www.ijecs.in
International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume 3 Issue 11 November, 2014 Page No. 9193-9196

Study Of Fault Prediction Using Quad Tree Based K-Means Algorithm And Quad Tree Based EM Algorithm

Swapna M. Patil, R.V.Argiddi
Dept.Of Computer science and Engineering,
Walchand Institute Of Technology,
Solapur,413006
swapn_thedreams@yahoo.co.in

Assistant Professor
Dept.Of Computer science and Engineering,
Walchand Institute Of Technology,
Solapur,413006
argiddi@gmail.com

***Abstract*—**The paper intends to do a comparative study of the two clustering algorithms, namely K-Means and EM. Quad tree is used as a common algorithm to initialize both the clustering algorithms. The dataset is then clustered and classified separately by K-Means and EM algorithms. The motive of this paper is to prove the effectiveness of EM over K-Means. Classification and clustering of the dataset done via EM is seen to have lower faults as compared to clustering and classification done via K-Means algorithm

Keywords— Quad Tree, K-Means clustering, Expectation Maximization Algorithm, Iris Dataset, Clustering, Classification, Hyper-Quad tree

1.INTRODUCTION

analysis(supervised classification).In supervised classification we are provided with a collection of labeled(preclassified) patterns; the problem is to label the newly encountered, y Clustering is a division of data into groups of similar objects. Each group, called cluster, consists of objects that are similar amongst themselves and dissimilar compared to objects of other groups. representing data by fever clusters necessarily looses certain fine details, but achieves simplification. It represents many data objects by fever clusters and hence it models data by its clusters[3].

Cluster analysis is the organization of a collection of patterns(usually represented as a vector of measurements or a point in a multidimensional space)into clusters based on similarity. Patterns within a valid clusters are more similar to each other than they are to a pattern belonging to different cluster. It is important to understand the difference between clustering(unsupervised classification) and discriminate et unlabeled patterns. In the case of clustering, the problem is to Clustering is a division of data into groups of similar objects. Each group, called cluster, consists of objects that are similar amongst themselves and dissimilar compared to objects of other groups. representing data by fever clusters necessarily looses certain fine details, but achieves simplification. It represents many data objects by fever clusters and hence it models data by its clusters[3].

Cluster analysis is the organization of a collection of patterns(usually represented as a vector of measurements or a point in a multidimensional space)into clusters based on similarity. Patterns within a valid clusters are more similar to each other than they are to a pattern belonging to different cluster. It is important to understand the difference between clustering(unsupervised classification) and discriminate analysis(supervised classification).In supervised classification we are provided with a collection of labeled(preclassified) patterns; the problem is to label the newly encountered, yet unlabeled patterns. In the case of clustering, the problem is to group the given collection of unlabeled patterns into meaningful clusters.

It’s usage is seen in multivarious fields, including machine learning, pattern recognition, image analysis, information retrieval, and bioinformatics.[4] Unsupervised techniques like clustering may be used for fault prediction [7].

This paper aims to predict faults in the classification of datasets. Many clustering methods exist to partition a dataset by some natural measure of similarity. In this paper a Quad Tree based EM algorithm [3][6] has been applied for predicting faults in the classification of datasets.

2. RELATED WORK

Related work mainly focuses on software fault prediction using different clustering techniques in data mining; concept of Simple K-Means, Quad Tree based K-Means Algorithm. Earlier research on faulty software mechanism permits verification specialists to concentrate their sources and time on the problem areas of the different software systems which are under development. One of the main purposes of these models is to help in software maintenance budgeting. Among various clustering techniques available in literature K-means clustering approach is most widely being used? Different authors apply different clustering techniques and expert-based approach for software fault prediction problem.. Seliya and Khoshgoftaar proposed a constrained based semi-supervised clustering scheme. They proved that this approach helped the expert in making better estimations as compared to predictions made by an unsupervised learning algorithm. A Quad Tree-based K-Means algorithm has been applied for predicting faults in program modules [1]. The aim of their topic is double. Earliest, Quad-Trees are applied for finding the initial cluster centers to be input to the K-Means Algorithm. Bhattacherjee and Bishnu [1] have applied unsupervised learning approach for fault prediction in software module. Bhattacherjee and Bishnu [1] takes an input threshold parameter delta which directs the number of initial cluster centers and by varying it generates the desired initial cluster centers.[1] The clusters obtained by Quad Tree-based algorithm were found to have maximum gain values. Next the Quad-tree based algorithm is applied for predicting faults in program modules. The overall error rates of this prediction approach are compared to other existing algorithms and are found to be better in most of the cases. J.Han and M.Kamber provide a detailed description of the widespread concepts of data mining and the tools required to manipulate data. Fault prediction using quad tree and Expectation Maximization clustering algorithm, limits the research in this book to the section of “Cluster Analysis”. The cluster analysis section in this book gives a detailed description of the different types of clustering methods. This paper concentrates on the working, pitfalls and advantages of the two clustering algorithms namely the K-Means clustering algorithm and Expectation Maximization algorithm.

3.OVERVIEW OF FAULT PREDICTION USING QUAD TREE AND EXPECTATION MAXIMIZATION ALGORITHM.

In this paper, a Quad Tree based *Expectation Maximization (EM)* algorithm has been applied for predicting faults in the classification of datasets. K-Means is a simple and popular approach that is widely used to cluster/classify data. However, K-Means does not always guarantee best clustering due to varied reasons. The proposed EM algorithm is known to be an appropriate optimization for finding compact clusters.

3.1 Quad Tree

This tree data structure was named a Quad tree by Raphael Finkel and J.L. Bentley in 1974. A similar partitioning is also known as a Q-tree. Quad tree (4-ary tree) is the recursive data structure, this tree stands for a division of the matrix into sub matrices (nodes). Leafs of the quad tree are separated into “complete" or "blank" nodes. The Quad Tree-based method assigns the suitable initial cluster centers and removes the outliers hence overcoming the second and third weakness of K-Means clustering algorithm.

3.1.1 General features of quad tree-

 They decompose space into adaptable cells.

 Each cell (or bucket) has a highest capacity.

 When highest capacity is reached, the bucket splits.

 The tree directory tracks the spatial decomposition of the Quad tree.

Fig. 1 shows the simple Quad tree illustration from [1]

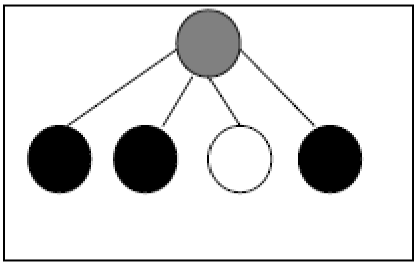


Fig1 Simple Quad Tree

Quad Tree Algorithm

Input Parameters :

1. Raw data (input points)
2. Min value (minimum number of points in a bucket)
3. Max value (maximum number of points in a bucket)

Algorithm :

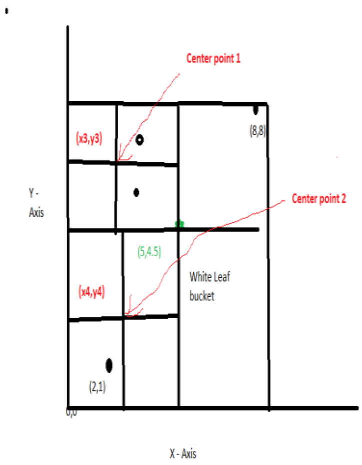
Step 1: Get the point which has the minimum (X,Y) coordinates (x1,y1).

Step 2: Get the point which has the maximum (X,Y) coordinates (x2,y2).

Step 3: Calculate the center point using formula (x1+x2 /2 , y1+y2 /2).

Step 4 : Draw all the points on X-Y axis.

Step 5: Divide the XY axis in 4 quadrant from the center point obtained in step3.



The above dig shows the green point as the center point.
Step 6: Identify the black leaf bucket (quadrant which has more than Max number of points) and White Leaf bucket(quadrant which has less than Min number of points.
Step 7: The centers of each black leaf bucket is calculated (x3,y3) and (x4,y4).
Step 8: The mean of all the center points obtained in the previous step is calculated (x3+x4/2,y3+y4/2).
Step 9 : This gives us our first centroid for K-Means algorithm.
Step10 : Repeat steps from 1 to 8 for black leaf buckets to get the other centroid points.

3.2 K-MEANS ALGORITHM

3.2 K-means

The *k*-means algorithm (KM) [2] partitions data into *k* sets. The solution is then a set of *k* centers, each of which is located at the centroid of the data for which it is the closest center. For the membership function, each data point belongs to its nearest center, forming a Voronoi partition of the data. The objective function that the

Properties of K-Means are outlined below:
1.There are always K clusters.
2. There is always at least one item in each cluster.
3. The clusters are non-hierarchical and they do not overlap.
4. Every member of a cluster is closer to its cluster than any other cluster.

Implementation of K-Means Algorithm

1. Input the centroid points obtained using the quad tree algorithm as the initial cluster centers for the first iteration.
2. Compute distance between each data point and each centroid using the distance formula:
|(x2-x1)|+|(y2-y1)| = distance
3. **Repeat**

- ➢ (Re)assign each data point to the cluster with which it has the minimum distance.
- ➢ Update the cluster means for every iteration.
- ➢ Until clustering converges.

3.3EXPECTATION MAXIMIZATION ALGORITHM

EXPECTATION MAXIMIZATION(EM) is a well established clustering algorithm in the statistics community. EM is a distance based algorithm that assumes the dataset can be modeled as a linear combination of multivariate normal distribution and the algorithm finds the distribution parameter that maximize a model quality measure, called log likelihood. The EM algorithm is an extension of the K-Means algorithm [3][7].

EM is chosen to cluster data for the following reasons among others:

1. It has a strong statistical basis.
2.It is linear in database size.
3.It is robust to noisy data.
4.It can accept the desired number of clusters as input.
5.It can handle high dimensionality.
6.It converges fast given a good initialization.

Implementation of E-M Algorithm

The general E-M algorithm is comprised of the following simple steps:

1. **Initialization**

Initialize the distribution parameters, such as the means, covariances and mixing coefficients and evaluate the initial value of the log-likelihood (the goodness of fit of the current distribution against the observation dataset)';

2. **Expectation**

Evaluate the responsibilities (i.e. weight factors of each sample) using the current parameter values;

$$p(x) = \frac{1}{(2\pi)^{k/2}|\Sigma|^{1/2}} \exp\left(-\tfrac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)\right)$$

3. **Maximization**

Re-estimate the parameters using the responsibilities found in the previous step;

4. **Repeat**

Re-evaluate the log-likelihood and check if it has changed; if it has changed less than a given threshold, the algorithm has converged.

4.CONCLUSION

This paper reviews the problems with using simple K-Means in the classification of datasets [6]. The effectiveness of Quad Tree based EM clustering algorithm in predicting faults while classifying a dataset, as compared to other existing algorithms such as, K-Means has been evaluated. The Quad Tree approach assigns appropriate initial cluster centers and eliminates the outliers. K-Means is considered to be one of the simplest methods to cluster data [1]. However, the proposed EM algorithm is used to cluster data effectively. Combining the Quad Tree approach and the EM algorithm gives a clustering method that not only fits the data better in the clusters but also tries to make them compact and more meaningful. Using EM along with Quad Tree makes the classification process faster. With K-means, convergence is not guaranteed but EM guarantees elegant convergence.

5.REFERENCES

[1] P.S. Bishnu and V. Bhattacherjee, Member, IEEE" Software Fault Prediction Using Quad Tree-Based K-Means Clustering Algorithm" IEEE Transactions on Knowledge and Data Engineering, Vol. 24, No. 6, June 2012

[2] J. MacQueen. Some methods for classification and analysis of multivariate observations. In L. M. LeCam and J. Neyman, editors, Proceedings of the Fifth Berkeley Symposium on Mathematical *Statistics and Probability*, volume 1, pages 281–297, Berkeley, CA,1967. University of California Press.

[3] J. Han and M. Kamber, Data mining Concepts and techniques, 2nd edition, Morgan Kaufmann Publishers, pp. 401-404, 2007.

[4] http://en.wikipedia.org/wiki/Cluster_analysis

[5] Leela Rani.P, Rajalakshmi.P," Clustering Gene Expression Data using Quad-tree based Expectation Maximization Approach" International Journal of Applied Information Systems (IJAIS) – ISSN : 2249-0868 Foundation of Computer Science FCS, New York, USA Volume 2– No.2, June 2012 – www.ijais.org

[6] Osama Abu Abbas, Computer Science Department, Yarmulke University, Jordan, "Comparisons between data clustering algorithms" The international Arab Journal of Information Technology,Vol.5,No.3,July 2008.

[7] P.S.Bishnu and V. Bhattacharjee, "Software Fault prediction using Quad tree based K-Means method," IEEE transactions on Knowledge and Data Engineering ,Vol. PP, No.99, May 2011