An Efficient Approach for Medical Image Classification using Association Rules

A. Veeramuthu
Research Scholar
Sathyabama University
Chennai, India.

S. Meenakshi, Ph.D
Prof. & Head, Dept. of Information Technology
SRR Engineering College
Chennai, India.

ABSTRACT

Medical images are crucial in diagnosis, therapy, surgery, reference, and training. The availability of Digital radiology equipment availability has ensured that digital medical images management gets more attention now. This paper presents an automatic classification system for Computed Tomography (CT) medical images are presented in this paper. In the presented methodology, Bi-orthogonal spline wavelet was used to extract features from the brain, chest and colon CT scan images. Association Rule Mining (ARM) was used for feature reduction leading to the selection of attributes with respect to class label based on frequent sets. On feature selection the CT images are classified using Naive Bayes and k-Nearest Neighbor algorithm. The classification accuracy obtained was compared with and without feature selection using Association Rule Mining.

Keywords

Image Classification, Computed Tomography (CT), Bi-orthogonal Spline Wavelets, Feature Extraction, Feature Selection, and Association Rule Mining (ARM).

1. INTRODUCTION

Content-Based Image Retrieval (CBIR) is retrieves images based on their content, as against that based on the metadata. CBIR system's objective in the medical domain is to help specialists in diagnosis, retrieving past cases with the images to reveal proven pathology. This is in addition to related information on clinical diagnoses [1] many works are suggested the content based access to the medical images to aid clinical decisions (evidence-based medical practice) [2, 3]. They also proposed scenarios integrating of CBIR with PACS [4], and also into clinical routine [5]. Past cases image retrieval is done through comparing present case images with those in the database. They are then sorted according to similarity criteria used.

There are two reasons for using CT scans in radiotherapy planning, the first being that images contain information about anatomical, which helps to plan both direction and entry locations of radiotherapy rays to target a tumor and avoid risk to organs. The second is that CT scan images use rays, a principle similar to radiotherapy. This gains importance as radiotherapy ray intensity is computed from scanner image intensity [6].

A CBIR system indexes and retrieves tasks using computed image features as against using total images as features are image processing algorithm computed numerical values that capture and store image descriptions in feature vectors. Images similar to query images are returned based on distance measure during retrieval or a similarity based searching process. Knowing the application domain based most relevant/non-redundant features is necessary to improve similarity search accuracy. But, correlated and irrelevant information leading to a dimensionality curse results when many feature extractors are used [7] thereby retrieval process efficiency deteriorates.

Two methods to reduce dimensionality problems: feature extraction (FE) and feature selection (FS). Feature extraction changes feature (f-space) space leading to a lower dimensional, called transformed space. Lack of original dimensions reduces results “interpretability.” Usually, some features extracted automatically are not correlated to analysis target distorting results. Such attributes reduce discriminative power of those relevant when they are not removed through extraction. Feature selection (FS) locates a highly relevant feature subset, from original feature space, based on a user-defined criterion. Hence, this paper selects the feature selection methods instead of the feature extraction as the approach of choice to improve CBIR procedures. Feature selection (FS) is an important and constantly used data mining pre-processing technique [8] that reduces features number by removing irrelevant and/or noisy data. Its advantages include faster data mining algorithm convergence and accurate results.

This paper presents an automatic CT medical image classification system based on bi-orthogonal spline wavelets and association rule mining based CT medical images. Bi-orthogonal spline wavelets are used for feature extraction with association rules being applied for feature reduction in this study. CT images are classified with Naive Bayes and k-Nearest Neighbor algorithm after feature selection.

The remaining content of this paper is arranged as follows: Section 2 reviews some of the related works available in the literature; section 3 included the problem statement of the paper, section 4 details the overall architecture of the system is presented. Section 5 explains the experimental setup and the results achieved for classification, precision and recall for the presented method and section 6 concludes the paper.

2. RELATED WORKS

Ribeiro et al [9] presented StARMiner (Statistical Association Rule Miner) is a new algorithm that attempts to locate the "essence" of medical images through identification of relevant features from those from an image using statistical association rules advantageously. It also presents two case studies which use extracted features from segmented/non segmented medical images. The first reveals StARMiner behavior on feature vectors while condensing segmented images texture information in 30 features (attributes). The second uses a conventional 256 attribute feature vector from not-segmented images received through Zernike moments. Both case studies show and analyze the proposed algorithm on its ability to discriminate images in categories (benign/malignant tumors) and retrieve them. StARMiner selected attributes were compared to those selected by the known C4.5. It was seen that the former's attributes ensure higher image retrieval ability. StARMiner reduced features by 50% in the first study and by 85% in the next. It was also observed that similarity queries precision through use of reduced feature vectors was higher than when using the original.

An image mining approach regarding brain tumor classification in CT brain scan images was presented by Rajendran et al [10]. The process involved pre-processing, feature extraction, association rule mining and the hybrid classifier. Pre-processing was undertaken through the median filtering process with edge features being extracted through canny edge detection. This paper proposed two image mining approaches with a hybrid manner. Frequent Pattern tree (FP-Tree) algorithm generates frequent CT scan image patterns that mine association rules. Decision tree classifies medical images for diagnosis and it shown that it enhances classification accuracy. The proposed method's efficiency is improved by the hybrid method when compared to conventional image mining methods. Results from experiments on pre-diagnosed brain images database revealed 97% sensitivity and 95% accuracy. This helps physicians make accurate decisions in classifying brain images as normal, benign and malignant in diagnosis.

An association rules (AR) and neural network (NN) based automatic diagnosis system to detect breast cancer is presented by Karabatak et al [11]. AR reduces breast cancer database dimensions with NN being used for classification. The AR + NN system's performance is compared to the NN model. AR reduces input feature dimension space from nine to four. A 3-fold cross validation method was applied to the Wisconsin breast cancer database to evaluate performance of the suggested system at test stage. The proposed system's classification rate is 95.6% proving that AR reduce feature space dimensions while the proposed AR + NN model can obtain fast automatic diagnosis for other kinds of diseases.

A novel association rules (ARs) based technique was presented by Chaves et al [12] to locate relations among activated brain areas in single photon emission computed tomography (SPECT) imaging. This work aims to find out attribute associations which characterize normal subject's perfusion patterns for use in Alzheimer's disease (AD) diagnosis. First, voxel-as-feature-based activation estimation procedures locate tri-dimensional activated brain regions of interest (ROIs) for all patients. These ROIs become input for second mining of ARs among activation blocks through use of a control set. Here, the support and the confidence measures are proportional to functional areas which are activated across the brain both singly and mutually. Then image classification is performed through a comparison of AR numbers verified the tested subjects to a specific threshold which in turn is based on previously mined rules number. Classification experiments evaluated the proposed methods using a SPECT database having 41 controls (NOR) and 56 AD physician labeled patients. Leave-one-out cross validation strategy validated the proposed method, yielding classification accuracy up to 94.87%, thereby outperforming the recent computer aided diagnostic methods.

Dua et al. [13] presented a method to classify mammograms through use of a weighted association rule based classifier. Pre-processed images reveal regions of interest. Texture components extracted from an image's segmented parts are discretized for rule discovery. Associations rules are derived between various image segment extracted texture components and used for classification based on the intra class/inter class dependency. A common mammography dataset is classified through these rules with experiments evaluating rules efficacy under varied classification scenarios. Experiments proved that this procedure worked well for such datasets with as high as 89% accuracy surpassing accuracy rates of other rule based classification systems.

3. PROBLEM DESCRIPTION

Though various techniques in literature consider feature extraction process, not much work has been done in the feature selection process. Techniques used in literature consider statistical analysis for feature selection like Information Gain, Chi Square and Correlation. However frequent patterns are commonly available in medical images and not much work has been done to utilize this data for feature selection. This work proposed to select features based on associations between the features.

4. SYSTEM ARCHITECTURE

Though various medical image classification and retrieval systems are available in the literature, works related to CT images are limited. There is a need for automatic classification system for handling different kinds of CT images with high accuracy. This paper addresses the requirement of CT image classification. It reports the on-going investigation on proposing a classification system with high accuracy. This paper reports the initial investigations wherein existing methods are evaluated for medical image classification.

The architecture of the presented classification system is shown (see in Figure 1). The features of the CT images in the database are extracted using bi-orthogonal spline wavelets and the feature vectors are obtained by reducing the extracted features using ARM. The feature vectors are stored in Feature database. To classify an image, the input image's features are

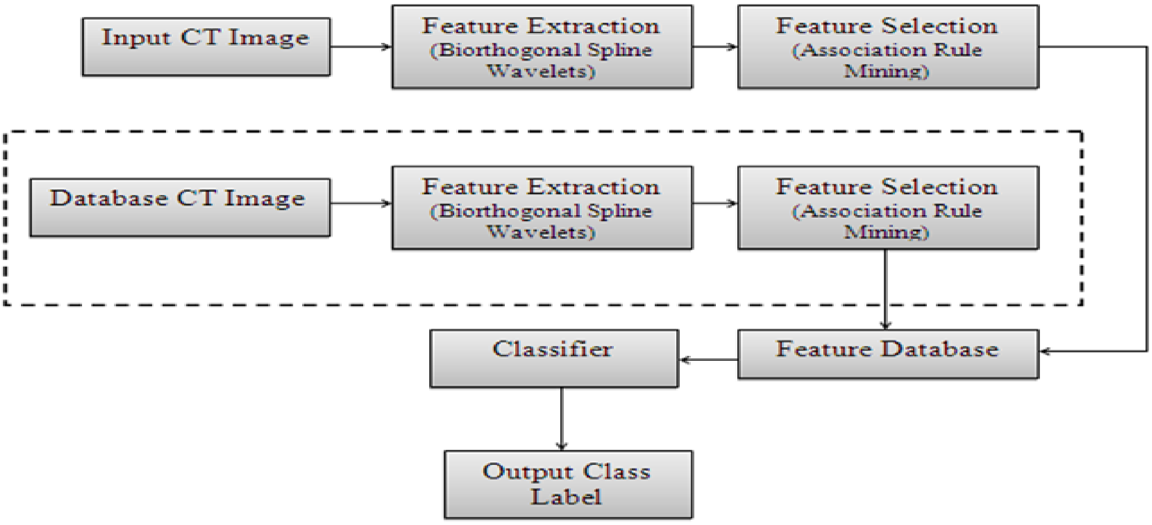
Input CT Image
Feature Extraction (Biorthogonal Spline Wavelets)
Feature Selection (Association Rule Mining)
Database CT Image
Feature Extraction (Biorthogonal Spline Wavelets)
Feature Selection (Association Rule Mining)
Classifier
Feature Database
Output Class Label

Fig 1: System Architecture for CT Image Classification

extracted and the extracted features are reduced to form feature vectors. The feature vector is then classified using Naive Bayes and k-Nearest Neighbors. The system outputs the class label of the input image.

4.1 Feature Extraction

Splines have explicit formulae in both the time and the frequency domain unlike other wavelet bases, ensuring easy manipulation. A progressive transition between two extreme multi resolutions is allowed. Spline wavelets are regular and generally symmetric or anti-symmetric and can also be designed to ensure compact support and achieve optimal time-frequency localization (B-spline wavelets). B-splines are underlying scaling functions which are short and regular scaling functions of order L. Finally, among all known wavelets of a given order L splines have best approximation properties. They ensure smooth approximating functions.

Cohen-Daubechies-Feauveau [14] introduced bi-orthogonal wavelet basis to obtain symmetric, regular and compactly supported wavelet pairs. But this is not compatible with orthogonality requirements to be dropped. Splines built bi-orthogonal wavelets are attractive due to their short support and regularity and are popular for coding [15]. Their symmetry and short support properties help reduce reconstructed images truncation artifacts. As such spline wavelets type is the least constrained of all it is impossible to build other non-compactly supported spline wavelets.

Construction of bi-orthogonal wavelet bases includes two multi-resolution analyses: one for analysis and the other for synthesis [14] usually denoted by $\{V_i(\tilde{\varphi})\}_{i\varepsilon Z}$ and$\{V_i(\varphi)\}_{i\varepsilon Z}$,

Where, $\tilde{\varphi}(x)$ and $\varphi(x)$are analysis and synthesis scaling functions. It is to be noted that $\tilde{\varphi}$ and φ can be a two-scale relation arbitrary solution.

Corresponding analysis and synthesis wavelets $\tilde{\Psi}(x)$and $\Psi(x)$ are constructed by taking scaling functions linear combinations equations (1) and (2) as follows:

$$\tilde{\Psi}\left(\frac{x}{2}\right) = \sqrt{2}\sum_k \tilde{g}(k)\tilde{\varphi}(x-k) \qquad (1)$$

$$\Psi\left(\frac{x}{2}\right) = \sqrt{2}\sum_k g(k)\varphi(x-k) \qquad (2)$$

4.2 Feature Reduction

Large set of data items have interesting association and/or relationships among Association Rule Mining (ARM). ARM show frequently occurring attributes value conditions in a dataset. Possible rules are allowed to be captured to explain the existing of some attributes based on the existing of others. Market Basket Analysis [16] is a typical example of ARM.

Let I = {i_1,i_2,….i_m} be a set of m distinct items. Transaction T is any items subset in I. A transaction database T is said to support item-set x ⊆ I if it contains all x items. The transaction fraction in D supporting x is called support value and if it is above some user defined minimum support threshold then item-set is frequent, otherwise it is infrequent. Maximum frequent item-sets are denoted F if all frequent item-sets superset is infrequent item-sets. Maximum frequent item-sets discovered are stored in maximum frequent item-sets [17]. Maximum frequent candidate set - the smallest item-sets -includes all known current frequent item-sets, but fails to include any infrequent item-sets. Maximum frequent item-sets identification earlier reduces generated candidate item-sets number. AR is useful in business applications, market basket analysis, store layout and item promotions, telecommunication alarm correlation, university course enrolment and texture and image processing.

Most ARM algorithms are differing variations of the Apriori algorithm, a state of the art algorithm [16]. It works iteratively first locating a set of large 1-item sets and then set of 2-itemsets and others. The scan number over a transaction database is equal to the length of the maximal item set. Apriori is based on the fact that simple but powerful observation generates a smaller candidate set using the large item sets set found in an earlier iteration. The Apriori algorithm is detailed below [17]:

```
Apriori() algorithm
$L_1$ = {large 1-itemsets}
k = 2
while $L_{k-1} \neq \phi$ do
begin
        $C_k$ = apriori_gen ( $L_{k-1}$)
          for all transactions t in D do
        begin
                $C^t$ = subset( $C_k$, t)
                For all candidate c ϵ $C^t$ do
                  c.count = c.count + 1
        end
        $L_k$ = { c ϵ $C_k$ | c.count >= minsup }
        k = k + 1
end
```

Apriori scans transaction databases D to count support of each item i in I to determine a set of large 1-itemsets. Iteration is then performed for each computation of set of 2-itemsets, 3-itemsets, and others.

5. EXPERIMENTAL SETUP AND RESULTS

Bi-orthogonal spline wavelet was used to extract features in experiments with 150 brain, chest and colon CT scan images. ARM was used for feature reduction leading to the selection of attributes with respect to class label based on frequent sets.500 item-sets were selected. The extracted features were classified using Naive Bayes and k-Nearest Neighbor with 10 fold cross validation. The classification accuracy obtained was compared with and without feature selection using proposed Association Rule Mining. Table 1 summarizes the results obtained. The classification accuracy and root mean square error (RMSE) obtained for both the classifiers is shown in Figure 2.

From Figure 2 it can be seen that the classification accuracy increases by 8.89% for k-Nearest Neighbor and 2.96% for Naive Bayes when ARM is used for feature reduction. The root mean squared error is the lowest for k-NN with ARM. Similarly, it is seen from figure 3 that the precision and recall for the k-NN is the highest.

Table 1 Summary of Results

Method	Classification Accuracy (%)	RMSE	Precision	Recall
Naive Bayes	88.15	0.2811	0.881	0.881
k-NN	84.44	0.3220	0.847	0.844
Naive Bayes with ARM	91.11	0.2434	0.911	0.911
k-NN with ARM	93.33	0.2108	0.939	0.933

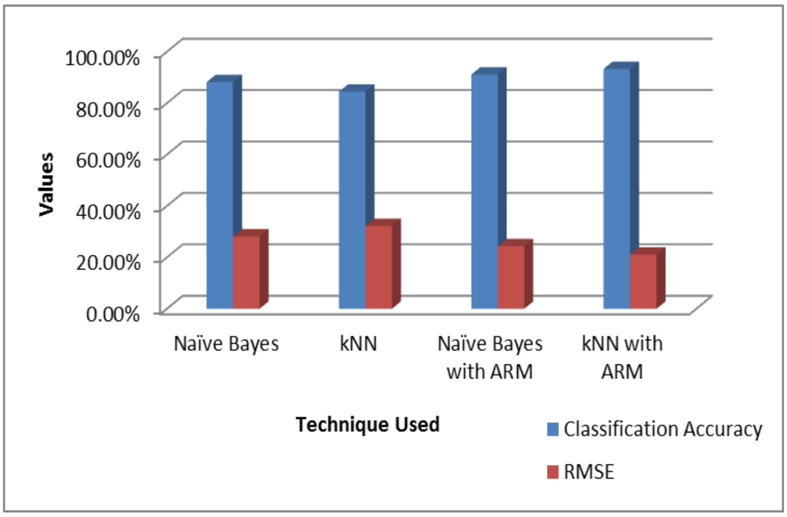

Fig 2: Classification and RMSE

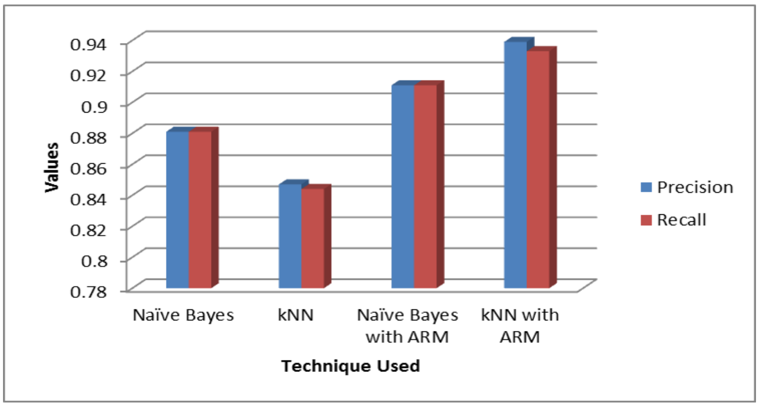

Fig 3: Precision and Recall

6. CONCLUSION

CBIR's objective in the medical domain is to help specialists in medical diagnosis retrieving relevant past cases with images that reveal proven pathology. This should be along with corresponding associated clinical diagnoses and related information. This paper investigates feature selection efficacy and reduction using ARM on CT medical images. Features extraction was through bi-orthogonal spline wavelets. Naive Bayes Classification and k-Nearest Neighbor classifiers evaluated the method's accuracy. Classification accuracy improved by 8.89% for k Nearest Neighbor and 2.96% for Naive Bayes when feature reduction was with the help of ARM. Precision and recall also shows great improvement. Further investigations to refine the feature selection process and classifiers are required to improve the classification accuracy.

7. REFERENCES

[1] H. Muller, N. Michoux, D. Bandon, A. Geissbuhler, 2004. A review of content-based image retrieval systems in medical applications – clinical benefits and future directions, International Journal of Medical Informatics - IJMI 73 pp.1–23. DOI: 10.1016/j.ijmedinf.2009.05.001.

[2] Aisen, L. Broderick, H. Winer-Muram, C. Brodley, A. Kak, C. Pavlopoulou, J. Dy, C.-R. Shyu, A. Marchiori, 2003, Automated storage and retrieval of thin-section CT images to assist diagnosis: system description and preliminary assessment, Radiology 228, pp. 265–270, DOI: 10.1148/radiol.2281020126.

[3] S.K. Kinoshita, P.M.d. Azevedo-Marques, R.R. Pereira Jr., J.A.H. Rodrigues, R.M. Rangayyan, 2007, Content-based retrieval of mammograms using visual features related to breast density patterns, Journal of Digital Imaging 20, pp.172–190, DOI: 10.1007/s10278-007-9004-0.

[4] C. Traina Jr., A. Traina, M. Araujo, J. Bueno, F. Chino, H. Razente, P. AzevedoMarques, 2005, Using an image-extended relational database to support content-based image retrieval in a pacs, Computer Methods and Programs in Biomedicine 80, S71–S83, DOI: 10.1016/S0169-2607(05)80008-2.

[5] M. Oliveira, W. Cirne, P. Azevedo-Marques, 2007, Towards applying content-based image retrieval in the clinical routine, Future Generation Computer Systems 23, pp.466–474, DOI: 10.1016/j.future.2006.06.009.

[6] Haiwei Pan, Xiaolei Tan, Qilong Han, Guisheng Yin, 2011, A Domain Knowledge Based Approach for Medical Image Retrieval, Information Engineering and Electronic Business, 3, pp.16-22, DOI: 10.1109/BICTA.2010.5645250.

[7] F. Korn, B. Pagel, C. Faloutsos, 2001, On the 'dimensionality curse' and the 'self-similarity blessing', IEEE Transactions on Knowledge and Data Engineering 13, pp. 96–111, DOI: 10.1109/69.908983.

[8] H. Liu, L. Yu, 2005, Toward integrating feature selection algorithms for classification and clustering, IEEE Transactions on Knowledge and Data Enginnering 17, pp.491–502, DOI: 10.1109/TKDE.2005.66.

[9] Ribeiro, M. X., Balan, A. G., Felipe, J. C., Traina, A. J., &TrainaJr, C., 2009, Mining statistical association rules to select the most relevant medical image features, Mining Complex Data, 165, pp. 113-131.

[10] Rajendran, P., &Madheswaran, M., 2010, Hybrid medical image classification using association rule mining with decision tree algorithm, Journal of Computing, 2, ISBN: 2151-9617, pp.127-136.

[11] Karabatak, M., & Ince, M. C., 2009, An expert system for detection of breast cancer based on association rules and neural network. Expert Systems with Applications, 36, pp.3465-3469, DOI: 10.1016/j.eswa.2008.02.064.

[12] Chaves, R., Górriz, J. M., Ramírez, J., Illán, I. A., Salas-Gonzalez, D., & Gómez-Río, M., 2011, Efficient mining of association rules for the early diagnosis of Alzheimer's disease, Physics in medicine and biology, 56, 6047, DOI: 10.1088/0031-9155/56/18/017.

[13] Dua, S., Singh, H., & Thompson, H. W., 2009, Associative classification of mammograms using weighted rules, Expert systems with applications, 36, pp.9250-9259, DOI: 10.1016/j.eswa.2008.12.050.

[14] A. Cohen, I. Daubechies and J.C. Feauveau, 1992, Bi-orthogonal bases of compactly supported wavelets,

Communications on Pure and Applied Mathematics, 45, pp. 485-560, DOI: 10.1002/cpa.3160450502.

[15] M. Vetterli and J. Kovacevic, 1995, Wavelets and Subband Coding, Prentice Hall, Englewood Cliffs, NJ.

[16] Agrawal, R., Imielinski T., & Swami, A., 1993, Mining association rules between sets of items in large databases, Proceedings of the ACM SIGMOD international conference, pp.1-10, DOI: 10.1145/170036.170072.

[17] N. Pasquier, Y. Bastide, R.Taouil, and L.Lakhal, 1999, Efficient mining of association rules using closed itemset lattices, Information Systems, 24, pp. 25-46, DOI: 10.1016/S0306-4379(99)00003-4.

[18] Agrawal, R., &Srikant, R., 1994, Fast algorithms for mining association rules in large databases, Proceedings of the 20th international conference on very large databases, pp. 487–499, DOI: 10.1023/A:1022852608280.