International Journal of Applied Information Systems (IJAIS) – ISSN : 2249-0868
Foundation of Computer Science FCS, New York, USA
Volume 2– No.4, May 2012 – www.ijais.org

Fault Prediction using Quad Tree and Expectation Maximization Algorithm

Meenakshi PC, Meenu S, Mithra M
Dept of Information Technology
Sri Venkateswara
College of Engineering,
Chennai- 602105

Leela Rani P
Assistant Professor
Dept of Information Technology
Sri Venkateswara
College of Engineering,
Chennai- 602105

ABSTRACT
The objective of the paper is to predict faults that tend to occur while classifying a dataset. There are various clustering algorithms that prevail to partition a dataset by some means of similarity. 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. EM guarantees elegant convergence. EM algorithm assigns an object to a cluster according to a weight representing the probability of membership. EM then iteratively rescores the objects and updates the estimates. The error-rate for K-Means algorithm and EM algorithm are computed, denoting the number of correctly and incorrectly classified samples by each algorithm. Result consists of charts showing on a comparative basis the effectiveness of EM algorithm with quad tree for fault prediction over the existing Quad Tree based K-Means (QDK) model.

General Terms
Algorithms, Clustering, Partition based clustering, Model based clustering, Classification

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

1. INTRODUCTION
Clustering can be considered the most dominant unsupervised learning technique. Clustering invariably focuses on finding a particular pattern or structure, in a collection of unlabeled data. Clustering falls under the domain of explorative data mining. It is a common technique for statistical data analysis. It's usage is seen in multifarious 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* [1][6] has been applied for predicting faults in the classification of datasets. The overall error-rates of this prediction approach are compared to other existing algorithms such as K-Means and are found to be better in most of the cases. This paper focuses on clustering by partition based method namely K-Means algorithm and model based method namely, EM algorithm. Fig.1 [4] indicates a sample scatter plot diagram clustered using the partitioning based K-Means algorithm. Fig.2 [4] indicates a sample scatter plot clustered using the EM algorithm.

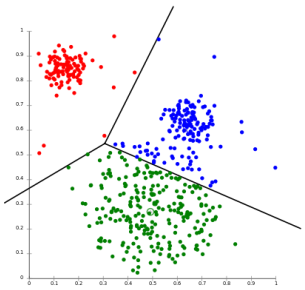

Fig.1

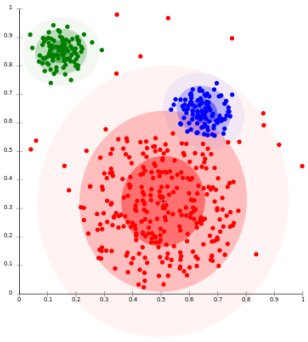

Fig.2

The objectives of this paper are as follows: First, Quad trees are applied for finding initial cluster centers for K-Means algorithm. User can generate desired number of cluster centers that can be used as input to the simple K-Means

algorithm. Second, the centroids obtained by the quad tree data structure are input to the EM algorithm to yield better results. Third, results are shown via charts indicating better throughput using the proposed system.

The ultimate goal is to improve the accuracy of fault prediction by using EM clustering algorithm. The increase in throughput is attributed to the fact that the EM algorithm is a soft clustering method. It is an extension of the K-Means clustering algorithm. In addition to clustering based on minimizing intra cluster distance, probability is calculated for each combination of data point and cluster. Clustering is done based on a weighted relationship thus derived. A significant decrease in error rates is observed through the proposed system implementation.

2. RELATED WORK

P.S. Bishnu and V. Bhattacherjee applied unsupervised techniques like clustering for fault prediction in software modules, more so in those cases where fault labels are not available. Their paper elicits a Quad Tree based K-Means algorithm for predicting faults in software modules or datasets. They have used a concept of clustering gain to determine the quality of clusters for evaluation of the Quad Tree based initialization algorithm as compared to other initialization techniques.

M. Laszlo and S. Mukherjee propose the usage of Hyper-Quad trees (HQ) as the initialization algorithm for finding the initial cluster centers/centroids that serve as input to various clustering algorithms such as K-Means, EM. The usage of HQ trees instead of Quad trees is however left to open research as the future work of the paper. Related work in this domain can be pursued to achieve increased efficiency in the computation of centroids derived from the initialization algorithm.

Osama Abu Abbas has made a vast comparative study of the various clustering algorithms. He has intended to study and compare different data clustering algorithms. The algorithms under investigation are: K-Means algorithm, Hierarchical Clustering algorithm, self organizing maps algorithm and Expectation Maximization clustering algorithm. All these algorithms are compared according to the following factors: size of dataset, number of clusters, type of dataset and type of software used. Some conclusions that are extracted belong to the performance, quality and accuracy of the clustering algorithms. This study by Osama Abu Abbas helps to compare K-Means and Expectation Maximization algorithms and proves that EM is more accurate than K- Means.

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 BASED K-MEANS ALGORITHM AND QUAD TREE BASED EM ALGORITHM

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.

3.1 Quad Tree - Properties

The initial cluster centers are found using a quad tree based algorithm [11] [8]. A quad tree is a tree data structure in which each internal node has exactly four children. Quad trees are most often used to partition a two dimensional space by subdividing it into four quadrants or regions [10]. The regions may be square or rectangular, or may have arbitrary. This data structure was named a quad tree by Raphael Finkel and J.L.Bentley in 1974. The cluster centers, thus found, serve as input to the clustering algorithms. Fig.3 [7] depicts a simple Quad tree

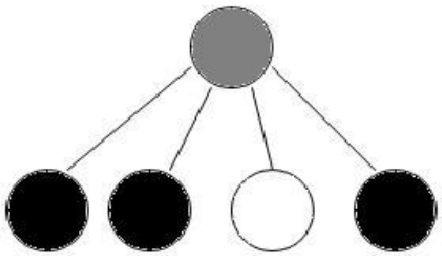

Fig.3

All forms of Quad trees share some common features:

- They decompose space into adaptable cells.
- Each cell (or bucket) has a maximum capacity. When maximum capacity is reached, the bucket splits.
- The tree directory follows the spatial decomposition of the Quad tree.

3.1.1 The Initialization Algorithm

1. Initialize MIN & MAX values.
2. Classify using Sepal Length as the parameter along the X axis.
3. Classify using Sepal Width as the parameter along the Y axis.
4. For each species:
   - Find the minimum and maximum x and y co-ordinates.

- Find the midpoint using the values obtained in the previous step.
- Divide the spatial area into four sub regions based on the midpoint.
- Plot the points and classify regions as white leaf buckets or black leaf buckets.
- The white leaf buckets are left as such.
- The Center data-points of each black leaf bucket are calculated for all black leaf buckets.
- The mean of all the center points obtained in the previous step is calculated.
- The computed mean gives the centroid point necessary for that species.

3.1.2 Parameters and Definitions

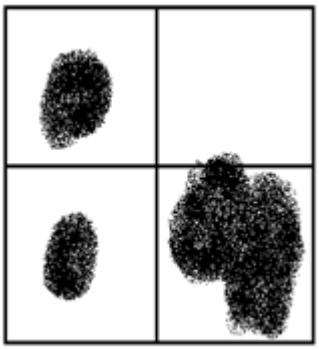

Fig.4

1. **MIN**

User defined threshold for minimum number of data points in a sub bucket.

2. **MAX**
User defined threshold for maximum number of data points in a sub bucket.

3. **White leaf bucket**
A sub bucket having less than MIN number of data points of the parent bucket.
Fig.4 shows an illustration of a white leaf bucket.

4. **Black leaf bucket**
A sub bucket having more than MAX number of data points of the parent bucket.

3.2 K-MEANS ALGORITHM

3.2.1 Description

K-Means [1] is an unsupervised clustering method where observations are iteratively relocated among a set of clusters until the convergence criterion is met. This popular algorithm follows a partitional clustering approach in which, partitioning method creates initial partitioning. It then uses iterative relocation technique that attempts to improve partitioning by moving objects from one group to another. K-Means clustering is simple, fast and widely used approach to classify or cluster data.

The algorithm begins with finding the initial centroids[10] for potential clusters ie, each cluster is associated with a centroid. The observations are assigned to each cluster based on its distance from the centroid. The partitioning of dataset is such that the sum of intra-cluster distances is reduced to an optimum value. A popular heuristic method is adopted where each cluster is represented by the mean value of observations in the cluster. The data points are reassigned and the algorithm runs until the convergence criterion is met or until relatively few observations change clusters.

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.

3.2.2 Working 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:

$$|(x2-x1)|+|(y2-y1)| = \text{distance}$$

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.

3.2.3 Shortcomings of K-Means Clustering

K-Means algorithm attempts to find best clusters for the observations or clusters that are well distributed. However, it may not always guarantee best clustering due to varied reasons. K-Means typically requires that the clusters be of spherical shape. It does not eliminate outliers and hence it requires that the data be free of noise. The algorithm is very sensitive to the initially selected centroids. It requires a huge training set to begin the clustering. K-Means does not assure best representation of the data in the clusters .Under some conditions, K-Means clustering algorithm may not converge at all [9].

3.3 EXPECTATION MAXIMIZATION ALGORITHM

3.3.1 Description

Expectation Maximization is a type of model based clustering method. It attempts to optimize the fit between the given data and some mathematical model. Such methods are often based on the assumption that the data are generated by a mixture of underlying probability distributions. The EM algorithm is an extension of the K-Means algorithm [1][7].

In practice, each cluster can be represented mathematically by a parametric probability distribution. The entire data is a mixture of these distributions. The Expectation Maximization algorithm is a popular iterative refinement algorithm that can be used for finding parameter estimates. It can be viewed as

an extension of the K-Means paradigm, which assigns an object to the cluster with which it is most similar based on the cluster mean. Instead of assigning each object to a dedicated cluster, EM assigns each object to a cluster according to a weight representing the probability of membership.

Hence in addition to working towards minimizing the Euclidean distance, it also takes into account the probability of membership of each data point to each cluster. Therefore clustering is performed based on two conditions. EM clustering hence shows higher throughput.

3.3.2 Working of Expectation Maximization Algorithm

1. Input the centroids obtained using the quad tree algorithm as the initial cluster centers.
2. Compute distance between each data point and each centroid using distance formula:

$$|(x2-x1)| + |(y2-y1)| = \text{distance}$$

3. Assign weights for each combination of data point and cluster based on the probability of membership of a data point to a particular cluster.

4. Repeat

- (Re)assign each data point to the cluster with which it has highest weight ie, highest probability of association.
- If a data point belongs to more than one cluster with the same probability, then (re)assign the data point to the cluster based on minimum distance.
- Update the cluster means for every iteration
- **Until** clustering converges [1].

3.3.3 Advantages of Fault Prediction using Quad Tree and Expectation Maximization Algorithm

The benefits of using EM as a replacement to K-Means algorithm are observed as follows:
1. The algorithm meets the convergence criterion faster and hence it results in lesser number of iterations.
2. Reduction in time and computational complexity
3. Will work despite limited memory (RAM)
4. Better throughput with lower error rates of classification

4. EXPERIMENTAL DESIGN

4.1 Dataset

The dataset that has been used for the purpose of experimental design in this paper is the popular Iris dataset [2]. It is a multivariate dataset. This is perhaps the best known database to be found in the pattern recognition literature. Fisher's paper is a classic in the field and is referenced frequently to this day. The data set contains 3 classes of 50 instances each, where each class refers to a type of iris plant. One class is linearly separable from the other 2; the latter are NOT linearly separable from each other. Predicted attribute is the class of iris plant. This is an exceedingly simple domain. There are 4 attributes in this dataset. The attribute information are as follows:
1. sepal length in cm
2. sepal width in cm
3. petal length in cm
4. petal width in cm
5. Class: Iris setosa, Iris versicolor, Iris virginica.

4.2 Comparison Metric/ Evaluation Parameter

Classification of the dataset using both the clustering algorithms proves that EM is a better fit [6] for clustering than K-Means. The two algorithms are compared for higher accuracy and efficiency using the metric "Error Rate". The evaluation parameters are the correctly classified and incorrectly classified data points. Based on these parameters, the error rate is evaluated using the following formula:

- Per Species:

$$\underset{\text{Total}}{\text{Actual}} - \underset{\text{Classified}}{\text{Correctly}} = \underset{\text{Classified}}{\text{Incorrectly}}$$

- Overall Error Rate,

$$\text{Error} = (TI) / (TC+TI)$$

Table 1

Actual Label	Predicted Label (Correctly Classified)	Predicted Label (Incorrectly Classified)
Species 1	Precisely Labeled	Mislabeled
Species 2	Precisely Labeled	Mislabeled
Species 3	Precisely Labeled	Mislabeled
TOTAL	Summation	Summation

Table 1 illustrates the blueprint of the table used to calculate the error rates for each algorithm. The actual number of instances in every species of flower is listed in the left corner of the table. After classification using any one of the clustering algorithms, the number of correctly classified and incorrectly classified data points are determined and noted in the right hand side of the table. This enables the computation of the error rates.

4.2.1 Parameters and Definitions

TI - total number of incorrectly classified species

TC - total number of correctly classified species

5. RESULTS

Using the formulae outlined above, error rates are calculated for both the clustering algorithms separately. The computed result is shown via charts for comparison purposes. Fig.5 indicates a bar chart that shows the comparison of error rates for both the algorithms. It proves that EM algorithm is more accurate than K-Means owing to lower error rates as shown. Visibly lower error rates are seen for the EM clustering algorithm, when compared to K-Means.

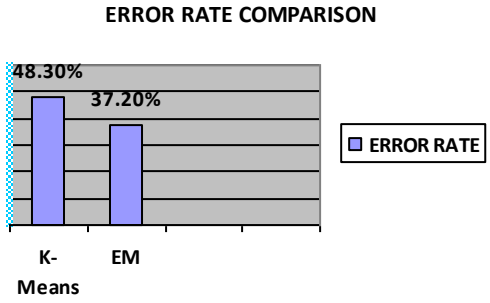

Fig.5

6. 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.

The proposed approach starts with a huge set (the popular Iris dataset [2]). The proposed system obtains the appropriate initial cluster centers through Quad Tree. These centroids serve as input to the EM algorithm, thus increasing the chances of finding the best clusters. The overall error rates of the proposed system are found comparable to other existing approaches. In fact, in the case of the Iris dataset, the overall error rates of the proposed approach have considerably reduced and are fairly acceptable. Results are shown, via charts indicating the effectiveness of the proposed approach.

7. FUTURE WORK

An extension of this paper would be to use a HQ Tree based EM clustering model [3]. The HQ tree is used as a replacement to the traditional Quad Tree approach so as to obtain even more precise cluster centers/centroids. A HQ tree is a D-dimensional analogue of a quad tree. Every node of a HQ tree is associated with a bounding hyper box and every non leaf node has 2D children. Thus HQ Trees are expected to yield better cluster centers as compared to the Quad Tree approach.

.

8. ACKNOWLEDGEMENTS

The authors thank Dr. R Ramachandran Ph.D, Principal Sri Venkateswara College Of Engineering for his motivation towards the accomplishment of this research. The authors also thank Dr. G Sumathi Ph.D., Head of the Department, Information Technology, for her suggestions and support.

9. REFERENCES

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

[2] http://archive.ics.uci.edu/ml/datasets/Iris

[3] M. Laszlo and S. Mukherjee, "A Genetic Algorithm Using Hyper-Quad trees for Low-Dimensional K-Means Clustering," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 28, no4, pp. 533-543, 2006.

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

[5]http://infolab.stanford.edu/~ullman/fcdb/oracle/or-jdbc.html#0.1_create

[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

[8]P.S.Bishnu and V. Bhattacherjee, "Outlier Detection Technique Using Quad Tree," Proc of Int. conf. on Computer Communication Control and Information Technology, pp. 143-148, Feb 2009.

[9]T.Kanungo, D.M. Mount, N.Netanyahu, C. Piatko, R.Silverman and A.Y. Wu, "An Efficient K-Means clustering Algorithm: Analysis and Implementation", IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 24, pp.881-892, 2002.

[10]P.S. Bishnu and V. Bhattacherjee, "A New Initialization method for K-Means using Quad Tree," Proc of National. conf. on Methods and Models in Computing, JNU, New Delhi, pp. 73-81, 2008.

[11]R.A. Finkel and J.L. Bentley, *Quad Trees: a Data Structure for Retrieval on Composite key*. Acta information, vol. 4, no. 1, pp. 1-9, 1974