Journal of Theoretical and Applied Information Technology
15th December 2011. Vol. 34 No.1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

CLASSIFICATION OF IMAGES USING JACCARD CO-EFFICIENT AND HIGHER-ORDER CO-OCCURENCES

[1]K. ANURADHA, [2] N. SAIRAM

[1]Asst Prof., School of Electrical Electronics Engineering, SASTRA University, Thanjavur, Tamil Nadu, 613401, India

[2] Prof., School of Computing, SASTRA University, Thanjavur, Tamil Nadu, 613401, India

E-mail: [1]kanuradha@ece.sastra.edu, [2]sairam@cse.sastra.edu

ABSTRACT

In this paper, we propose a clustering algorithm to deal with the medical image categorization. The algorithm implements an approach, which accepts a set of medical images as categorical input. The categorical inputs are compared with the given pool of images and the result is given in the form of boolean points. The Jaccard co-efficient similarity method does the classification by identifying the neighbors. The higher order co-occurrences with χ-sim Algorithm[2], which has been used for text categorization is implemented for identifying the similarity between images. The proposed approach is tested on images of different sets

Keywords: *Image Categorization, Clustering, Higher-Order Co-Occurrence*

1. INTRODUCTION

Medical images are compared by extracting some of the features in them. In this paper, each of the sample images is compared with the images in the database. The main issue is, traditional algorithms use euclidian distance to find the similarity between points, which results in metric space. The proposed approach solves this issue, by using non-metric space, since similarity measure is non-metric. The proposed approach does image comparison through feature extraction, thereby exhibit the relationships among various groups. In this paper, content-based comparison is done for images. The term 'content' in this context refers to colors, shapes, textures, or any other information that can be derived from the image itself. Based on the results of the comparison, a boolean matrix is formed. Each element in the boolean matrix is checked for finding the links. All the links together forms a link matrix. Then, the links are grouped to form clusters. The higher order similarity between categorical inputs and similarity between pool of images and vice versa is determined.

2. RELATED WORK

2.1 Image Classification For Content-Based Indexing [1]

They use a Bayesian classification approach, using vector quantization to learn the class-conditional probability densities of the features. This approach has the following advantages:

1) Small number of codebook vectors represent a particular class of images, regardless of the size of the training set;

2) It naturally allows for the integration of multiple features through the class-conditional densities;

3) It not only provides a classification rule, but also assigns a degree of confidence in the classification, which may be used to build a reject option.

The disadvantage of this method is they are not taking the input as categorical attributes.

2.2 Clustering

Clustering is similar to classification in order to group the given data set. The grouping is accomplished by finding similarities between them. The groups are called clusters. The clustering algorithms are classified into several types. Some of them are discussed here. For further details on clustering see[4],[5].

2.2.1 Hierarchical algorithms

Hierarchical algorithms find successive clusters using previously established clusters. These

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

algorithms usually are either agglomerative ("bottom-up") or divisive ("top-down"). Agglomerative algorithms begin with each element as a separate cluster and merge them into successively larger clusters. Divisive algorithms begin with the whole set and proceed to divide it into successively smaller clusters.

2.2.2 Partitional algorithms

Partitional algorithms typically determine all clusters at once. Under this category there are different types of clustering such as minimum spanning tree, k-means clustering[7], nearest neighbor algorithm.

2.2.3 Clustering with categorical attributes

Traditional algorithms do not always work with categorical data. The higher order co-occurrences with χ-sim Algorithm [2], which has been used for text categorization. The clustering algorithms for data with boolean and categorical data as text is discussed in ROCK algorithm[6]. Here the images, which are taken as the input is, treated as categorical attributes and clustering is made.

3. PROPOSED APPROACH

Medical images are compared by extracting some of the features in them. Each of the sample images is compared with the images in the database. Based on the results of the comparison, a boolean matrix is formed. Each element in the boolean matrix is checked for finding the links. All the links together forms a link matrix. Then, the links are grouped to form clusters. The higher order similarity between categorical inputs and similarity between pool of images and vice versa. The proposed technique contains three phases.

3.1 Image Comparison

The images, which are taken as categorical inputs, are compared with the pool of images to be classified, using RGB feature. Each of the categorical input is compared with the pool of images. The result of the comparison is a r x c array with the elements as 1 for each j^{th} categorical input contained in the i^{th} pool of image, where j=1 to c and i=1 to r. The remaining elements are 0s.

3.2 Identifying The Neighbors

Based on the number of links between images, similarity between images is identified. Images are said to be neighbors, if their similarity is above some threshold. To check whether the similarity exceeds the threshold or not, the similarity is to be measured. So, Jaccard coefficient is used to measure the similarity.

$$sim(ti,tj) = \frac{|Ti \cap Tj|}{|Ti \cup Tj|}$$

An adjacency matrix is formed and converted into a boolean matrix. In the boolean matrix, if the two corresponding points are neighbors, the entry is 1. The next step is, a link matrix is formed. Then, clusters are created by comparing each row in the link matrix with the other. If there are similar rows in the link matrix, they are placed in the same cluster.

3.3 Applying Higher Order Co-Occurrence

The proposed algorithm takes as input the boolean matrix along with the categorical information of the pool of images. The next step finds the higher order similarity between images using χ-sim Algorithm[2]

4. IMPLEMENTATION DETAILS

4.1 Image Comparison

The images, which are taken as categorical inputs, are compared with the pool of images to be classified, using RGB feature. Each of the categorical input is compared with the pool of images. The result of the comparison is a r x c array with the elements as 1 for each j^{th} categorical input contained in the i^{th} pool of image, where j=1 to c and i=1 to r. The remaining elements are 0s.

4.1.1 Pseudocode for comparison

1. Get the c categorical inputs
2. FOR EACH categorical input j
   FOR EACH image i in pool do the following
3. FOR EACH ri row
   IF row rj of j image is subset with row ri of i THEN
   rj = rj +1
   END IF
4. ENDFOR
5. IF rj = total row of j image then
   v(i,j)=1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

```
        ELSE
        v(i,j)=0
        END IF
    6.  ENDFOR
    7.  ENDFOR
```

4.2 Identifying The Neighbors

Based on the number of links between images, similarity between images is identified. Images are said to be neighbors, if their similarity is above some threshold. To check whether the similarity exceeds the threshold or not, the similarity is to be measured. So, Jaccard coefficient is used to measure the similarity.

$$sim(mi,mj) = \frac{|Mi \cap Mj|}{|Mi \cup Mj|}$$

An adjacency matrix is formed and converted into a boolean matrix. In the boolean matrix, if the two corresponding points are neighbors, the entry is 1. The next step is, a link matrix is formed. Then, clusters are created by comparing each row in the link matrix with the other. If there are similar rows in the link matrix, they are placed in the same cluster.

4.2.1 Pseudocode for identifying the neighbor

```
1.  FOR EACH ROW i in  M
2.  FOR EACH ROW j in  M
3.   SET   sim(i,j)=s_in(i,j)/s_un(i,j);
4.  IF sim(i,j) >= theta
        neighbor(i,j)=1
    ELSE
      neighbor(i,j)=0
    END IF
5.  ENDF OR
6.  ENDFOR
7.  SET link = neighbor x  neighbor'
```

4.2.2 Pseudocode for forming cluster

```
1.  FOR EACH ROW  i in link
2.   k1=1;
3.  FOR  ROW  j in link

4.      IF row i= row  j THEN
5.          set clus(i,k1)=j
6.          k1=k1+1
    END IF
7.  ENDFOR
8.  ENDFOR
9.  DELETE duplicate row
```

4.3 Applying Higher Order Co-occurrence

This part of algorithm takes input as the boolean matrix along with the categorical information of the pool of images. The next step finds the higher order similarity between images using χ-sim Algorithm[2]

1. Initialize $SR^{(0)}$ matrix and $SC^{(0)}$ matrix as identity matrix.
2. FOR t = 1 to n times do steps 3 and 4
3. Calculate new similarity matrix at each iteration

$$\mathbf{SR}^{(t)} = (\mathbf{M} \bullet \mathbf{SC}^{(t-1)} \bullet \mathbf{M}^{\mathrm{T}}) \otimes \mathbf{NR},\ nr_{ij} = \frac{1}{\mu_i \mu_j}$$

$$\mathbf{SC}^{(t)} = (\mathbf{M}^{\mathrm{T}} \bullet \mathbf{SR}^{(t-1)} \bullet \mathbf{M}) \otimes \mathbf{NC},\ nc_{ij} = \frac{1}{\mu^i \mu^j}$$

4. Set diagonal elements of $SR^{(t)}$ and $SC^{(t)}$ to 1

4.4 Advantages And Limitations Of The Proposed Approach

The advantages of the proposed approach are :

1. Overcome the problems that arise in image meta search
2. Outliers can be effectively processed. When we choose a value for *theta*, the outliers are separated from the rest of the images.
3. Adding additional columns in similarity matrices, to identify images of different category.

The limitations of the proposed approach is, it is relevant in situations where a domain expert/similarity table is the only source of knowledge.

5. EXPERIMENTAL RESULTS

The experiments were done with huge data. The following are output with seven images in pool and 6 categorical input images as a sample.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

```
m1 =

     0     1     0     0     0     0     0
     0     1     1     1     1     1     0
     1     1     0     0     0     1     1
     0     1     0     0     0     0     0
     0     1     0     1     0     0     0
     0     1     1     1     0     0     0
     1     0     0     0     0     0     1


sim =

    1.0000    0.2000    0.2500    1.0000    0.5000    0.3333         0
    0.2000    1.0000    0.2857    0.2000    0.4000    0.6000         0
    0.2500    0.2857    1.0000    0.2500    0.2000    0.1667    0.5000
    1.0000    0.2000    0.2500    1.0000    0.5000    0.3333         0
    0.5000    0.4000    0.2000    0.5000    1.0000    0.6667         0
    0.3333    0.6000    0.1667    0.3333    0.6667    1.0000         0
         0         0    0.5000         0         0         0    1.0000


link =

     3     1     0     3     3     1     0
     1     3     0     1     3     3     0
     0     0     2     0     0     0     2
     3     1     0     3     3     1     0
     3     3     0     3     5     3     0
     1     3     0     1     3     3     0
     0     0     2     0     0     0     2
```

Figure 5.1 m1 gives comparison of pool of image with categorical input. Link gives neighbors between images in pool.

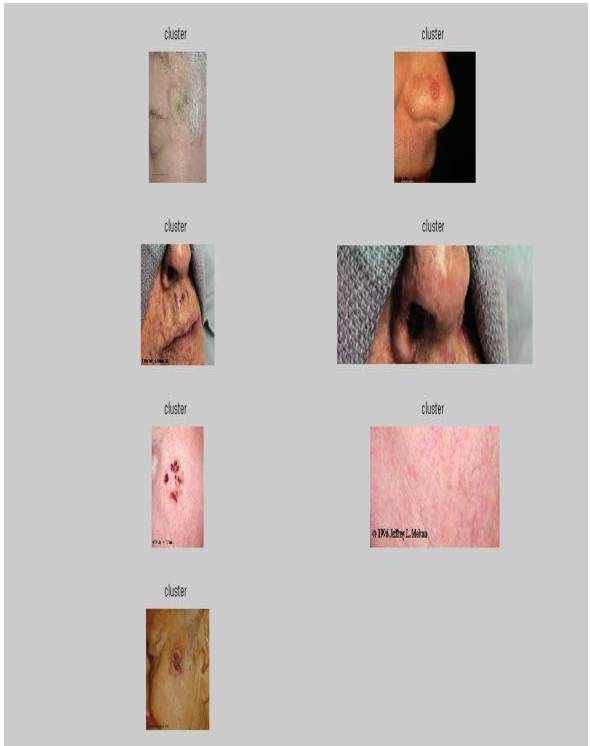

Figure 5.2 this figure shows how pool of image is clustered together based on the categorical inputs.

```
clus1 =

     1     4
     2     6
     3     7
     5     0


m =

     0     1     0     0     0     0     0     1     0     0     0
     0     1     0     0     0     0     0     1     0     0     0
     0     1     1     1     1     1     0     0     1     0     0
     0     1     1     1     0     0     0     0     1     0     0
     1     1     0     0     0     1     1     0     0     1     0
     1     0     0     0     0     0     1     0     0     1     0
     0     1     0     1     0     0     0     0     0     0     1

Similarity between images and categorical inputs

sr =

    1.0000    0.5000    0.0833         0         0         0         0
    0.5000    1.0000         0         0         0         0         0
    0.0833         0    1.0000         0         0         0         0
         0         0         0    1.0000    0.0500         0    0.1667
         0         0         0    0.0500    1.0000    0.2000    0.0667
         0         0         0         0    0.2000    1.0000         0
         0         0         0    0.1667    0.0667         0    1.0000
```

Figure 5.3 Clus1 gives the cluster information. M gives the Boolean matrix with cluster information. sr gives the similarity between images and categorical inputs.

Thus it is possible in find the similarities between the images in the pool and categorical images.

6. CONCLUSION

The proposed approach differs from the existing approaches by taking categorical inputs, as image segments, which extracts features and identifies the similarity between images. This approach improves the efficiency of image clustering based on color sensation. The experiment results depicts that it would work fast, and the clustering and categorization of images is done effectively. The proposed approach also overcomes the drawbacks of content based indexing.

REFRENCES:

[1] Aditya Vailaya, A. T. Figueiredo, Anil K. Jain and Hong-Jiang Zhang, "Image Classification for Content-Based Indexing", *IEEE Transactions On Image Processing*, Vol. 10, No. 1, 2001,PP 117-130

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[2] G. Bisson and F. Hussain, "Text Categorization Using Word Similarities Based on Higher Order Co-occurrences", *Proceedings of tenth SIAM international Conference on Data mining 2010*, pp. 1-12.

[3] G. Bisson and F. Hussain, "Chi-Sim: A New Similarity Measure for the Co-clustering Task", *Proceedings of the 2008 Seventh International Conference on Machine Learning and Applications* #Volume 00, 2008, pp. 211–217.

[4] M. Inaba, H. Imai, and N. Katoh, ªExperimental Results of a Randomized Clustering Algorithm,º Proc. 12th Ann. ACM Symp. Computational Geometry, pp. C1-C2, May 1996.

[5] M.N. Murty, A.K. Jain and P.J. Flynn, ªData Clustering: A Review,º ACM Computing Surveys, vol. 31, no. 3, pp. 264-323, 1999.

[6] Sudipto Guha, Rajeev Rastogi, Kyuseok Shim, "ROCK: A Robust Clustering Algorithm for Categorical Attributes," icde, pp.512, 15th *International Conference on Data Engineering (ICDE'99)*, 1999.

[7] Tapas Kanungo, David M. Mount, Nathan S. Netanyahu, Christine D. Piatko, Ruth Silverman, and Angela Y. Wu, "An Efficient k-Means Clustering Algorithm: Analysis and Implementation", *IEEE Transactions On Pattern Analysis And Machine Intelligence*, Vol. 24, No. 7, July 2002, pp 881-892

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

AUTHOR PROFILES:

K. Anuradha received the M.Tech degree in Computer Science and Engineering from SASTRA University, Thanjavur, in 2009. She is a research student of Dr. N.Sairam. Currently, she is an Assistant Professor at SASTRA University, Thanjavur . Her interests are in Data mining and Genetic algorithm

Dr. N. Sairam received the M.Tech degree in Computer Science and Engineering from SASTRA University, Thanjavur. He received the Ph.D. degree in Computer Science from SASTRA University, Thanjavur. Currently, he is a professor at SASTRA University, Thanjavur. His research interests include distributed algorithms,data mining and genetic algorithms.