objective function with multivariables based on gradient information. It is illustrated below.

Minimize the function $f(y_1,y_2)= y_1-y_2+2y_1^2+2y_1y_2+ y_2^2$

From initial point $y^1=[y_1^1, y_2^1]=[0\ 0]$ and $\mathcal{E} = 10^{-2}$ let $C_1= 1/4$,

$C_2=2$

Iteration 1:

$f(y^1)=[0\ 0]$, $\frac{\partial f}{\partial y_1} = 1$, $\frac{\partial f}{\partial y_2} = -1$ and $\nabla f(y^{(1)}) = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$

$\|\nabla f(y^{(1)})\| = \sqrt{1^2 + -1^2} = 1.414 > \mathcal{E}$

$H^{(1)} = \nabla^2 f(y^{(1)}) = \begin{bmatrix} 4 & 2 \\ 2 & 2 \end{bmatrix}$

$y^{(2)} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} - \begin{bmatrix} 4+10^4 & 2 \\ 2 & 2+10^4 \end{bmatrix}^{-1} \begin{pmatrix} 1 \\ -1 \end{pmatrix} = \begin{pmatrix} -0.999 \\ 1.000 \end{pmatrix} 10^{-4}$

$f(y^2) = -1.99 \times 10^{-4} < y^1$. hence set $\beta^{(2)} = 10^4/4 = 2500$, i=2.

Iteration 2:

$\nabla f(y^{(2)}) = \begin{pmatrix} 0.999 \\ -1.000 \end{pmatrix}$, $\| \nabla f(y^{(2)}) \| = 1.41 > \mathcal{E}$

Hence compute $y^{(3)} = \begin{pmatrix} -4.99 \\ 5.00 \end{pmatrix} 10^{-4}$

$f(y^3) = -0.99 \times 10^{-3} < y^2$. hence set $\beta^{(3)} = 2500/4 = 625$, i=3.

Repeat this iterative process until the convergence criterion is satisfied. If at y^4 convergence is reached then $f(y^4)$ will become minimum point that is initial cluster center. When this approach is applies to the same objects as in Fig 1, the minimum point achieved with the above illustration act as initial cluster centers represented with green color thus by reducing the iteration process of Fig 1(a). The result of this approach is shown in Fig (2).

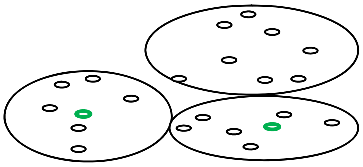


Figure 2: Identification of initial cluster centers

So using this approach the iterative process of Fig 1(a) can be avoided and also accuracy of the clusters can be improved because in the existing system clustering results are based on the object that is randomly selected may lead to different results. One object as initial cluster centroid may form different clusters with its objects and at the same time another object as initial cluster centroid may form different clusters with its different objects. The objects in the resulted cluster depend on initial cluster centroid selected.

V. CONCLUSION AND FUTURE WORK

Existing methods for the problems of this type of multivariable functions uses line search which is very time consuming process. This paper proposes the trustworthy approach with absence of line search to find optimal centroid of multivariable functions using gradient information. But it is necessary to compute the Hessian Matrix for every iteration. After finding minimum point, applied suitable distance measure to form the clusters. This method reduces time of line search compared to existing methods and minimizes the impact of randomness. This method is not limited to single type of clustering algorithm, but it can be applied to minimize the impact of randomness on clustering results.

REFERENCES

[1] Boland, W. R., E. R. Kamgnia, and J. S. Kowalik. 1979. A Conjugate Gradient Optimization Method Invariant to Nonlinear Scaling,'' *JOTA*, 27(2), 221–230.

[2] Broyden, G. G. 1970. The Convergence of a Class of Double-Rank Minimization Algorithms. *J. Inst. Math. Appl.*, 6, 76–90, 222–231.

[3] Chen H.L., Chen M.S., and Chen Lin SU. 2009. Frame work for clustering Concept–Drifting categorical data. *IEEE Transaction Knowledge and Data Engineering* Vol. 21, no.5.

[4] Davison, E. J., and P. Wong. 1975. A Robust Conjugate-Gradient Algorithm Which Minimizes L-Functions,'' *Automatica,* 11, 297–308.

[5] Fletcher, R. 1970. A New Approach to Variable Metric Algorithms. *Computer J.,* 13,317–322.

[6] Fletcher, R., and C. M. Reeves. 1964. Function Minimization by Conjugate Gradients. *Computer J.,* 7, 149–154.

[7] Fried, I., and J. Metzler. 1978. SOR vs Conjugate Gradients in a Finite Element Discretization. Int. J. Numer. Methods Eng., 12, 1329–1342.

[8] Goldfarb, D. 1977. Generating Conjugate Directions without Line Searches Using Factorized Variable Metric Updating Formulas. *Math Prog.,* 13, 94–110.

[9] Han. J. and Kamber. M. 2001. Data Mining: Concepts and Techniques, Morgan Kaufmann.

[10] Hestenes, M. R., and E. Stiefel. 1952. Methods of Conjugate Gradients for Solving Linear Systems,'' *NBS Res. J.,* 49, 409–436.

[11] Liang Bai, Xueqi Cheng, Jiye Liang, Huawei Shem, ―An Optimization
Model for Clustering Categorical Data Streams with Drifting Concepts‖, IEEE Transactions on Knowledge and Data Engineering, 2871 – 2883, Volume: 28, Issue: 11, Nov. 1 2016.

[12] Miele, A., and J. W. Cantrell. 1969. Study on a Memory Gradient Method for Minimization of Functions. *JOTA,* 3(6), 459.

[13] Ragsdell, K. M., R. R. Root, and G. A. Gabriele. 1974. Newton's Method: A Classical Technique of Contemporary Value. *ASME Design Technol. Transfer Conf. Proc.,* New York, p. 137.

[14] Ravindran.A. and Ragsdell. Engineering Optimization Methods and applications. 2nd edition, *wiley* edition.

[15] Ravi Sankar Sangam, Hari Om, ―K-mode stream algorithm for clustering categorical data streams‖, CSI Transactions on ICT, pp 1–9,2017

[16] K.Reddy Madhavi, ―Graphical Method to find Optimal Cluster Centroid for Two-variable Linear Functions of Concept drift Categorical Data‖ 2nd International Conference on Computational Science, Engineering and Information Technology (CCSEIT-2012), ACM ICPS,Coimbatore, ISBN: 978-4503-1310-0, India, Oct 26-28, 2012.

[17] K.Reddy Madhavi, ―Identification of Optimal Cluster Centroid for Multi-Variable Functions for Concept-drift Categorical Data,‖ International Conference on Advances in Computing, Communications and Informatics (ICACCI), ACM ICPS and DBLP indexed, Chennai, Aug 3-5, 2012.

[18] K.Reddy Madhavi, ―Identification of Optimal Cluster Centroid for Unconstrained Nonlinear Multivariable Functions,‖ ART Com- Fourth International Conference on Advances in Recent Technologies in Communication and Computing, by ACEEE (Association of Computer, Electronics and Electrical Engineers) IET digital library, Bangalore, Oct 19-20, 2012.

[19] K.Reddy Madhavi, ―Clustering of Concept-drift Categorical data implementation in Java‖, The Fourth International Conference on Recent trends in Computing, Communication and Information Technologies- ISSN:1865-0929, ISBN: 978-3-642-29215-6, Springer CCIS series, VIT University, Vellore, ObCom Dec 2011.

[20] K.Reddy Madhavi, ―Modified S^2 and Pattern Search Methods to Find Optimal Cluster Centroid for Multi-Variable Functions,‖ International journal of Computer Technology and Electronics Engine+ering(IJCTEE) , placed in National Science Library(NSL) ISSN: 2249 6343, Vol 2, Issue 2, pp: 173-178, Texas, USA, 2012.
[21] K.Reddy Madhavi, ―Extension of Direct Search Methods to Find Optimal Cluster Centroid for Constrained Multi-Variable Functions,‖ SearchDL, Fourth International Joint Journal Conference in Engineering, IJJCE 2012 organized by ACEEE-_The Association of Computer, Electronics and Electrical Engineers, Int. J. on Recent Trends in Engineering and Technology, Vol. 7, No. 1, July 2012, ISSN 2158-5563 (Online) ISSN 2158-5555 (Print), USA.

[22] Shanno, D. F. 1970. Conditioning of Quasi-Newton Methods for Function Minimization. *Math. Comp.,* 24, 647–657.