International Journal of Pure and Applied Mathematics
Volume 116 No. 22 2017, 375-385
ISSN: 1311-8080 (printed version); ISSN: 1314-3395 (on-line version)
url: http://www.ijpam.eu
Special Issue

ijpam.eu

Multiple Class-Association Rules for Content Based Image Retrieval with Efficiency

[1]R.I. Heaven Rose and [2]A.C. Subhajini

[1]Department of Computer Application,

Noorul Islam University, Kumaracoil, Tamil Nadu, India.

drheavenrose@gmail.com

[2]Department of Computer Applications,

Noorul Islam University, Kumaracoil, TamilNadu, India.

acsubhajini@yahoo.com

Abstract

Association rule based classification is one of the techniques for making more accurate decision and bringing more confidence to the diagnosis of medical images. Indeed, the performance of the association rule based classification degrades rapidly when the database size increases. In such situation FP-Tree (Frequent Pattern Tree) is the best option, but difficult with this approach is that it produces redundant FP-Tree. It requires lot of memory to store all transactions in the memory. In addition, the large size database continually produces large volume of association rules. It is difficult to accumulate, reclaim, reduce and arrange these rules well before applying to a classifier. To overcome these issues this paper proposes a new associative classification technique named CHiSAR(Classification based on Highly Strong Association Rules). This technique contains two phases, generation of rules and classification. First phase proposes a new compressed data structure known as Compact Frequent Pattern-Tree (CFP-Tree) for rule generation. It will speed up the rule mining process and consumes less memory to accommodate all transactions. The generated rules are organized in a CR-Tree and the rules are reduced efficiently by confidence, association and database coverage thresholds. Second phase, classification is done based on a weighted χ^2 analysis with numerous highly strong association rules. Our experiment shows that CHiSAR is highly effective and accurate at classification of US (Ultra Sound) Kidney images into normal, medical renal disease and cortical cysts while compare with CBA, C4.5, KNN and Naïve Bayes.

Key Words: Image mining, FP-Tree, CFP-Tree, association rules mining, classification, associative classification, CBA.

1. Introduction

In medicinal field two kinds of computer systems are commonly used; the Content-Based Image Retrieval (CBIR) and Computer-Aided Diagnosis (CAD). CAD is aiming at providing second opinions to radiologist to aid in diagnosis. The aim of CBIR in medicine is to assist physicians in diagnosis, retrieving cases or images to increase the physician's self-confidence on explaining test image details. At present, image mining is one of the demanding field which combines both data mining and image processing techniques. The important task of mining image is successfully connecting pixel based information to object based information. It is difficult to relate the pixel information of image to object information. This is semantic gap. The aim of image retrieval is to reduce the gap of semantic and progress the precision of image retrieval. There are many technologies have been developed to reduce semantic gap such as object ontology, machine learning, relevance feedback, web image retrieval. Generally, the image retrieving task consists of extracting several features like concentration, contour, shade, texture, dimension and position, indexing the features using image individuality and retrieving imagery from database that are related to the input image. The Supervised machine learning method associates low level visual information to query concepts and predicts the query image class label on the basis of user input. In supervised machine learning, large amount of images is collected, and classifier is trained to find out the new image's category label. Bayesian classifier is a famous method to automatically classify general type image database as indoor or outdoor i.e. city or landscape, etc.

In medicine, the important area of the diagnosis is to develop a sub set rules for classifying images by using its characteristics. To enhance the diagnosis of medical images, association rule mining based methods are widely used. The aim of this method is to deriving knowledge from image database by using pixel based information [1]. The association rule based classification provides an idea of diagnosis to physician for diagnosing of medical images. In this method of kidney failure detection, each training image is combined with a set of keywords which are the representative terms preferred by the specialist for accurate results. Association rule mining involves in efficient classification of US kidney images into three categories: normal, medical renal disease and cortical cysts [2]. In this paper, an efficient technique CHiSAR have been proposed which assists to make accurate and efficient classification of medical images. This technique also improves the precision of image retrieval system. The major influences of the proposed technique in classification are following. First, rather than applying a simple rule for classification, CHiSAR applies a class name wise rules for classification. CHiSAR choose a set of strong rules then examines the association between those rules. Second, CHiSAR uses a new tree structure CR_Tree which helps to record all rules compactly and to retrieve sub set rules for classification efficiently. Third, instead of making FP-Tree for mining rules, CHiSAR creates a CFP_Tree to increase the speed of the mining

and consumes less memory to accommodate all the transactions. The remaining contents of the paper is structured like follows. Section 2 explains about related works and technique used for preprocessing. Section 3 describes the proposed system which includes multiple association rule mining method and multiple rules based classification. The experimental results and the performance report are explained in Section 4. Conclusion of this paper is presented on Section 5. Section 6 deals with References.

2. Related Works

In data mining, association rule mining is one the most important technique to find out the pattern that are not explicitly stored in the data base. Many researchers develop computer aided systems for medical image classification and diagnosis. Association rule based classification method provides accurate classification than other well-known classification methods. Association rule mining technique is used to find out association between objects using transactional database [3]. Boolean association rule mining technique is good for classification. The Boolean rule mining process contains three stages. In the first stage the databases used for mining are transformed into a transactional database. Second, mining technique is applied on the data base and generates the frequent patterns. Finally, generates the association rules from the frequent patterns using support and confidence thresholds. Apriori and FP-Growth are the two well-known data mining algorithms used to extract association rules form transaction data base [4]. In paper [5] the detail of content based image recovery frame work using binary association rules is given. Binary association rules represent the significant feature of images in rule form, regarding a specific criterion. In paper [6], the authors propose a classification descriptor using moment invariants values. The author explained about the revised efficient numerical algorithms and established some concrete examples of real time application with moment invariance values. Furthermore, the authors explained about the methods to develop moment invariant functions for medical image diagnosis. In paper[7], the association rule based image classification is described. In this paper, association rules are mined from Frequent Pattern tree and medical images are classified by using decision tree based method. In paper [8], the author proposed a Modified Frequent Pattern Mining algorithm which uses a MFP-Tree data structure to mine association rules. This is more efficient than FP-Tree based mining algorithms. In paper [9], the authors introduced a Novel Fuzzy Association rule mining algorithm (NFARM) which is used for medical image diagnosis. This method supports to diagnose brain tumor exist on brain images. The NFARM algorithm provides multiple keywords to detect the cancerous cells. This helps the physicians for making good decision. In paper [2], the authors proposed an algorithm ARCKi which classifies the given images and provides the second opinion to physician for making accurate diagnosis. In paper [10], the authors offered the various reviews summary on breast cancer diagnosis and prediction. All the above, the authors have used Apriori or FP-Growth algorithms for rule mining. While compare both of these

FP_Growth is efficient than Apriori[13]. The main issue about Apriori algorithm is it requires many database scan to find frequent items. But, FP-Growth algorithm needs only two scans to find frequent items. But problem on FP-Growth algorithm is, for the large database FP-tree structure used by the FP-Growth can not fit into memory. As for the large database, FP-Growth cannot generate all set of rules for classification. The limited rules generated from the tree do not support accurate classification. To overcome these problems, this paper proposes a Compact Frequent Pattern tree (CFP-tree) data structure to store all transaction before mining classification rules. The proposed technique classification based on Highly Strong Association Rules (CHiSAR) uses the CFP-Tree data structure for rule generation. The generated rules are organized in a compact structure called CR-Tree. It supports to retrieve rules efficiently. The rules are pruned using confidence, correlation and database coverage thresholds. The classification process is performed using weighted χ^2 test using a set of highly strong association rules.

Preprocessing Method for Association Rule Mining

During preprocessing continuous values are converted into discrete ones using IICBMergeFS (Interval Inconsistency Based Merge Feature Selection) algorithm presented in [11]. The aim of preprocessing in Association Rule mining is to reduce the number of intervals and keeps less inconsistency between two adjacent intervals. In our work, during pre-processing step the IICBMerge algorithm scan the feature vector and removes all inconsistent and irrelevant features. The improved feature vector for each image is formed and organized into transaction database. The transaction that model the training data are simplified into $T_i = \{C, f_1, f_2, \ldots f_n\}$, where C denotes the class label and f_1, f_2 $\ldots f_n$ denotes the values of features.

3. Proposed System

We propose a supporting Content Based Image Retrieval through Classification with efficiency.

The proposed technique CHiSAR (Classification based on Highly Strong Association Rules) added with CBIR helps physicians to make better diagnosis. CHiSAR consists of two phases.

i. Generation of Association Rules
ii. Classification Using Subset of Association Rules

Figure 1 shows the design of the suggested technique. In rule generation step, CHiSAR generates group of rules of the form R_i: A, B→c, where A, B is a frequent item sets of the training data, and c is the class name. Support and confidence are two important thresholds. Moreover, CHiSAR performs rule pruning operation and selects only a sub set of highly strong classification rules.

During second stage, CHiSAR selects only a subset of rules which are matching the rules generated from the test image and predicts the class name of the test

image.

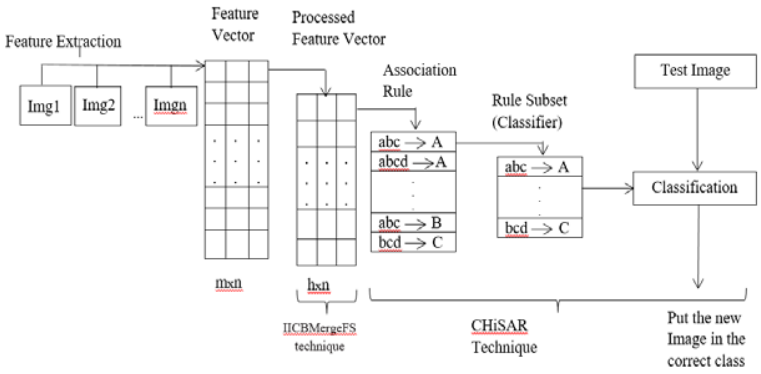


Figure 1: Classification based on Highly Strong Association Rules

Discovering Association Rules Using Multiple Thresholds

Discovering classification association rules consumes lot of time. To improve the execution time and data storage we propose a new method named CHiSAR which helps to store all transactions efficiently and also helps to retrieve rules efficiently. The basic hints of rule mining in CHiSAR is follows.

Input: Database D, min_support threshold

Output: CFP-tree

i. Scan the database D once and finds the item that support count is greater than the min_support threshold. Next, arrange them in descending order of their support count. i.e., FI_list (Frequent Item list)
ii. Map the FI-List items in the index form with index number and its support.
iii. For each transaction T
iv. Map the frequent items in the index form, and arrange in ascending order of their index value.
v. The item from FI-list having maximum support is taken as root node of the tree and assign index 1 to it. Then add the other items in the left most branch of the tree. The final node in the path indicates the class name.
vi. If the starting item index is new, then adjust the pointer and create sub trees and increase the index value or level.
vii. End for

The constructed CFP_tree is called as global CFP-tree. Once the global CFP_tree has been built then, construct local CFP_tree for every index starting from the minimum support count value.

Algorithm for mining classification rule from CFP_tree is:

Input: global CFP_tree, support threshold

Output: frequent patterns

i. Start from the least index item, collect the paths formed by this index in the global tree.
ii. For each combination of the indexes in the path.
iii. Check the support count of all indexes. The index items having less minimum support count are removed.
iv. Mapping the remaining index items with the support.
v. Creates a projection for all transaction ending with these indexes. Local CFP-tree represents this projection.
vi. Making link between the items similar to the link between items in global CFP_tree.
vii. Assign new index ids to all frequent items of that projection.
viii. Continue the above steps until use all index.

As a result, the CFP-tree supports to mine the frequent items using local and global CFF_tree, remove the local non-frequent items and utilize less memory space by mining local CFP_tree one by one. Hence, large database can also easily fit into main memory.

Pruning Methods

The large amount of rules does not support for efficient classification. So that the rules must be pruned by removing repeated and unwanted rules.

The methods used for pruning rules is described below.

Method 1:

General and high confidence rules are used to prune specific and low confidence rules. While the new rule is inserted, this method verifies if the inserted rule can be pruned or it can be prune other rules that are already inserted.

Method 2:

Choosing positively correlated rules. We make test for each rule if its antecedent is positively correlated with its precedence. This is performed by χ^2 test. If the rule is positively correlated means those with χ^2 result exceeds a significance range of thresholds.

Method 3:

Pruning rules using database coverage. In this method for each rule, first find all objects of the training data set matching rule R. If r correctly classifies at least one data object, then select R and increase the cover _count of those data object. Remove the data object when its cover_count exceeds coverage threshold.

Highly Strong Association Rules Based Classification

Sub set of rules selected after pruning step that represents the classifier. For a given image, the classifier searches the set of rules to find the class name.

The algorithm used for classifying new image is described below.

Algorithm: CHiSAR (Classification based on Highly Strong Association Rules)

Input: feature vector of the test image V, confidence α.

Output: keywords K

Algorithm

i. For each rule r in CHiSAR is the form: X→C
ii. do
iii. {for each itemset x∈X
iv. do
v. {
vi. Check C matches V, then
vii. increase matches count by 1
viii. else
ix. increase number of non matches by 1
x. }}
xi. If number of matches>1 then
xii. Categorize the rules according to class names.
xiii. For each category do
xiv. The rule with weighted $\mathcal{X}^2$ is selected.

For rule R , we define weighted $\mathcal{X}^2$ using formula (1):

$$\text{weighted } \mathcal{X}^2 = (\text{minimum}\{\sup(I), \sup(C)\} - \frac{\sup(I)\sup(C)}{|N|})^2 |N| a \dots\dots\dots\dots (1)$$

Where

$$a = \frac{1}{\sup(I)\sup(C)} + \frac{1}{\sup(I)(|N| - \sup(C))} + \frac{1}{(|N| - \sup(I))\sup(C)} + \frac{1}{(|T| - \sup(I))(|N| - \sup(C))} \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots (2)$$

Weighted $\mathcal{X}^2$ calculates the upper limit of $\mathcal{X}2$ value of the rule with respect to fixed of other settings. The weighted $\mathcal{X}2$ value of the group for each group of rules is defined as $\sum \frac{x^2 x^2}{\text{weighted } x^2}$

4. Experimental Results and Performance Analysis

We have tested our proposed system with the Ultra Sound Kidney images obtained from signal processing laboratory (http://www.splab/en/).The database contains various types of US kidney images such as normal, cortical cyst (cc), medical renal diseases (mrd) are used for our experiments. Texture features are extracted and put in a training database. To find classification rules, CHiSAR first scans the training data set to find complete frequent patterens using support threshold. And mines association rules from the frequent item set using support and confidence thresholds. Then the mined rules are organized in a CR_Tree. For effective classification, we have pruned rules by deleteting redundant and

unwanted rules. After prepare a sub set rules, CHiSAR redy to classify the test image. The accuray of the proposed method is evaluated by comparing the classfication accuracy of the exisiting classification methods like CBA, C4.5, KNN and Naïve Bayes. The five methods paramenters are set as follows.

All parameters of C4.5 are assigned to defaulting values. The C4.5 method based on association rule gives better accuracy than C4.5 decision tree and rule based methods. We have assign support threshold to 40% and confidence threshold to 50% and no limit on number of rules for CBR. For Naïve Bayes and KNN all parameters are set to default values. For CHiSAR, the support and confidence thresholds are set as same as CBA. The database coverage threshold is assigned to 4 and the confidence threshold to 50%. We have taken US kidney images from signal processing laboratory (http://www.splab/en/). A total of 350 images which belongs to three categories: normal, medical renal disease and conical cyst are used to the performance of the proposed method. There are 208 normal images, 84 medical renal disease images and 58 conical cyst images. Out of 350 images 300 images are used as training set and the remaining 50 images are used as testing. During training phase, we compute all the 140 features from each image. The feature values are organized as feature vector. Feature vector is reduced by using IICBMerge algorithm [11]. The processed feature vector is called training data set. The proposed technique mines the transaction database to find the complete set of rules and selecting rules based on database coverage. The set of rules are selected after pruning is represent the actual classifier. During the testing phase, the new classifier classifies the test image as normal kidney or medical renal disease kidney image or conical cyst kidney image using selected rules. As can be seen from the table, CHiSAR outperforms than CBA, C4.5, KNN and Naïve Bayes on accuracy. CHiSAR wins 87.5% of test data sets. Accuracy of the process is computed by

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \ldots\ldots\ldots \quad \ldots(3)$$

The results are shown in Table 1.

Table 1: The Comparison of C4.5,CBA,Naïve Bayes,KNN and CHiSAR on Accuracy

Dataset	attr	class	rec	CBA(%)	C4.5 (%)	KNN (%)	Naïve Bayes (%)	CHiSAR (%)
Kidney image	140	3	3500	86.5	86.14	85.24	82.48	87.5

From Figure 2 it can be seen that the classification accuracy increased by 1% than CBA, 1.36% than C4.5, 2.26% than KNN and 5.02% than Naïve bayes.

Table 2: The Comparison of CBA and CHiSAR on Main Memory Usage

Dataset	attr	class	rec	CBA mem(M)	CHiSAR mem(M)	Saving(%)
Kidney image	140	3	3500	590	88	85.09%

This shows that, association ruled based classification gives better accuracy than with out using rule mining.

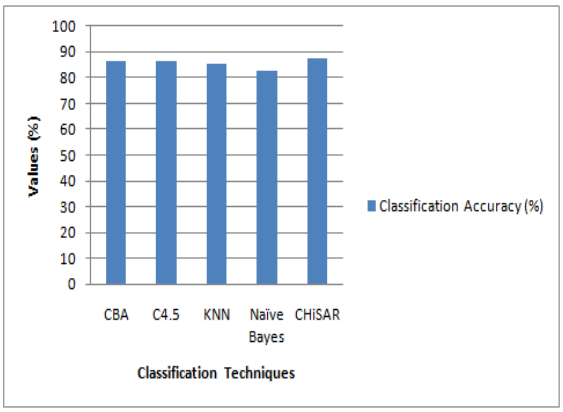


Figure 2: Classification

CR_tree is a compact structure used to record all mined rules. The effectiveness of CR_tree is achieved by comparing the main memory usage of CBA and CHiSAR using large data base. Memory usage status of both CBR and CHiSAR are shown in Table 2.

5. Conclusion

This paper focused two major challenges of associative classification. Fist one is efficiency at handling large number of mined association rules. Second one is effectiveness at predicting new class names with high accuracy. We have proposed a novel associative classification technique, CHiSAR, i.e., Classification based on Highly Strong Association Rules. This technique contains some distinguished features such as the classification is performed by using weighted χ^2 study imposed on set of association rules. This is important to improve overall classification accuracy. And, it prunes rules effectively based on confidence, correlation and database coverage. Moreover, the efficiency of the classification is achieved by using the extension of efficient frequent pattern mining method, FP-Growth, construction of CFP-Tree and applying a CR-Tree to store and retrieve mined rules efficiently. Our experiment on US Kidney image data base in repository show that CHiSAR is considerably, highly effective at classification and has better average accuracy in comparison with CBA, C4.5, KNN and Naïve Bayes. Moreover, it is more efficient, effective and scalable than other associative classification methods.

References

[1]. Ribeiro M.X., Traina A.J., Traina Jr C., Azevedo-Marques P.M., An association rule-based method to support medical image diagnosis with efficiency, IEEE transactions on multimedia 10(2) (2008), 277-285.

[2]. Dhanalakshmi K., Rajamani V., An Efficient Rule Based Method for diagnosing Ultrasound Kidney Images, IEEE International Conference on Computational Intelligence and Computing Research (2010).

[3]. Imielinski T.T. Virmani A., MSQL: A Query Language for Database Mining, Data Mining and Knowledge Discovery 3(2) (1999), 373-408.

[4]. Goebel M., Gruenwald L., A Survey of Data Mining and Knowledge Discovery Software Tools, SIGKDD Explorations 1(5) (1999), 20-33.

[5]. Akila, Uma Maheswari, Content Based Image Retrieval Using Binary Association Rules, Journal of Computer Science 9(6) (2013), 678-689.

[6]. Flusser J., Moment Invariants in Image analysis, World Academic Science Engineering Technology 1 (2012), 376-381.

[7]. Rajendran P., Madheswaran M., Hybrid Medical Image Classification Using Association Rule Mining with Decision Tree Algorithm, Journal of Computing 2(1) (2012).

[8]. Jun Gao, A New Association Rule Mining Algorithm and its Application, 3rd International Conference on Advanced Computer Theory and Engineering (2010).

[9]. Rajendran P., Madheswaran M., Novel Fuzzy Association Rule Image Mining Algorithm for Medical Decision Support System, International journal of Computer Applications 1(20) (2012).

[10]. Shweta Kharya, Using Data Mining Techniques for Diagnosis and Prognosis of Cancer Disease, International Journal of Computer Science, Engineering and Information Technology 2(2) (2012).

[11]. Heaven Rose R.I., Subhajini A.C., Improving Feature Selection Through Global Discretization for Content Based Image Retrieval, International Journal of Control Theory and Applications 10 (2017).

[12]. Agrawal R., Srikant R., Fast Algorithms for Mining Association Rules, Proc. 20th int. conf. very large data bases (1994), 487-499.

[13]. Han J., Pei J., Yin Y., Mining Frequent Patterns without Candidate Generation, SIGMOD (2000).