International Journal of Data Mining Techniques and Applications
Volume: 04 Issue: 01 June 2015 Page No.48-51
ISSN: 2278-2419

Overviewof Data Mining Techniquesand Image Segmentation

P.Sujatha[1], K.K.Sudha[2]
[1]AssociateProfessor, Department of Computing Science,Vels University, Chennai, India
[2]Research Scholar, Department of Computing Science, Vels University , Chennai, India
Email: suja.research@gmail.com, sudhaphd.cs@gmail.com

Abstract - Today data mining has become a vital role in all fields. This is because they discover interesting patterns and relationship in a data repository. Data mining is suitable for various fields such as image processing, artificial intelligence, machine learning, statistics and computation capabilities. Image segmentation is the fundamental step in various image processing tasks such as image analysis, visualization, and object representation and so on. The goal of image segmentation is to simplify, to partition image into meaningful regions and easier to analyze. This paper presents an overview of various data mining techniques associated with image segmentation. The data mining techniques such as clustering, classification and association are very easy to implement image segmentation that delivers valuable results. The above techniques are helpful to retrieve information from the images that are used to diagnose diseases, face detection, to improve the segmentation quality and so on. This paper provides crispy report about the data mining techniques and how they fetch the information from the images.

Keywords: Data Mining, Class, Association Rules, Clustering, Classification, Segmentation.

I. INTRODUCTION

"A picture is worth a thousand words" is a popular notation which means a lot of information can be easily conveyed with the help of an image than descriptive text. Images are widely used in various fields including medical, automotive, textile, computer and more. Data mining is an emerging technology that helps inretrieving information from large databases. Data mining is complimenting various industries including business, science, medical, education, finance and more. Image segmentationisused with several data mining techniques to diagnose diseases, identify and locate tumors, to detect the fault portion in yarn and patterns in textile, forensic, applications based on satellite images and so on. This paper showcases concepts about image segmentation, data mining and a literature survey on how data mining techniques are used with image segmentation to solve a particular problem.

II. LITERATURE SURVEY

Kun-CheLuet.al[3] proposed an effective and efficient model in image data mining using image segmentation. Decision tree, a data mining technique is used for segmenting the image in pixels and stored in a database. Each pixel in the database holds a label and feature. Using these labels and features author developed a model for discovering the relationship between the attributes of pixels and their target labels. ChinkiChandhoket.al[6] had used k-means algorithm to improve the segmentation quality of precision and computational time. They have derived clusters based on the color and spatial features. Good results are obtained for smaller k – values rather than the bigger value. This model was run for several number of times with the same and different values to analyze the quality of clusters obtained.The result of this experimentimplies to improve the quality of precision and computational time in image segmentation.Kun-CheLuet.al[1] found the important pixels of an image using the decision tree technique of data mining.

Then they classify the pixels as important and unimportant using simple rules and appliesa Huffman algorithm for image compression. Finally the resulting image should be stored with lesser space complexity. Difficulties during the practical implementation ofthismodel are: Due to huge redundancy of the given image pixel in the image segmentation quality is not achieved fully, it need label information of an image pixel in advance and it required to find the actual hidden properties for undetermined image pixel.The refined work of this proposed model is to be an unsupervised one. The model should automatically analyze and choose the label information for further use.P. Rajendranet.al[5] proposed a hybrid image mining technique which enhances the classification process of brain tumor is to be more accurate and sensitive. In this technique, Frequent Pattern Tree (FP – Tree) algorithm is used to mine association rules and decision tree algorithm is used to classify the brain tumor cells in a dynamic way. The hybrid image mining technique results in 97% sensitivity and 95% accuracy to classify the brain tumor. Petra Perner[9] designedadecision tree toolfordiagnosing lung cancer. This tool extracts information from large image databases (x-ray). This method has garnered good result and therefore it may be used lymph node diagnosis.

III.IMAGE SEGMENTATION

Image segmentation is the process of segregating the images into pixels according to the features. Various image segmentation algorithms and techniques are used to partition the image. It is also considered to be an important aspect in image analysis. In order to avoid false information during image segmentation, image de-noising process is implemented. Basically, image segmentation process is broadly classified into two categories – Detecting the Discontinuities and Detectingthe Similarities that are based on image properties.In detecting the discontinuities, the images are partitioned mainly based on isolated points, lines and edges.In detecting the similarities, the

International Journal of Data Mining Techniques and Applications
Volume: 04 Issue: 01 June 2015 Page No.48-51
ISSN: 2278-2419

images are partitioned into regions using thresholding, region growing, region splitting and merging techniques.

A. Threshold Based Segmentation

Techniques like histogram thresholding and slicing are used to segment the image. These techniques are directly enforced in the image or combined with pre and post processing techniques. Figure-1 shows the boundaries of obtained by thresholding.

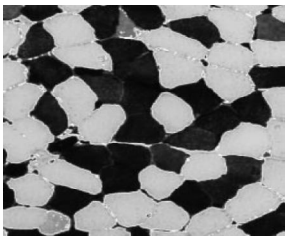

FIGURE-1 THRESHOLD BASED SEGMENTATION

B. Edge Based Segmentation

In this segmentation, the pixels in the image are classified as edge or non-edge by applying the edge filter. Pixels which are detached from an edge are set aside as same category. Figure-2 shows the connected region boundaries and all non-border segments are eliminated.

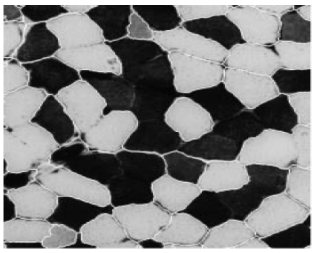

FIGURE-2 EDGE BASED SEGMENTATION

C. Region Based Segmentation

It is an iterative based technique, which cluster the neighbor pixel having similar values and pixels having dissimilar values forms another cluster. Figure-3 shows the boundaries are marked using this technique.

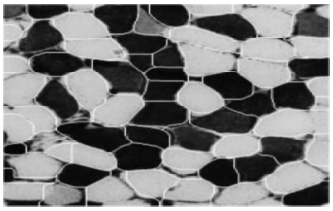

FIGURE-3 REGION BASED SEGMENTATION

III. DATA MINING TECHNIQUES

The main objective of data mining is to discover the hidden knowledge from huge databases. Generally, data mining task can be classified into two categories – descriptive data mining and predictive data mining. The objective of descriptive data mining is to derive patterns (correlation,trends,trajectories) that summarizes the underlying relationship between data. For predictive data mining, the value of a specific attribute (target/dependent variable)is predicted on the bases of other attribute (explanatory) values. The goal of a data mining system can be achieved by using the following data mining tasks.

A. Association Rules

It is used to detect the relationship or associations between specific values of categorical variables in large data sets. The goal of association rule is to form a large number of rules from data repository which are useful to the users. Confidence and support are the two criteria in association rules to form relationship from a large database. An association rule is the significance of two item sets, X → Y.

- *Support:*The frequent number of transactions that contain both X and Y.

Support = Number of Occurrence / Total Support

TABLE-1 MEASURING SUPPORT

Number	Items	Support
1	XYZ	Total Support = 5
2	XYW	Support {XY} = 2/5 =40%
3	YZ	Support {YZ} = 3/5 = 60%
4	XZ	Support {XYZ} = 1/5 = 20%
5	YZW	

- *Confidence:* It measures how often items in Y appear in transactions that contain X.

Confidence = Number of 'X' Occurrence / Number of 'Y' Occurrence.

TABLE-2 MEASURING CONFIDENCE

Number	Items	Confidence
1	XYZ	Confidence {X => Y} = 2/3 = 66%
2	XYW	Confidence {Y => Z } =3 / 4 = 75%
3	YZ	Confidence {XY => Z} = 1 / 2 = 50%
4	XZ	
5	YZW	

This powerful exploratory techniquehas a wide range of applications in many areas of business practice and also research - from the analysis of consumer preferences or human resource management, to the history of language. These techniques enable analysts and researchers to uncover hidden patterns in large data sets, such as "customers who order product A often also order product B orC".

B. Clustering

Cluster analysis group objects (observations, events) based on the information found in the data describing the objects or their relationships. The goal is that the objects in a group will be similarly (or related) to one another and different from (or unrelated to) the objects in other groups. The greater the similarity (or homogeneity) within a group, and the greater the difference between groups, the "better" or more distinct the clustering.Clustering methods are classified into 5 approaches – partitioning algorithms, hierarchical algorithm, density – based method, grid-based method, model-based method.

International Journal of Data Mining Techniques and Applications
Volume: 04 Issue: 01 June 2015 Page No.48-51
ISSN: 2278-2419

C. Partitioning Based Algorithm

This algorithm minimizes the data clustering criterion by iterative relocating data points between clusters until a (locally) optimized partition is obtained. In a Data set D, with 'n' objects and 'k' number of clustersto be designed, the partitioning based algorithmcoordinates the objects into k –partitions (k≤ n), where in each partition represents a cluster C1,….Ck. K-means and k-medoids are commonly used partitioning based algorithm. In k-means each cluster is characterized by the center of the cluster and in k-medoids each cluster is characterized by one of the objects in the cluster.In K-means technique the centroid of a cluster is treated as the center point. The centroid Ci can be found by mean or medoids of the objects and referred to the cluster. The difference between an object 'p' (belongs to Ci) and Ci, is calculated by Dist (p, Ci), the distance between the objects are forecast by Euclidean distance. The main aim of k-means clustering is to reduce the total intra cluster variance or squared error function.

$$E = \sum_{i=1}^{k} \sum_{p \in C_i} dist(p, c_i)^2$$

Where 'E' is the sum of squared error, 'k' is the number of clusters, Ci is the centroid of acluster, and 'p' is the object assigned to the cluster.

D. K-means Algorithm

Step 1: Partition objects into 'k' number of clusters.
Step 2: Each cluster is identified with a centroid (center point) and every object is assigned to cluster with closest centroid.
Step 3: when all the objects are selected in a cluster, the centroid position Ci is recalculated.
Step 4: Step 2 and 3 are repeated until the centroid position is said to standard.

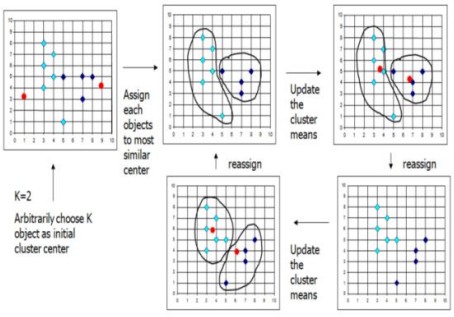

FIGURE-4 K - MEANS ALGORITHM

Data objects are placed in 2 dimensional space and 'k' value is 2 where the objects are partitioned into two clusters. On the basis of the algorithm, the centroids of the clusters are marked as red. Each object is grouped at the base of minimal distance from thecentroid. The centroid of each cluster is reassign for several iterations until the intra cluster variance or squared error function is reduced.K-medoid is another partitioning technique of cluster analysis that is similar to k-means approach. It clusters the data set consisting of 'n' objects into 'k' number of clusters where 'k' value is given by the user. This technique chooses an object randomly from a dataset 'D'as an initial representative object called medoid. The medoid is centrally located object in a datasetand objects with minimal distance from the centroid form acluster. The approach minimizes the sum of dissimilarities between objects and its centroid.

$$E = \sum_{i=1}^{k} \sum_{p \in C_i} dist\,(p, O_i)$$

Where 'E' is the sum of the absolute error, 'k' is the number of clusters, Ci is the centroid of a cluster, 'p' is the object assigned to the cluster and Oi is present centroid of the cluster.

E. K-medoidsAlgorithm

Step 1: Arbitrarily choose 'k' objects in D as the initial representative objects.
Step 2: Assign each remaining object to the cluster with the nearest representative objects
Step 3: Randomly select a non-representative objects, Orandom
Step 4: Compute the total cost, S, of the swapping representative object, Oj, with Orandom
Step 5: If S < 0 then swap Oj with Orandom to form the new set of 'k' representative objects.
Step 6: Repeat Step 2 to 5 until no change in centroid.

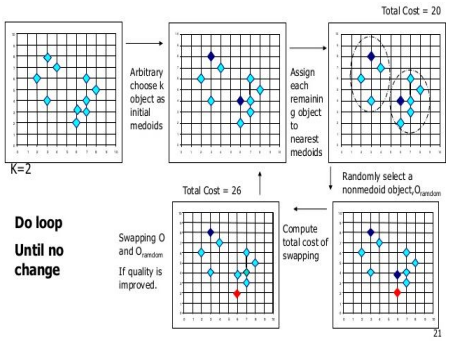

FIGURE-5 K - MEDIOD ALGORITHM

Data objects are placed in 2 dimensional space and 'k' value is 2 where the objects are partitioned into two clusters. Initially, 'k' objects are selected as the initial medoids and thenearest objects to the medoids form clusters. Now, non-medoidobjectOrandom is selected and the total cost of swapping the objects is computed. Oj and Orandom are swapped to improve the quality of a cluster. This process will be executed until the sum of dissimilarities between objects and its centroid is minimized. A good clustering method will produce high quality clusters in which – the intra class (i.e intra cluster) similarity is high and the inter class similarity is low. The quality of clustering is measured by the definition and representation of clustering chosen, its ability to discover some or all of the hidden patterns.

F. Classification

Classification is a Data Mining (machine learning) technique that predicts the group of data items. A trained data set is the input for classification whose class label is already known. The classification technique analysis the trained data set and frame a pattern to accredit a class label for a subsequent unlabeled data set. Since classification use known dataset to frame patterns it comes under supervised learning. Decision Tree, Support Vector Machine (SVM), Neural networks, nearest neighbor (kNN), Naïve Bayes classifier are some of popular

International Journal of Data Mining Techniques and Applications
Volume: 04 Issue: 01 June 2015 Page No.48-51
ISSN: 2278-2419

classification techniques. Classification is used in image segmentation to retrieve information from images.The decision tree is a powerful technique in classification model and it uses a tree structure to solve complex problems. By using this technique, simple and multiple possible solutions can be produced. In decision tree the relationship between different events or decision are in an easyto understand format. Each internal node in adecision tree (non-leaf node) is the condition of an attribute, whereas each branch is the result of the test condition, and each leaf node (or terminal node) indicates a class label. Finally, the root node is thetopmost node of a tree. Figure - 6explains how a general structure of a decision tree will work.

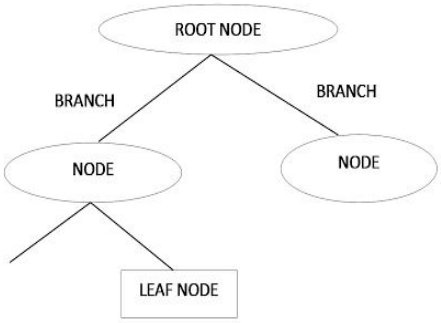

FIGURE-6 GENERAL STRUCTURE OF A DECISION TREE

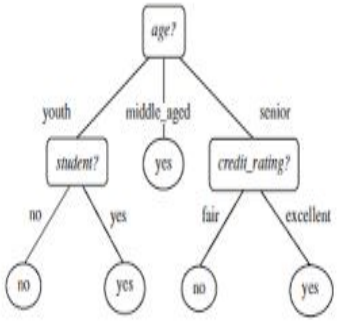

FIGURE-7 SIMPLE DECISION TREE EXPLAINED

In Figure – 7decision tree is constructed to predict if a customer will buy an itemized computer or not. A decision tree is created by adding series of questions and possible solutions. Each node has a subset explaining the possible outcomes. In the above given example, a purchase decision is evaluated based on age (youth, middle_aged, senior).

IV. CONCLUSION

In this paper, image segmentation and various data mining techniques are reviewed that are extensively used in diverse fields.This review confered deeper insightsabout clustering technique along with image segmentation. It has become a disruptive technology, which has been widely used in enormous fields including Medical, Textile ,Meteorology and more. However, data miningisphenomenallysuccessful, some bottlenecks that may arise which has to be addressed in the near future.

REFERENCES

[1] Kun-Che Lu, Don-Lin Yang.,'Image Processing and Image Mining using Decision Trees', International Conference on Extended Database Technology, 2006, pp. 35-40.
[2] http://www.bioss.ac.uk/people/chris/ch4.pdf.
[3] Kun-Che Lu, Don-Lin Yang and Ming-Chuan Hung., 'Decision Tree Based Image Data Mining and Its Application on Image Segmentation', Journal Information Science and Engineering 25, 989 – 1003 (2009).
[4] http://www.cs.toronto.edu/~jepson/csc2503/segmentation.pdf
[5] P. Rajendran, M. Madheswaran., 'Hybrid Medical Image Classification Using Association Rule Mining with Decision Tree Algorithm', Journal of Computing, Volume 2, Issue 1, January 2010, ISSN 2151 – 9617.
[6] ChunkyChandhok, SoniChaturvedi, A. A. Khurshid., 'An Approach to Image Segmentation using K-means Clustering Algorithm', International Journal for Information Technology (IJIT), Volume – I, Issue – I August 2012, ISSN 2279 – 008X.
[7] http://www.academia.edu/648890/Support_vsConfidence_in_Association_Rule_Algorithms
[8]]http://www.cs.put.poznan.pl/jstefanowski/sed/DM-7clusteringnew.pdf
[9] Petra Perner,'Mining Knowledge in Medical Image Databases' In Data Mining and Knowledge Discovery: Theory, Tools, and Technology, Belur V. Dasarathy (Eds), Proceeding of SPIE Vol. 4057 (2000), 359 – 369.
[10]https://courses.cs.washington.edu/courses/cse576/book/ch10.pdf
[11]S. Hameetha Begum, 'Data Mining Tools and Trends – An Overview', International Journal of Emerging Research in Management & Technology', February 2013, ISSN: 2278 – 9359.
[12]Xiao – Feng Wang, De – Shuang Huang, HuanXu., 'An Efficient Local Chan-Vese Model for Image Segmentation', Elsevier, Volume 43, Issue 3, March 2010, Pages 603 – 618.
[13]George Karypis, Vipin Kumar., 'A Fast and High Quality Multilevel Scheme for Partitioning Irregular Graphs' SIAM Journal of Scientific Computing, Volume 20, Issue 1, Pages 359 – 392.
[14]Anil K.Jain, 'Data Clustering: 50 Years beyond K – Means', 19th International Conference in Pattern Recognition (ICPR), Volume 31, Issue 8, 1 June 2010, Pages 651 – 666.