AIVP ADVANCES IN IMAGE AND VIDEO PROCESSING
SOCIETY FOR SCIENCE AND EDUCATION UNITED KINGDOM
VOLUME 4, ISSUE 2
ISSN 2054-7412

Discovering Optimized Association Rules based on Image Content

Jyoti Jayprakash Deshmukh and Udhav Bhosle
Department of Electronics and Telecommunication Engineering, Rajiv Gandhi Institute of Technology, Mumbai, University of Mumbai, India;
jyotideshmukh11@gmail.com; udhavbhosle@gmail.com

ABSTRACT

Authors present the concept of image mining, an extension of data mining for discovering semantically meaningful information and image data relationship from a large collection of images. Association rule mining is the process of discovering useful and interesting rules, representing frequent patterns from large datasets, depends on user specified minimum support and confidence values. These constraints lead to exponential search space and dataset dependent minimum support and confidence values. The authors propose an optimization technique for overcoming these problems using multi-fitness function Genetic algorithm and constrained nonlinear minimization and minimax optimization method. Synthetic image set containing geometric shapes and standard MIAS medical image dataset are used to validate the proposed optimization algorithm.

Experimental results show that, Genetic algorithm generates more efficient, effective and strong association rules than constrained nonlinear minimization and minimax optimization method. Genetic algorithm achieves 50% and 90%, constrained nonlinear minimization and minimax optimization method achieves 22% and 74 %, reduction in association rules for synthetic image set and standard MIAS medical image dataset respectively.

Keywords: Image Mining; Association rule mining; Correlation measures; Apriori algorithm; mammogram; Genetic algorithm.

1 Introduction

Advanced image acquisition and storage technologies enabled the development of extensive image database. Every day, a large number of images are generated. Digital libraries are enriched with the addition of various images. Proper analysis of these images reveals useful information to the users. Humans find it difficult to discover underlying knowledge and patterns in large collection of images due to unavailability of effective tools. Image mining system extracts semantically meaningful information (knowledge) from image data automatically. Association rule mining is the process of finding useful and interesting rules from large datasets. These association rules depend on user specified minimum value of support and confidence. This leads to exponential search space and dataset dependent minimum support and confidence value.

Ji Zhang et al. [1] discussed image mining frameworks, current developments, issues in image mining, state-of-the-art techniques. C. Ordonez et al. [2] introduce data mining for knowledge discovery in image database. Authors concentrate on the problem of finding associations rules in 2-dimensional color images. Carson et al. [3] presents image representation to provide a transformation to a small

DOI: 10.14738/aivp.42.1994
Publication Date: 23rd April, 2016
URL: http://dx.doi.org/10.14738/aivp.42.1994

set of localized coherent regions from a raw pixel data in color and texture space. A framework to achieve higher retrieval efficiency using texture information is reported in [4]. Jawad Nagi et al. [6] proposed a method using morphological processing and seeded region growing algorithm for automated breast profile segmentation for region of interest (ROI) detection. The spatial arrangement of pixel intensities characterizes texture information. Texture is used as a visual feature to retrieve similar patterns from image database [7], [8]. Beyer et al. [9] proposed a technique to reduce number of features for increasing significance of each feature and thus to improve the discrimination accuracy. Agrawal et al. [11] first time discussed the problem of Association rule mining. From any market-basket type database, extraction of some useful and interesting rules is performed by a Pareto based genetic algorithm [16]. Manish Saggar et al. [17] optimize the association rules using Genetic algorithm and predict the rules which contain negative attributes. Wakabi-Waiswa et al. [18] proposed a multi-objective approach for generating optimal association rules using syntactic superiority and transactional superiority quality metrics.

The Authors proposed a method to find frequent significant patterns in a given collection of images using association rule mining. In this paper, association rules are found using Apriori algorithm. Among these rules, many are redundant and uncorrelated which give misleading information and hence exponential search space. So, the authors proposed an optimization method to get non-redundant, highly correlated and strong association rules. It includes optimization of association rules using Genetic algorithm and constrained nonlinear minimization and minimax optimization method.

The Proposed method is validated on synthetic image set containing geometric shapes and standard MIAS medical image dataset. It follows four steps. First step is pre-processing, image segmentation and extraction of objects. In second step, features are extracted from segmented object and feature vector is generated. Third step is formation of transaction database which is given as input to Apriori algorithm for generating association rules. In fourth step, generated association rules are optimized using two different optimization algorithms. First includes optimization of association rules using Genetic algorithm by refining them using multi-fitness function for the interesting measures such as Cosine, All-Confidence, Accuracy and Jaccard to get strong association rules. In second algorithm, the association rules are optimized using constrained nonlinear minimization and minimax optimization method.

The rest of this paper is organised as follows. Section two presents the proposed method. In section three, experimental results are discussed and section four comprises conclusion and scope for future work.

2 Proposed Method

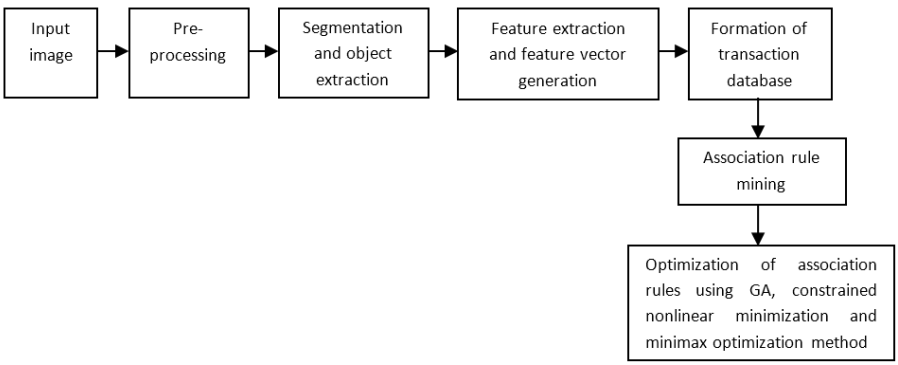


Figure 1: Block diagram of the proposed method

Figure 1 shows the block diagram of the proposed method. Here the novelty is in optimization of association rules using Genetic algorithm and constrained nonlinear minimization and minimax optimization method. It gives non-redundant, highly correlated and strong association rules, and hence reduces search space and consequently improves the processing time of its future applications.

2.1 Image Pre-processing and Segmentation

In pre-processing step for synthetic images, median filtering is used to remove external noise introduced in images. The objects from each image are extracted using segmentation algorithm. The image is segmented into K regions using the spatial grouping of pixels. Using summary information, blob for each connected region having area greater than 2% of the image area is generated. These blobs are called objects [2].

For standard MIAS medical image dataset, median filtering is used to remove digitization noise introduced during image acquisition process. Figure 5(a) shows original mammogram image, figure 5(b) shows median filtered image. Radiopaque artifacts such as labels and wedges in mammogram images are removed using thresholding and morphological operations. Through experimentation, a global threshold with a value of T=100 is selected for transforming the grey-scale mammogram image into binary [0,1] format. Figure 5(c) shows thresholded image. Morphological operations such as dilation, erosion, opening, and closing are carried out on the binary images for suppression of artifacts, labels and wedges. Contrast enhancement is performed on the processed mammogram image. Figure 5(d) shows the resultant image obtained after applying contrast enhancement technique. Pectoral muscles are segmented by using the region growing technique. For region growing a seed is placed inside the pectoral muscle of input mammogram. Segmented image with region of interest (ROI) is shown in figure 5(e) [5], [6].

2.2 Feature Extraction and Feature Vector Generation

For synthetic images, features are extracted from the segmented objects and organized into feature vectors. Ten-dimensional feature vectors are produced which contain summary information about color, texture and area. Statistical texture features as average grey level, uniformity, average contrast, smoothness, third moment and entropy are used. For standard MIAS medical image dataset, Grey Level Co-occurrence Matrix (GLCM) is used to extract the features from the segmented images and are organized into feature vectors. A Co-occurrence matrix M (d, θ) is given by the relative frequency of occurrences of two grey level pixels i and j separated by d pixels in the θ orientation. Co-occurrence matrices are calculated for 0^0, 45^0, 90^0, and 135^0 directions, and for the distances 1, 2, 3, 4 and 5. In this process, 20 matrices of 16 by 16 integer elements per image are produced. For each matrix, seven features as presented in Table 1 are calculated. Feature vector of size 140 elements is produced to represent each image [7]. Also Table 1 gives grey level texture features positions in feature vector. Figure 2 describes the directionality used in GLCM.

2.3 Formation of Transaction Database and Association Rule Mining

For synthetic image set, objects in the input images are identified by segmentation algorithm and labeled using the image query processing algorithm. Similarity function is used to compare the current object with earlier identified and labeled object. Similarity measure is 1 if there is exact match for all desired features and it is 0, if the match becomes worse. If the match is found, same ID is assigned to current object else new ID will be created and assigned. The result of the above step is

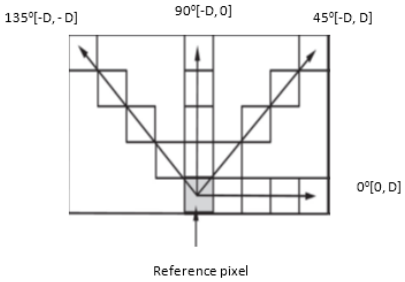


Figure 2: Directions used in GLCM matrix

Table 1. Texture features and their positions in feature vector [10]

Feature	Equation	Meaning	Position
Step	$\sum i \sum j\ P(i,j)$	Distribution	1-20
Variance	$\sum i \sum j\ P(i-j)^2 P(i,j)$	Contrast	21-40
Entropy	$\sum i \sum j\ P(i,j)\log(P(i,j))$	Suavity	41-60
Energy	$\sum i \sum j\ P(i-j)^2$	Uniformity	61-80
Homogeneity	$\sum_i \sum_j\ P(i-j)/(1+\lvert i-j \rvert)$	Homogeneity	81-100
3° Moment	$\sum i \sum j\ (i-j)^3 P(i,j)$	Distortion	101-120
Inv. Variance	$\sum_i \sum_j\ P(i,j)/(i-j)^2$	Inv. Contrast	120-140

a set of transactions, one for each image, which contains the object identifiers for the objects, contained in the image. Thus above created image transaction database has one entry per image containing object ID's. This is given as input to association rule mining algorithm to generate association rules. These object ID's appear in the association rule, showing correlation between these objects, present in the input images.

For standard MIAS medical image dataset, keywords of input mammogram images i.e. benign or malignant and feature vectors are used to build the transaction database. Transaction database has a transaction record for every input image and it is submitted to Apriori algorithm.

2.4 Association Rule Mining

Let R be the set of objects, called as items, and T be a set of data cases. *Association rule* is represented as A→B, where A, B⊂ R and A ∧ B=ϕ. A and B are called, the antecedent and the consequent of the rule respectively. Item set is a set of items, containing the antecedent and the consequent. Support and confidence gives the strength of an association rule. Support value gives how frequently a rule is applicable for a given data set. Value of confidence determines how frequently items in B appear in transactions that contain A [11]. These metrics are defined as,

$$\text{Support, } S\ (A \rightarrow B) = P(A \cup B) \quad (1)$$

$$\text{Confidence, } C\ (A \rightarrow B) = P(B|A) \quad (2)$$

The rule A→ B has support s in T, if s % of the data case in T contains both A and B. The rule valid for T with confidence c if c % of the data case in T that support A also support B. Association rule mining is to find all rules that have support and confidence value greater than some minimum threshold value for support and confidence, specified by the user.

A Brute-force approach determines the support and confidence of very possible rule for association rule mining. It is an expensive approach, as there exist many rules that can be extracted from a given data set. Let d be the items present in a data set, total number of possible rules extracted are,

$$R = 3^d - 2^{d+1} + 1 \quad (3)$$

Most of the rules are discarded after applying minimum support 20% and confidence 50%, making most of the computations waste. So, it is good to prune the rules early without finding support and confidence values. To improve performance of association rule mining algorithms, the support and confidence requirements should be decoupled [12]. To overcome this, Apriori algorithm [11] divides the problem into two tasks as: frequent item set generation and from the generated frequent item set, extraction of all the high-confidence rules. So, in proposed system, Apriori algorithm is used for mining frequent item sets for generating strong association rules.

2.5 Optimization of Association Rules using Genetic Algorithm

Association rule mining by using Apriori algorithm gives all rules in data which satisfy the minimum support and confidence threshold value. Information interpreted varies according to rule. Many times, rule having high value of support and confidence give conflicting or redundant information, which makes it uninteresting rule. Thus the confidence and support measures are insufficient for filtering out uninteresting association rules. To overcome this limitation, interestingness correlation measures are used for augmenting the support and confidence framework for association rules [14]. This leads to correlation rules of the form,

$$A \rightarrow B \text{ [support, confidence, correlation measures]} \quad (4)$$

Multi-fitness function Genetic algorithm is used, to optimize association rules, generated in image mining process using Apriori algorithm. It includes interestingness correlation measures as fitness function rather than support and confidence to obtain strong association rules.

2.5.1 Outline of Basic Genetic Algorithm

Random population comprising n chromosomes is generated and the fitness function f(x) for each chromosome x in the population is evaluated. New population is created by repeatedly executing following steps until creation of the new population is finished [13].

In selection step, two parent chromosomes are selected according to their fitness, from a population. Further, crossover is performed between parents with a crossover probability, for creation of new offspring. If there is no crossover, produced offspring is an exact copy of parents. With a mutation probability, new offspring are mutated at each locus (position in chromosome) and placed in a new population. New generated population is used for a further processing of algorithm. When the end condition is satisfied, it stops there, and returns the best solution in current population. Then follow crossover step.

2.5.1 Operators in Proposed Genetic Algorithm

In proposed system, for selecting individuals with respect to fitness function, roulette wheel selection method is used. Single-point crossover method is used. In this crossover method one crossover point is selected, binary string is copied from beginning of chromosome to the crossover point of one parent and rest is copied from the second parent. Mutation gives a chance for flipping a gene within a chromosome. Crossover and mutation probabilities are empirically selected for the proposed algorithm, as 0.85 and 0.006 respectively.

Fitness Function: Multi-fitness function used in the proposed algorithm comprises interestingness correlation measures as All-Confidence, Cosine, Accuracy and Jaccard for refining the association rules. Many correlation measures are listed in the literature. However, we consider All Confidence,

Cosine, Accuracy and Jaccard for Multi-Fitness function. Interestingness correlation measures between two item sets, A and B, are defined as

$$\text{All-Confidence (A, B)} = \min [P(\frac{A}{B}), P(\frac{B}{A})] \quad (5)$$

$$\text{Cosine (A, B)} = \frac{P(A \cup B)}{\sqrt{P(A).P(B)}} \quad (6)$$

$$\text{Accuracy (A, B)} = P(A, B) + P(\overline{A}, \overline{B}) \quad (7)$$

$$\text{Jaccard (A, B)} = \frac{P(A \cup B)}{P(A)+P(B)-P(A \cup B)} \quad (8)$$

Algorithm 1 gives optimized association rules generated using Genetic algorithm with interestingness correlation measures as fitness function.

Algorithm 1: Association rule optimization using Genetic algorithm

Input: n: total number of rules, S = 0, minimum Support threshold, minimum Confidence threshold and avg_opt_threshold_GA = average optimized threshold value using Multi-fitness function Genetic algorithm for Cosine, All-Confidence, Accuracy and Jaccard correlation measures.

Output: Strong association Rules

1. For every rule r_i, and $i \leq n$
2. Compute supp (r_i), Conf (r_i), Cosine(r_i), All-Confidence(r_i), Accuracy(r_i), Jaccard(r_i)
3. If Supp(r_i) > minimum Support threshold and Conf (r_i) > minimum Confidence threshold
    a. If Cosine(r_i), All-Confidence(r_i), Accuracy(r_i) and Jaccard(r_i) > avg_opt_threshold_GA
        i. Print 'Strong Association Rule'
        ii. S + +
    b. End
4. End
5. End For

2.6 Optimization of Association Rules using Constrained Nonlinear Minimization and Minimax Optimization Algorithm

Correlation measures such as cosine, all-confidence, accuracy and jaccard are nonlinear functions. So optimization of individual correlation measure is done by using constrained nonlinear minimization and minimax optimization method. Constrained nonlinear minimization is used to find a minimum of a constrained nonlinear multivariable function. Minimax optimization method is used to find a minimum of the worst-case value of a set of multivariable functions. Algorithm 2 gives optimized association rules generated using constrained nonlinear minimization and minimax optimization algorithm.

Algorithm 2: Association rule optimization using constrained nonlinear minimization and minimax optimization algorithm

Input: n: total number of rules, S = 0, minimum Support threshold, minimum Confidence threshold and avg_opt_threshold_CNMMOA = average optimized threshold value using constrained nonlinear minimization and minimax optimization algorithm for Cosine, All-Confidence, Accuracy and Jaccard correlation measures.

Output: Strong association Rules

1. For every rule r_i, and $i \leq n$
2. Compute supp (r_i), Conf (r_i), Cosine(r_i), All-Confidence(r_i), Accuracy(r_i), Jaccard(r_i)
3. If Supp(r_i) > minimum Support threshold and Conf (r_i) > minimum Confidence threshold
   a. If Cosine(r_i), All-Confidence(r_i), Accuracy(r_i) and Jaccard(r_i) > avg_opt_threshold_CNMMOA
      i. Print 'Strong Association Rule'
      ii. S + +
   b. End
4. End
5. End For

3 Experimental Results and Discussions

To validate the proposed method, experiment is performed on two different data sets, carried on MATLAB environment.

3.1 Experiment I- The Synthetic Image Dataset

In this experiment, set of co-related synthetic images are selected. The objects from each image are extracted using segmentation algorithm and scale color selection is estimated. Ten-dimensional feature vectors are produced and these vectors contain summary information about color, texture and area. Statistical texture features as average grey level, uniformity, average contrast, smoothness, third moment and entropy are used. Using Expectation Maximization (EM) method, several clustering of feature vectors are produced and decided the best using Minimum description length principle. The image is segmented into K regions using the spatial grouping of pixels. Using summary information, blob for each connected region having area greater than 2% of the image area is generated. These blobs are called objects. Objects in the images are identified and labeled using the image query processing algorithm. Similarity function is used to compare the current object with earlier identified and labeled object. Similarity measure is 1 if there is exact match for all desired features and it is 0, if the match becomes worse. If the match is found, same ID is assigned to current object else new ID will be created and assigned to it. The result of the above step is a set of transactions, one for each image, which contains the object identifiers for the objects, contained in the image. This image transaction database is used as input to association rule mining algorithm to generate association rules.

10 representative images are considered for this experiment [2]. Figure 3 shows the original image on the left with different geometric shapes with white background. The images are the input data for our program.

By using image mining algorithm as discussed in section two, the various objects (blobs) present in these images are segmented. These blobs are labeled using similarity measure based on color and texture features. The same label is given to similar blobs based on their color and texture features. Color standard deviation is set to 0.29 and contrast standard deviation to 0.5 for object identification and matching. These parameters are tuned after several experiments. The results obtained are treated as transactions for association rule mining.

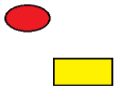
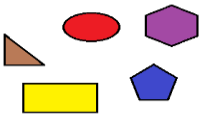
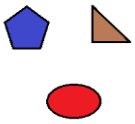
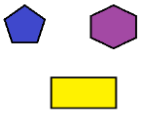
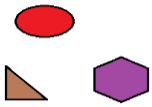
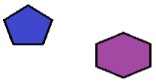
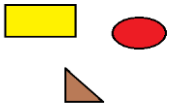
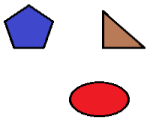
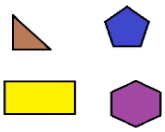
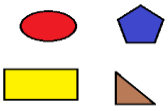
Image	Objects Identified	Image	Objects Identified
	-> 1,2		-> 1,2,3,6,7
	-> 1,3,6		-> 2,6,7
	-> 1,3,7		-> 6,7
	-> 1,2,3		-> 1,3,6
	-> 2,3,6,7		-> 1,2,3,6

Figure 3: Images and objects identified [2]

Rules Generated:

Parameters are:

Support: 20%

Confidence: 70%

Number of records: 10

Number of Association Rules Generated: 18

1. {3} → {1} S=60% C=86%
2. {1}→ {3} S=60% C=86%
3. {6} → {3} S=50% C=71%
4. {3} → {6} S=50% C=71%
5. {1,6} → {3} S=30% C=100%
6. {3,6} → {1} S=40% C=80%
7. {7} → {6} S=40% C=80%
8. {2,7} → {6} S=30% C=100%
9. {2,3} → {1} S=30% C=75%
10. {1,2} → {3} S=30% C=75%
11. {2,6} → {3} S=30% C=75%
12. {2,3} → {6} S=30% C=75%
13. {6,7} → {2} S=30% C=75%
14. {2,6} → {7} S=30% C=75%
15. {1,2,6} →{3}S=20% C=100%
16. {3,6,7} →{2}S=20% C=100%

17. {2,3,7} →{6}S=20% C=100%
18. {1,7} → {3} S=20% C=100%

Let's analyze some of the above rules. The first rule {3} → {1} state that if there is triangle in the image then there is also a circle. This rule has support 60% and confidence 86% which are greater than minimum threshold of support and confidence. The rule {1, 6} → {3} means that if circle and pentagon is present in the image, then triangle is also present with support 30% and confidence 100%. The rule {1, 2, 6 } → {3} means that if circle, square and pentagon is present in the image, then triangle is also present with support 20% and confidence 100%.

3.1.1 Optimization of Association Rules using Genetic Algorithm

Correlation measures Cosine, All-Confidence, Accuracy and Jaccard are calculated for these rules, as listed in Table 2. Multi-fitness function Genetic algorithm is used to achieve strong association rules. Interestingness correlation measures such as Cosine, All-Confidence, Accuracy and Jaccard are selected empirically, as fitness function, as they show feasible results to refine the rules generated in image mining.

Table 2. Support, confidence and correlation measure

Rule No	Support	Confidence	Correlation Measure			
			Cosine	All-Confidence	Accuracy	Jaccard
1	60	86	0.8571	0.8571	0.8	0.75
2	60	86	0.8571	0.8571	0.8	0.75
3	50	71	0.7142	0.7142	0.6	0.555
4	50	71	0.7142	0.7142	0.6	0.555
5	40	100	0.7559	0.5714	0.4	0.5714
6	40	80	0.6761	0.5714	0.4	0.5
7	40	80	0.6761	0.5714	0.6	0.5
8	30	100	0.654	0.4285	0.3	0.4285
9	30	75	0.5669	0.4285	0.4	0.375
10	30	75	0.5669	0.4285	0.4	0.375
11	30	75	0.5669	0.4285	0.3	0.375
12	30	75	0.566	0.4285	0.3	0.375
13	30	75	0.6123	0.5	0.3	0.4285
14	30	75	0.67	0.6	0.3	0.5
15	20	100	0.5345	0.2857	0.2	0.2857
16	20	100	0.5773	0.333	0.2	0.333
17	20	100	0.5345	0.2857	0.2	0.2857
18	20	100	0.5345	0.2857	0.2	0.2857

It is observed that number of strong association rules generated depend on support, confidence and optimized value of interestingness correlation measures such as Cosine, All-Confidence, Accuracy and Jaccard. These correlation measures are linearly independent of each other. So optimization of individual correlation measure is carried out by using Genetic algorithm. By using Genetic algorithm, optimized values for Cosine, All-Confidence, Accuracy and Jaccard are listed in Table3. By using Multi-Fitness function Genetic algorithm, average optimized threshold value for correlation measures is 0.4. Rules satisfying average optimized threshold criteria are considered to be strong rules. Table 3 lists number of optimized association rules obtained by using interestingness correlation measures as fitness function of Genetic algorithm.

Table 3. Correlation measures and number of optimized rules by GA

Interestingness correlation measures as fitness function parameter of GA	Optimized value	No of optimized rules Using GA
Cosine	0.5741	11
all-confidence	0.4986	9
Accuracy	0.21	14
Jaccard	0.3619	14

Optimized association rules by using multi-fitness function Genetic algorithm are:

1. {3} → {1} S=60% C=86%
2. {1} → {3} S=60% C=86%
3. {6} → {3} S=50% C=71%
4. {3} → {6} S=50% C=71%
5. {1,6} → {3} S=30% C=100%
6. {3,6} → {1} S=40% C=80%
7. {7} → {6} S=40% C=80%
8. {2,3} → {1} S=30% C=75%
9. {1,2} → {3} S=30% C=75%

3.1.2 Optimization of Association Rules using Constrained Nonlinear Minimization and Minimax Optimization Method

Correlation measures such as cosine, all-confidence, accuracy and jaccard are nonlinear functions. So optimization of individual correlation measure is done by using constrained nonlinear minimization and minimax optimization method. Optimized value for cosine, all-Confidence, accuracy and jaccard are listed in Table 4. For constrained nonlinear minimization and minimax optimization method, average optimized threshold value obtained for correlation measures is 0.225. Rules satisfying average optimized threshold criteria are considered to be strong rules. Table 4 lists number of strong association rules obtained by optimization of interestingness correlation measures using constrained nonlinear minimization and minimax optimization method.

Table 4. Correlation measures and number of optimized rules by constrained nonlinear minimization and minimax optimization method

Interestingness correlation measures	Optimized value	No of optimized rules
Cosine	0.1	18
all-Confidence	0.50	9
Accuracy	0.20	18
Jaccard	0.1	18

Optimized association rules by using constrained nonlinear minimization and minimax optimization method are:

1. {3}→{1} S=60% C=86%
2. {1}→{3} S=60% C=86%
3. {6}→{3} S=50% C=71%
4. {3}→{6} S=50% C=71%
5. {1,6}→{3} S=30% C=100%
6. {3,6}→{1} S=40% C=80%
7. {7}→{6} S=40% C=80%
8. {2,7}→{6} S=30% C=100%
9. {2,3}→{1} S=30% C=75%

10. {1,2}→{3} S=30% C=75%
11. {2,6}→{3} S=30% C=75%
12. {2,3}→{6} S=30% C=75%
13. {6,7}→{2} S=30% C=75%
14. {2,6}→{7} S=30% C=75%

Figure 4 explains comparison of optimized association rules using Genetic algorithm, constrained nonlinear minimization and minimax optimization method. It shows that Genetic algorithm achieves better optimization of association rules than constrained nonlinear minimization and minimax optimization method. Thus, Genetic generates more efficient, effective and strong association rules than constrained nonlinear minimization and minimax optimization method.

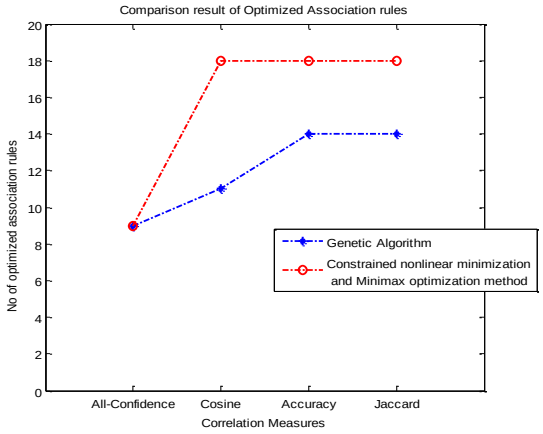


Figure 4: Comparison of optimized association rules using Genetic algorithm, constrained nonlinear minimization and minimax optimization method

3.2 Experiment II- Mammogram Image Dataset: MIAS

Mammography Image Analysis Society (MIAS) database is used to test proposed optimization method. The MIAS database is built by Suckling et al. [15], and is openly available for scientific research. MIAS dataset provide appropriate information based on types of background tissues, and the class of abnormalities present in the mammograms. The class of abnormality consists of normal-abnormal class, and again based upon the severity of abnormality; the abnormal class is divided into two sub-classes such as benign and malignant. The MIAS database contains 322 images, which are categorized into three according to tissue types like fatty, fatty-glandular and dense-glandular. Out of 322 images, 208 images are normal, 114 images are abnormal, and again among abnormal images the numbers of benign and malignant types are 63 and 51 respectively. All the abnormal images are considered for our experiment from this database. Results of segmentation step are shown in figure 5(a-e).

(a) (b) (c) (d) (e)

Figure 5: Mammogram segmentation process results; (a) original mammogram; (b) filtered image after noise removal; (c) Thresholded image; (d) image after contrast enhancement; (e) final segmented image

Textual feature are extracted from the segmented mammogram images (ROI) by using Grey Level Co-Occurrence Matrix (GLCM) method and these features are organized into feature vectors. Table 5 gives feature vector generated for MIAS database. Feature vector and the keyword of the input mammogram images i.e. benign or malignant are submitted to transaction database. Table 6 shows the transaction database for MIAS database, where first column presents class of image i.e. benign or malignant, is represented by 1001 and 1002 respectively. Second column onwards present feature value interval label. Unique label for each interval are assigned, which increases sequentially for next interval. The transaction representations of all the images in the training set are submitted to Apriori algorithm to generate association rules. At output, 14232 association rules are obtained by using Apriori algorithm for MIAS database. Examples of association rules mined are:

5,129 ->1001 i.e. Benign Image (Support=13% and Confidence=100%)

2, 25, 47 ->1002 i.e. Malignant image (Support=9% and Confidence=100%)

First rule explains that the image having the feature value interval label as 5 and 129 tend to be a benign image. Second rule explains that the image having the feature value interval label as 2, 25 and 47 tend to be a malignant image.

Next, multi-fitness function Genetic algorithm and constrained nonlinear minimization and minimax optimization algorithm are applied on all association rules generated by Apriori algorithm in earlier step to determine strong, effective and highly correlated association rules. For every rule support, confidence, and correlation measures as Cosine, All-Confidence, Accuracy and Jaccard are calculated. Average optimized threshold value for correlation measures Cosine, All-Confidence, Accuracy and Jaccard are 0.4 and 0.225 by using genetic algorithm and constrained nonlinear minimization and minimax optimization algorithm respectively. Rules satisfying average optimized threshold criteria are considered to be strong rules. Table 7 shows number of association rules obtained from Brute-force approach, Apriori algorithm and by both optimization algorithms.

Table 5. Feature vector generation

Image No. (1-114)	Features (1-140)											
	1	2	3	4	...	134	135	136	137	138	139	140
1	65792	65535	65278	65021	...	2.9111	3.07555	3.05777	3.08888	3.13777	3.16888	3.13777
2	65792	65535	65278	65021	...	0.59555	0.71111	0.82222	0.44000	0.54666	0.62666	0.66666
3	65792	65535	65278	65021	...	0.93777	1.07555	1.15111	0.74666	1.13333	1.50666	1.89777
4	65792	65535	65278	65021	...	3.66222	3.80444	3.89333	3.88888	4.01777	3.94666	3.73333
5	65792	65535	65278	65021	...	5.73777	6.27555	6.67111	4.16888	5.25777	6.02666	6.54666
6	65024	64770	64516	64262	...	3.84888	4.16888	4.34666	3.63555	3.68444	3.81333	4.07111
7	65792	65535	65278	65021	...	5.27555	5.89333	6.35111	3.56888	4.74666	5.43555	6.09333
8	65792	65535	65278	65021	...	2.43111	2.57333	0	0	0	0	0
9	65792	65535	65278	65021	...	2.96444	3.28000	1.19555	0.48888	0.60000	0.64888	0.65777
10	65792	65535	65278	65021	...	3.07555	3.05777	0	0	0	0	0
11	65792	65535	65278	65021	...	1.04000	1.20000	2.57333	1.82222	2.19111	2.70222	3.09777
12	65792	65535	65278	65021	...	3.10666	3.36888	3.28000	2.21777	2.89333	3.29333	3.53333
.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.		.	.	.	.	.	.
114	65792	65535	65278	65021	...	4.88888	5.35555	3.32444	4.49333	5.08000	5.28000	4.37333

Table 6. Transaction database

Image Class	Feature value interval label											
1001	2	26	50	74	120	144	168	170	194	218	242	266
1001	7	31	56	75	107	137	160	176	200	223	248	267
1001	11	31	56	85	112	137	160	179	202	223	248	277
1001	21	43	64	91	117	125	152	189	211	235	256	283
1001	20	45	70	95	99	123	147	188	211	237	262	287
1001	20	43	67	92	97	125	149	188	211	235	259	284
1001	19	43	67	94	100	125	149	187	211	235	259	286
1001	2	26	50	74	120	144	168	170	194	218	242	266
1001	8	31	56	75	105	137	160	176	202	223	248	267
1001	2	26	50	74	120	144	168	170	194	218	242	266
1001	15	37	62	89	103	131	154	183	207	229	254	281
1001	17	41	64	91	103	127	152	185	208	233	256	283
.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.		.	.	.	.	.	.
1002	13	34	59	88	111	134	157	181	205	226	251	280
1002	2	26	50	74	120	144	168	170	194	218	242	266

Table 7. Association rule optimization result

Dataset used	Brute-force approach	Apriori Algorithm	Optimization by constrained nonlinear minimization and minimax optimization algorithm	optimization by Genetic Algorithm
Synthetic image dataset	180	18	14	9
MIAS dataset	523250	14232	3702	1410

4 Conclusion

The authors proposed the association rule optimization method using two different optimization algorithms. First includes optimization of association rules using Multi-fitness function Genetic algorithm. Second is optimization of association rules using constrained nonlinear minimization and minimax optimization method. Proposed optimization algorithm is validated on synthetic image set containing geometric shapes and standard MIAS medical image dataset. Association rules referring to specific objects are found regardless of object position and object orientation. Interestingness correlation measures as Cosine, All-Confidence, Accuracy and Jaccard are linearly independent of each other. Using multi-fitness function Genetic algorithm these correlation measures are optimized.

Multi-fitness function Genetic algorithm generates more efficient, effective and strong association rules than constrained nonlinear minimization and minimax optimization method for image mining. The algorithm proved to be a critical approach in reducing number of redundant rules and complexity of system. Moreover, it can reduce the computation cost of image analysis and can easily be applied to other image analysis applications. Future scope includes, use of optimized association rules for mammogram classification.

ACKNOWLEDGEMENT

The authors would like to thank Dr. J. Suckling and co-authors for providing access to the dataset entitled “Mammographic Image Analysis Society (MIAS)”.

REFERENCES

[1]. J. Zhang, W. Hsu and M. L. Lee, *Image mining: trends and developments*. Journal of Intelligent Information Systems, 2002. 19(1): p. 7-23.

[2]. C. Ordonez and E. Omiecinski, *Discovering association rules based on image content.* Research and Technology Advances in Digital Libraries, 1999. Proceedings. IEEE Forum on, IEEE, 1999. p. 38-49.

[3]. C. Carson, S. Belongie, H. Greenspan and J. Malik, *Region-based image querying.* Content-Based Access of Image and Video Libraries, 1997. Proceedings. IEEE Workshop on, IEEE, 1997. p. 42-49.

[4]. M. Sahu, M. Shrivastava, *Image mining: a new approach for data mining based on texture.* IEEE International Conference on Computer and Communication Technology, 2012. p. 7-9.

[5]. R. Gonzalez and R. Woods, *Digital image processing*. Pearson Addison-Wesley Publications Co., Second Edition, March 1992.

[6]. J. Nagi, SA. Kareem, F. Nagi, SK. Ahmed, *Automated breast profile segmentation for ROI detection using digital mammograms.* Biomedical Engineering and Sciences (IECBES), 2010 IEEE EMBS Conference on, IEEE, 2010.

[7]. RM. Haralick, K. Shanmugam, IH. Dinstein, *Textural features for image classification.* Systems, Man and Cybernetics, IEEE Transactions on, 1973. 6: P. 610-621.

[8]. JC Felipe, AJ Traina, Jr C. Traina, *Retrieval by content of medical images using texture for tissue identification*. Computer-Based Medical Systems, 2003. Proceedings. 16th IEEE Symposium. IEEE, 2003. p. 175-180.

[9]. K. Beyer, J. Goldstein, R. Ramakrishnan, U. Shaft, *When is “nearest neighbor” meaningful?*. Database theory—ICDT’99, Springer Berlin Heidelberg, 1999. P. 217-235.

[10]. MX. Ribeiro, C. Traina, PM. Azevedo-Marques, *An association rule-based method to support medical image diagnosis with efficiency.* Multimedia, IEEE Transactions on, 2008. 10(2): p. 277-285.

[11]. R. Agrawal, T. Imieliński, A. Swami, *Mining association rules between sets of items in large databases*. ACM SIGMOD Record, 1993. 22(2): p. 207-216.

[12]. PN. Tan, M. Steinbach, V. Kumar, *Introduction to data mining.* Boston: Pearson Addison Wesley, Jun 2006.

[13]. S. N. Sivanandam and S. N. Deepa, *Introduction to genetic algorithm.* Springer Science and Business media, 2008.

[14]. J. Han, M. Kamber and J. Pei, *Data mining: concepts and techniques.* Elsevier, Jun 2011.

[15]. J. Suckling, J. Parker, D. Dance, S. Astley, I. Hutt, C. Boggis, I. Ricketts, E. Stamatakis, N. Cerneaz, S. Kok, P. Taylor, D. Betal, and J. Savage, *The mammographic image analysis society digital mammogram database.* In IWDM, 1994. P. 211–221.

[16]. A. Ghosh and B. Nath, *Multi-objective rule mining using genetic algorithms*. Information Sciences, Elsevier, 2004. 163(1): p. 123–133.

[17]. M. Saggar, AK. Agrawa and A. Lad, *Optimization of association rule mining using improved genetic algorithms.* IEEE International Conference on Systems, Man and Cybernatics, 2004. p. 3725-3729.

[18]. P. P. Wakabi-Waiswa, V. Baryamureeba and K. Sarukesi, *Optimized association rule mining with genetic algorithms*. Natural Computation (ICNC), 2011 Seventh International Conference on, IEEE, 2011. 2: p. 1116-1120.