International Journal of Computer and Communication Technology

Volume 6 | Issue 4 Article 1

October 2015

FEATURE SELECTION USING ASSOCIATION RULES FOR CBIR AND COMPUTER AIDED MEDICAL DIAGNOSTIC

C. BHUVANESWARI
Department of computer science , Theivanai ammal college for women (Autonomous) Villupuram,
bhuvana.csdept@gmail.com

P. ARUNA
Department of, Computer Science and Engineering , Annnamalai University chidambaram,
arunapuvi@yahoo.co.in

D. LOGANATHAN
Department of Computer Science and Engineering , Pondicherry Engineering college, pondicherry,
dloganathan@pec.edu

FEATURE SELECTION USING ASSOCIATION RULES FOR CBIR AND COMPUTER AIDED MEDICAL DIAGNOSTIC

C.BHUVANESWARI[1], P.ARUNA[2] & D.LOGANATHAN[3]

[1]Department of computer science , Theivanai ammal college for women (Autonomous) Villupuram
[2]Department of, Computer Science and Engineering , Annnamalai University chidambaram
[3]Department of Computer Science and Engineering , Pondicherry Engineering college, pondicherry
Email:bhuvana.csdept@gmail.com, arunapuvi@yahoo.co.in,dloganathan@pec.edu

Abstract-Digital images are now the basis of visual information in medical applications. The advent of radiology which employs imaging for diagnosis generates great amount of images. Automatic retrieval of images based on features like color, shape and texture is termed Content Based Image Retrieval. The increasing dependence of modern medicine on diagnostic techniques such as radiology, computerized tomography has resulted in a sudden increase in the number and significance of medical images. Content Based Image Retrieval techniques are being extensively used to aid diagnosis by comparing with similar past cases and improvising Computer Aided Diagnosis. In this paper, it is proposed to extract features in the frequency domain using Walsh Hadamard transform and use FP-Growth association rule mining to extract features based on confidence. The extracted features are classified using Naïve Bayes and CART algorithms and the proposed method's classification accuracy is evaluated. Experimental results show that classification accuracy for Naïve Bayes is 100 and 96.8 for CART on application of proposed method.

General Terms-Content Based Image Retrieval, Association Rule Mining

Keywords-*Walsh Hadamard Transform, FP Growth algorithm, Information Retrieval, Naïve Bayes, CART.*

1. INTRODUCTION

The expeditious growth of digital image databases motivated Content Based Image Retrieval which in turn requires efficient search schemes. Low level visual features including color, texture and shape, are automatically extracted to represent images. But low-level features might not accurately characterize high-level semantic concepts. Image retrieval techniques based on visual image content are widely researched for more than a decade. Many web search engines retrieves similar images through a search and match of textual metadata related to digital images. This paper addresses and analyses challenges and issues of CBIR techniques/systems, which has evolved over the past few years and which covered different segmentation methods; edge, boundary, region, color, texture, and shape based feature extraction; object detection and identification. For improved precision in retrieved images, searching requires associating meaningful image descriptive text labels as metadata with all database images. In real-world image retrieval systems, feedback relevance given by the user is usually limited; to typically less than 20, whereas image space dimensionality could range from several hundreds to thousands. Manual image labeling, generally called manual image annotation is practically difficult when it comes to exponentially increasing the image database.

The need evolves for two solutions with regard to automatic image annotation and content based image retrieval. Content based image retrieval techniques aims for efficient retrieval, in response to the query image (or sketch) with similar resultant images obtained from the image database. The database images are preprocessed for extraction, indexed and then stored corresponding to the image features. CBIR overcame the difficulties of manual annotations by using visual feature based representations like color, texture, shape, etc.

In this paper, it is proposed to use Walsh Hadamard transform for extracting features in the frequency domain and use FP-Growth association rule mining to extract features based on confidence. The extracted features are classified using Naïve Bayes and CART algorithms and the proposed method's classification accuracy is evaluated. The rest of the paper is organized as follows: Section 2 deals with some of the related works available in literature. Section 3 explains the methodology; experimental setup is detailed in section 4 and section 5 concludes the paper.

2. LITERATURE REVIEW

Ribeiro, et al., [5] proposed a method of Feature selection through Association Rules (FAR) in CBIR to improve performance. The dimensionality of the feature vectors were reduced using Association rules. The rules improved the precision of the similarity queries. Experiments performed validated the efficiency of the proposed algorithms. Results showed that the association rules successfully enhance the CBIR and support medical image analysis in medical systems.

Rajendran, et al., [6] proposed two hybrid image mining approaches for classification of medical images. The proposed method involved pre-

processing, feature extraction, association rule mining and hybrid classifier. Median filtering process and canny edge detection techniques were used in the preprocessing step. Frequent pattern tree (FP-Tree) algorithm was used to generate the frequent patterns in the images. The FP-Tree algorithm also mines the association rules. The proposed system improves the classification process. Experimental results show that 97% sensitivity and 95% accuracy was achieved using the proposed method on the pre-diagnosed database of brain images.

Bugatti, et al., [7] presented continuous feature selection techniques to improve the precision of content-based queries in image databases by removing noisy features. Continuous weights to each feature were assigned according to their relevance in the proposed method. Association rules were used to find patterns relating to low-level image features to high-level knowledge about the images. The weights for the features were determined by patterns mined. The feature weighting through the statistical association rules reduces the semantic gap, thus, improving the precision of the content-based queries. Experiments show that the proposed method improves the precision of the query results up to 38%.

Ribeiro, et al., [8] proposed StARMiner (Statistical Association Rule Miner) that aims at identifying the most relevant features from those extracted from each image, taking advantage of statistical association rules. The feature vectors condense the texture information of segmented images in just 30 features. The attributes selected by StARMiner with those selected by the C4.5 were compared, and found that the attributes selected by StARMiner maintains the retrieval ability of images higher. The proposed StARMiner achieved a reduction of 50% to 85% in the number of features.

Ordonez, et al., [9] presented a data mining algorithm to find association rules in 2-dimensional color images. The proposed algorithm has four major steps: feature extraction, object identification, auxiliary image creation and object mining. The proposed algorithm finds association rules in images without using domain knowledge and is fast and automated. The association rules are combined with automatically identified objects obtained from a matching process on segmented images. Rules refer to specific objects are obtained regardless of object position, object orientation. Experimental shows that image mining is feasible to obtain simple rules from not complex images with a few simple objects.

3. METHODOLOGY

In this paper, it is proposed to extract features using the Walsh Hadamard Transform (WHT) and select features using Feature Pattern Growth Feature Class (FPGFC) association rule such that the associability between each feature to the class is determined and apply Naïve Bayes and Classification and Regression Tree (CART) classification algorithm.

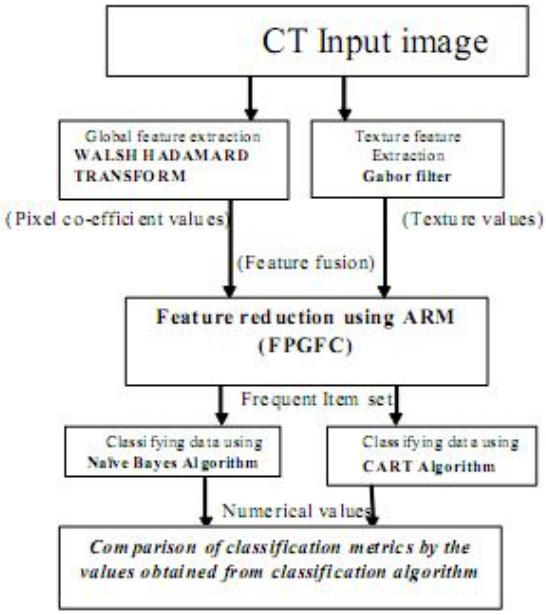

Fig 1 Flow diagram of feature selection

3.1 Walsh Hadamard Transform

The Walsh Hadamard Transform is the best known of the non sinusoidal orthogonal transforms which gained wide use in digital signal processing, as its application was easy with shortened processing time. The WHT shares the following with the discrete Fourier transform (DFT) [10]:

•It is involved with periodic finite series and provides a spectrum whose period contains the same sample number as the temporal sequence.

•It uses fast and efficient methods to compute algorithms in a few operations.

•It could be extended to multidimensional signals.

Additionally the WHT was advantageous due to the following reasons: (a) it has a real nature and (b) only additions and subtractions are needed to compute coefficients. The Hadamard transform Hm is a 2^m X 2^m matrix, and the $(k, n)^{th}$ entry is given in equation (1) and (2):

$$k = \sum_{0}^{i<m} k_i 2^i = k_{m-1} 2^{m-1} + ... + k_1 2 + k_0 \qquad (1)$$

$$n = \sum_{0}^{i<m} n_i 2^i = n_{m-1} 2^{m-1} + ... + n_1 2 + n_0 \qquad (2)$$

Where the k_j and n_j are the binary digits (0 or 1) of k and n.

The MRI image is given as input and the pixel co-efficient values are calculated using the Walsh hadamard transform for feature reduction.

3.2 Gabor filter

The Gabor filter is widely used in image processing, especially in texture analysis. The function is based on 'Uncertainty Principle' and can provide accurate time-frequency location [13]. The Gabor filters has

optimal localization properties in both spatial and frequency domain. The Gabor function is a harmonic oscillator, made of sine wave enclosed in a Gaussian envelope. A 2-D Gabor filter over the image domain (x,y) is given by eqn (3)

$$G(x,y) = \exp\left(-\frac{(x-x_0)^2}{2\sigma_x^2} - \frac{(y-y_0)^2}{2\sigma_y^2}\right) \quad (3)$$
$$\text{X}\exp\left(-2\pi i\left(u_0(x-x_0)+v_0(y-y_0)\right)\right)$$

where

(x_0, y_0) is location in the image,

(u_0, v_0) specifies modulation which has

frequency $\omega_0 = \sqrt{u_0^2 + v_0^2}$ and

orientation $\theta_0 = \arctan\left(v_0 / u_0\right)$

σ_x and σ_y are standard deviation of Gaussian envelope

The texture values are extracted using the Gabor filter and the values are given as a input for feature reduction using FPGFC.

3.3 Association Rule Mining

Association Rule Mining (ARM) task is to discover the hidden association relationship between the different itemsets in transaction database [11]. Let $I=\{i_1,i_2,\ldots i_m\}$ be a set of m distinct items. A transaction T is defined as any subset of items in I. A transaction database T is said to support an itemset x $\subseteq$ I if it contains all items of x. The fraction of the transaction in D that support x is called support value is above some user defined minimum support threshold then the itemset is frequent, otherwise it is infrequent. Maximum frequent itemsets have been denoted as F if all superset of frequent itemsets is infrequent itemsets. The maximum frequent itemsets that are discovered have been stored in the maximum frequent itemsets [12]. Maximum frequent candidate set, which is the smallest itemsets, it includes all current frequent itemsets known, but it does not include any infrequent itemsets. The identification of maximum frequent itemsets in earlier stage can reduce the number of candidate itemsets generated.

The Frequent Pattern Growth Feature Class (FPGFC) association rule builds a compact tree based on the following rules.

•Feature extracted correspond to items and form nodes of the tree. Class form the leaves.

•Features are normalized if they are continuous.

•Each sequence of image attributes is read as a 3x3 matrix and mapped.

•Counters are incremented when the attributes overlap with the same details

•Frequent items are selected as attributes for the classification algorithm.

Figure 1 shows the process. Assuming the attributes is given by equation (4)

$$A = \{a_1, a_2, a_3, a_4\} \quad (4)$$

Table1: Extracted dataset containing the overlaps

Class x	$\{a_2, a_3, a_4\}$
Class x	$\{a_3, a_4, a_5\}$
Class y	$\{a_1, a_3, a_4\}$

The FPGFC is derived from the tree shown in Figure 1 for the three overlaps extracted from the original transformed image data

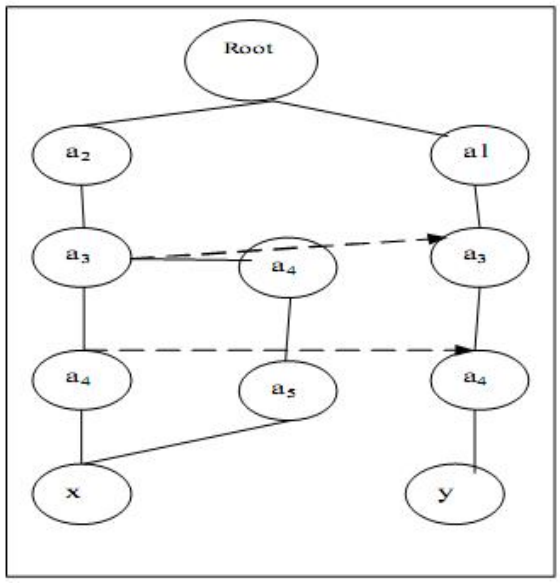

Figure 2: The frequent item sets discovery method using FPGFC

Naive Bayes Naïve Bayes are used for classifying the extracted features in this study. The extracted features are classified to the most likely class. Learning in Naïve Bayes is simplified by assuming that the features are independent for a given class. The feature is classified as shown in equation (5):

$$P(X|C) = \Pi_{i=1}^{n} P(X_i|C) \quad (5)$$

Where $X=(X_1,\ldots, X_n)$ is the feature vector and C is a class.

3.4 Classification and Regression trees (CART)

CART uses tree structures for classifying data. It can be applied to any dataset without setting any parameters and don't require any specification of functional form. Advance selection of variables is not necessary as splitting rules are determined in stepwise method. Although proper feature selection enhances CART's performance, the CART splits nodes on single variables and optionally uses linear combinations of non-categorical variables. Alternate

splits are created for each split to classify an object as shown in Fig 2.

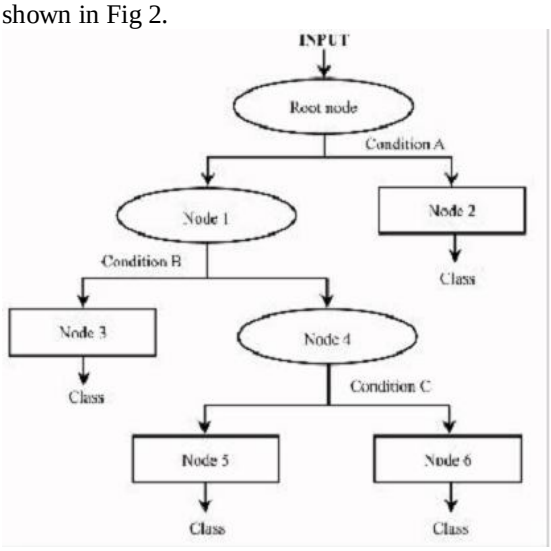

Figure 3: Frequent item sets discovery method using FPGFC

EXPERIMENTAL SETUP

In image processing, feature extraction is a special form of dimensionality reduction, when the input data to an algorithm is too large to be processed and it is suspected to be notoriously redundant then the input data will be transformed into a reduced representation set of features (also named features vector). Transforming the input data into the set of features is called feature extraction. Feature extraction involves simplifying the amount of resources required to describe a large set of data accurately.

35 Lung cancer images consisting of two class labels of images showing lung cancer and image without lung cancer were selected in this work. The lung cancer image is taken as input, resizing and conversion of gray scale of images are done, median filter is applied to remove noises, the Walsh Hadamard Transform is performed to return the co-efficients of discrete Walsh Hadamard Transform of the input X.

Texture feature are extracted using Gabor filter. Its impulse response is defined by a harmonic function multiplied by a Gaussian function. Because of the multiplication-convolution property (Convolution theorem), the Fourier transform of a Gabor filter's impulse response is the convolution of the Fourier transform of the harmonic function and the Fourier transform of the Gaussian function. Gabor filter results the values of the texture values.

3.5 Figure 4 shows some of the medical images used as input and Figure 5 shows the images of the feature set.

Figure 4: Lung Cancer Image

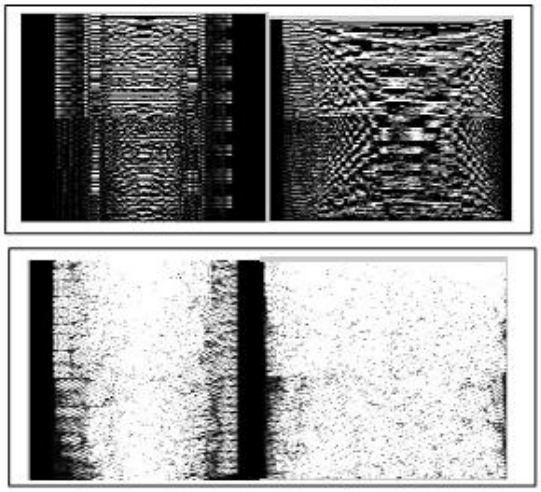

Figure 5 Feature Extracted from input image.

Feature selection refers to extracting pertinent features (or descriptors) from images that are important for differentiating one class of objects from another. The selected features are given as an input for association rule mining .The associations among the attributes with respect to the class label based on frequent set were selected. 200 item sets were selected. The extracted features were extracted and classified using Naïve Bayes and Classification and Regression Tree with 10 fold cross validation. The classification accuracy obtained was compared with and without feature selection using proposed Association Rule Mining. The accuracy increases while the association rule mining is performed which is shown below.

The classification accuracy obtained for both the classifiers is shown in

Figure 6. From Figure 7 it can be seen that the classification accuracy increases by 12.63 for Naïve Bayes and 7.37 for CART which was previously 87.37% and 89.47% respectively the resulting accuracy are 100 %and 96.8% .The root mean squared error is shown in Figure 7.

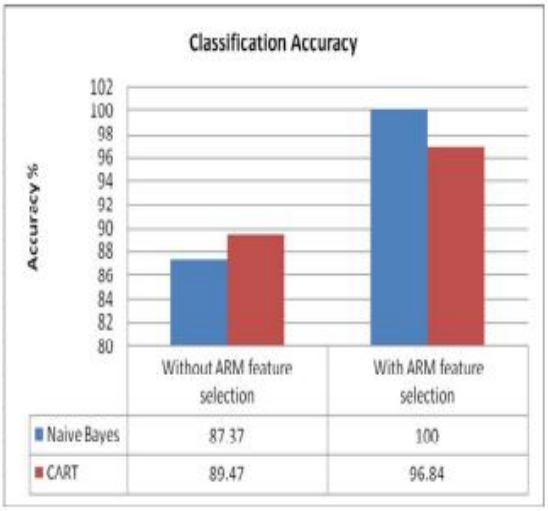

Fig 6: Classification accuracy.

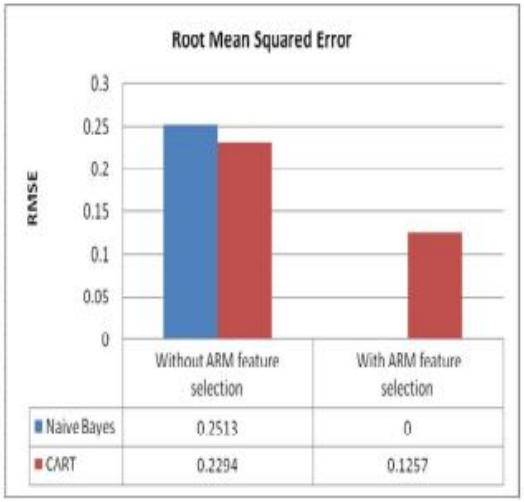

Fig 7: Root mean square error

5. CONCLUSION

In this paper, it was proposed to investigate the efficacy of feature selection and reduction using Association Rule Mining (ARM) on medical images. The features were extracted using the Walsh Hadamard Transform (WHT), features are selected using Feature Pattern Growth Feature Class (FPGFC) association rule such that the associativity between each feature to the class is determined. Naïve Bayes and Classification and Regression Tree (CART) classifiers were used for evaluating the accuracy of proposed method. It can be seen that the classification accuracy increases by 12.63 for Naïve Bayes and 7.37 for CART on application of proposed method which was previously 87.37% and 89.47% respectively the resulting accuracy are 100% and 96.8.**6.**

REFERENCES

[1] Smeulders, A., Worring, M., Santini, S., Gupta, A., Jain, R.: "Content-based image retrieval at the end of the early years". IEEE Trans.Pattern Anal. Mach. Intell. (2000), 22(12), 1349–1380.

[2] Lu, Y., Hu, C., Zhu, X., Zhang, H.J., Yang, Q.: "A unified framework for semantics and feature based relevance feedback in image retrieval systems". In: Proceedings of the 8th ACM International Conference on Multimedia, (2000), ACM Press, pp. 31–37.

[3] Smith, J.R., Basu, S., Lin, C.Y., Naphade, M.R., Tseng, B.: "Integrating features, models and semantics for content-based retrieval". In: Proceedings of the International Workshop on MultiMedia Content-Based Indexing and Retrieval (2001) (MMCBIR'01), pp. 95–98

[4] Zhou, X.S., Huang, T.S.: "Unifying keywords and visual contents in image retrieval". IEEE Multimedia (2002), 9(2), 23–33.

[5] Marcela X. Ribeiro, Pedro H. Bugatti, Caetano Traina Jr, Paulo M.A. Marques, Natalia A. Rosa, Agma J.M. Traina, "Supporting content-based image retrieval and computer-aided diagnosis systems with association rule-based techniques", Data & Knowledge Engineering, December 2009, Volume 68, Issue 12, Pages 1370–1382.

[6] P. Rajendran, M. Madheswaran, "Hybrid Medical Image Classification Using Association Rule Mining with Decision Tree Algorithm", Computer Vision and Pattern Recognition (cs.CV), January 2010, Journal of Computing, Vol. 2, Issue 1, PP:127-136.

[7]Pedro H. Bugatti, Marcela X. Ribeiro, Agma J. M. Traina, Caetano Traina Jr. 2008. "Content-based Retrieval of Medical Images by Continuous Feature Selection". 21st IEEE International Symposium on Computer-Based Medical Systems, June 17-19 2008, PP: 272-277.

[8] Ribeiro, M. X.; Balan, A. G. R.; Felipe, J. C.; Traina, A. J. M.; Traina Jr., C. "Mining statistical association rules to select the most relevant medical image features". In: First International Workshop on Mining Complex Data (IEEE MCD'05), Houston, USA: IEEE Computer Society, 2005, p. 91–98.

[9] C. Ordonez and E. Omiecinski. "Discovering Association Rules Based on Image Content". Proceedings of the IEEE Advances in Digital Libraries Conference (ADL'99), 1999, PP: 38-49.

[10] J. R. Johnson, R. W. Johnson, D. Rodriguez, and R. Tolimieri. "A methodology for designing, modifying, and implementing Fourier transform algorithms on various architectures". Circuits, Systems, and Signal Processing, 9(4):449–500, 1990.

[11] Sergey Brin, Rajeev Motwani, Jeffrey D. Ullman, Shalom Tsur, "Dynamic itemset counting and implication rules for market basket data", Proc. of the ACM SIGMOD Int'l Conf. on Management of Data, Tucson,AZ, USA, 1997, PP: 255-264.

[12] N. Pasquier, Y. Bastide, R.Taouil, and L.Lakhal, "Efficient mining of association rules using closed itemset lattices", Information Systems, Vol. 24, No. 1, 1999, pp. 25-46.

◈◈◈