International Journal of Information Technology and Knowledge Management
January-June 2011, Volume 4, No. 1, pp. 143-146

ASSOCIATION RULE MINING BASED ON IMAGE CONTENT

Deepa S. Deshpande

Image mining is concerned with knowledge discovery in image databases. We present a data mining approach to find association rules based on image content. The Data mining approach has four major steps: Preprocessing, Feature Extraction, Preparation of Transactional database and Association rule mining. The purpose of our experiments is to explore the feasibility of data mining approach.. Results will show that there is promise in image mining based on content.
Mammography is one of the best methods in breast cancer detection, but in some cases, radiologists cannot detect tumors despite their experience. Computer-aided method using association rule could assist medical staff and improve the accuracy of detection. It is well known that data mining techniques are more suitable to larger databases than the one used for these preliminary tests. In particular, a Computer aided method based on association rules becomes more accurate with a larger dataset. Traditional association rule algorithms adopt an iterative method to discovery frequent item set, which requires very large calculations and a complicated transaction process. Because of this, a new association rule algorithm is proposed in this paper. Experimental results show that this new method can quickly discover frequent item sets and effectively mine potential association rules.

1. Introduction

Advances in image acquisition and storage technology have led to tremendous growth in very large and detailed image databases. A vast amount of image data such as satellite images, medical images, and digital photographs are generated every day. These images, if analyzed, can reveal useful information to the human users. Unfortunately, it is difficult or even impossible for human to discover the underlying knowledge and patterns in the image when handling a large collection of images. Image mining deals with the extraction of implicit knowledge, image data relationship, or other patterns not explicitly stored in the image databases. The images from an image database are first preprocessed to improve their quality. These images then undergo various transformations and feature extraction to generate the important features from the images. With the generated features, mining can be carried out using data mining techniques to discover significant patterns. The resulting patterns are evaluated and interpreted to obtain the final knowledge, which can be applied to applications.

In traditional association rule mining, an association rule is represented in LHS => RHS form with both LHS and RHS allowed to contain multiple items. Support of an association rule is defined as the percentage of transactions that contains all items (both LHS and RHS) in an association rule and confidence of an association rule is defined as the percentage of LHS items that also contains RHS. An association rule holds if its support is greater than minsup and confidence is greater than minconf, where minsup and minconf are configurable. The problem of finding association rules is decomposed into sub-problems of finding all item sets with at-least minimum support (also called large item sets) and using these large item sets to generate the desired rules (tested for minimum confidence). Large item set generation is achieved by generating candidate item sets and keeping the ones with minimum support. This requires very large calculations and a complicated transaction process. The discovery of association rules is typically done in two steps: discovery of frequent item sets and the generation of association rules. The second step is rather straightforward, and the first step dominates the processing time, so this paper explicitly focuses on the first step by proposing new algorithm.

Department of Computer Science & Engineering, MGM's Jawaharlal Nehru Engineering College Aurangabad (Maharashtra), Dr.B.A.M.U. University, India

Email: deepadsd@yahoo.com

2. Scheme for Experiment

Step-I : Pre-processing Phase: Since real-life data is often incomplete, noisy and inconsistent, pre-processing becomes a necessity. In our case, we had images that were very large (typical size was 1024 x 1024) and almost 50% of the whole image comprised of the background with a lot of noise. In addition, these images were scanned at different illumination conditions, and therefore some images appeared too bright and some were too dark. The first step toward noise removal was pruning the images with the help of the crop operation in Image Processing. Cropping cuts off the unwanted portions of the image. Thus, we eliminated almost all the background information and most of the noise.. The next step towards pre-processing the images was using image enhancement techniques. Image enhancement helps in qualitative improvement of the image with respect to a specific application. Enhancement can be done either in the

spatial domain or in the frequency domain. Here we work with the spatial domain and directly deal with the image plane itself. In order to diminish the effect of over-brightness or over-darkness in images, and at the same time accentuate the image features, we applied the Histogram Equalization method, which is a widely used technique. The noise removal step was necessary before this enhancement because, otherwise, it would also result in enhancement of noise.

Step-II : Feature Extraction Process: Once the pre-processing is applied, an extraction process is used in order to extract texture feature using GLCM technique Statistical parameters such as Standard deviation, Mean, Moments, Smoothness, Uniformity, Entropy can be extracted from the preprocessed images by using GLCM(Gray Level Co-occurrence Matrix).

GLCM of an image is computed using a displacement vector d, defined by its radius ä and orientation è. Frequency normalization can be employed by dividing value in each cell by the total number of pixel pairs possible. Hence the normalization factor for 0° would be $(N_x - 1) \times N_y$ where N_x represents the width and N_y represents the height of the image. The quantization level is an equally important consideration for determining the co-occurrence texture features. Also, neighboring co-occurrence matrix elements are highly correlated as they are measures of similar image qualities. Each of these factors is discussed ahead in detail.

Choice of radius δ:

δ value ranges from 1, 2 to 10. Applying large displacement value to a fine texture would yield a GLCM that does not capture detailed textural information. It has been observed that overall classification accuracies with δ = 1, 2, 4, 8 are acceptable with the best results for δ = 1 and 2. This conclusion is justified, as a pixel is more likely to be correlated to other closely located pixel than the one located far away.

Choice of angle θ: **Every pixel has eight neighboring pixels allowing eight choices for θ, which are 0°, 45°, 90°, 135°, 180°, 225°, 270° or 315°. However, taking into consideration the definition of GLCM, the co-occurring pairs obtained by choosing è equal to 0° would be similar to those obtained by choosing è equal to 180°. This concept extends to 45°, 90° and 135° as well. Hence, we have four choices to select the value of ¸. Sample texture measures of mammogram images are given below:**

Moment	Expression	Measure of texture
Mean	$m = \sum_{i=0}^{L-1} z_i p(z_i)$	A measure of average intensity
Standard deviation	$\sigma = \sqrt{\mu_2(z)} = \sqrt{\sigma^2}$	A measure of average contrast
Smoothness	$R = 1 - 1/(1 + \sigma^2)$	Measures the relative smoothness of the intensity in a region.
Third Moment	$\mu_3 = \sum_{i=0}^{L-1} (z_i - m)^3 p(z_i)$	Measures the skew ness of a histogram
Uniformity	$U = \sum_{i=0}^{L-1} p^2(z_i)$	Measures the uniformity of intensity in the histogram
Entropy	$e = -\sum_{i=0}^{L-1} p(z_i) \log_2 p(z_i)$	A measure of randomness

Step-III: Preparation of Transactional Database: The extracted features are organized in a database in the form of transactions, which in turn constitute the input for deriving association rules. The transactions are of the form [Image ID, F1; F2; :::; Fn] where F1:::Fn are n features extracted for a given image. Sample Texture measures of mammogram images are given below:

Image Samples	Average intensity	Average contrast	Smoothness	Third moment	Uniformity	Entropy
Mam 1	39.6760	42.8696	0.0275	0.6056	0.1663	4.7401
Mam 2	47.9076	1.9005	0.0736	6.2341	0.1910	4.6683
Mam 3	43.7049	46.3144	0.0319	0.4708	0.2156	4.4888
Mam 4	43.3234	40.3894	0.0245	0.2425	0.1030	5.4656
Mam 5	43.3946	40.4359	0.0245	0.2419	0.1036	5.4638
Mam 6	62.3899	68.4661	0.0672	2.1793	0.2332	3.2310
Mam 7	68.0774	71.3436	0.0726	1.6967	0.2472	3.0586
Mam 8	61.9692	74.2953	0.0782	3.7407	0.2058	4.9878
Mam 9	55.0435	81.8304	0.0934	8.8683	0.2557	4.4263
Mam 10	43.1755	69.3156	0.0688	6.1621	0.3507	3.9049

Step-IV: Association Rule Mining: Discovering frequent item sets is the key process in association rule mining.

In order to perform data mining association rule algorithm, numerical attributes should be discretized first, i.e. continuous attribute values should be divided into multiple segments. Traditional association rule algorithms adopt an iterative method to discovery, which requires very large calculations and a complicated transaction process. Because of this, a new association rule algorithm is proposed in this paper. This new algorithm adopts a Boolean vector method to discovering frequent item sets.

In general, the new association rule algorithm consists of four phases as follows:

1. Transforming the transaction database into the Boolean matrix.
2. Generating the set of frequent 1-itemsets L1.
3. Pruning the Boolean matrix.
4. Generating the set of frequent k-item sets Lk(k>1).

The detailed algorithm, phase by phase, is presented below:

1. Transforming the transaction database into the Boolean matrix: The mined transaction database is D, with D having m transactions and n items. Let T={T1,T2,…,Tm} be the set of transactions and I={I1,I2,…,In}be the set of items. We set up a Boolean matrix Am*n, which has m rows and n columns. Scanning the transaction database D, we use a binning procedure to convert each real valued feature into a set of binary features.

   The 0 to 1 range for each feature is uniformly divided into k bins, and each of k binary features record whether the feature lies within corresponding range.

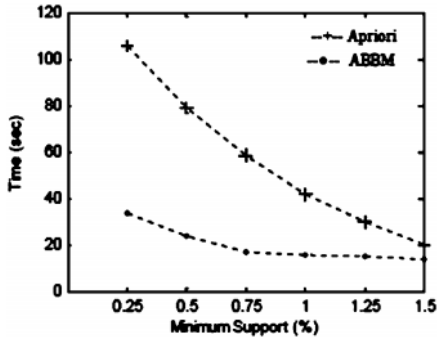
2. Generating the set of frequent 1-itemset L1: The Boolean matrix Am*n is scanned and support numbers of all items are computed. The support number Ij.supth of item Ij is the number of '1s' in the jth column of the Boolean matrix Am*n. If Ij.supth is smaller than the minimum support number, itemset {Ij} is not a frequent 1-itemset and the jth column of the Boolean matrix Am*n will be deleted from Am*n. Otherwise itemset {Ij} is the frequent 1-itemset and is added to the set of frequent 1-itemset L1. The sum of the element values of each row is recomputed, and the rows whose sum of element values is smaller than 2 are deleted from this matrix.

3. Pruning the Boolean matrix: Pruning the Boolean matrix means deleting some rows and columns from it. First, the column of the Boolean matrix is pruned according to Proposition 2. This is described in detail as: Let I• be the set of all items in the frequent set LK-1, where k>2. Compute all |LK-1(j)| where j belongs to I2, and delete the column of correspondence item j if |LK – 1(j)| is smaller than k – 1. Second, recompute the sum of the element values in each row in the Boolean matrix. The rows of the Boolean matrix whose sum of element values is smaller than k are deleted from this matrix.

4. Generating the set of frequent k-itemsets Lk: Frequent k-item sets are discovered only by "and" relational calculus, which is carried out for the k-vectors combination. If the Boolean matrix Ap*q has q columns where $2 < q \leq n$ and minsupth $\leq p \leq$ m, k q c, combinations of k-vectors will be produced. The 'and' relational calculus is for each combination of k-vectors. If the sum of element values in the "and" calculation result is not smaller than the minimum support number minsupth, the k-itemsets corresponding to this combination of k-vectors are the frequent k-itemsets and are added to the set of frequent k-itemsets Lk.

3. EXPERIMENTAL RESULTS

In order to appraise the performance of the new association rule mining algorithm, we conducted an experiment using the Apriori algorithm and the proposed algorithm. The algorithms were implemented in C Here presents the experimental results for different numbers of minimum supports. The results show that the performance of the new association rule mining algorithm is much better than that of the Apriori algorithm. Moreover, the better the performance efficiency of new association rule mining algorithm is, the smaller the minimum support is. This is because the smaller the minimum support, the more candidate item sets the Apriori algorithm has to determine, and also the Apriori algorithm's join and pruning processes take more time to execute. However, the new association rule mining algorithm does not produce candidate item sets, and it spends less time calculating k-supports with the Boolean matrix pruned.



*ABBM : Algorithm Based on Boolean Matrix.

Major steps to improve the performance of the new method for association rule mining:

- Adding more robust features, which are capable of generalizing more effectively? This can reduce a lot of the inaccuracies in the detection process. We intend to look into more image detection features to get more generalized view of the images. This would help us in detection of different types of association rules.

  The transactional database is constructed by merging some already existing features in the original database with some new visual content features that we extracted from the images using image processing techniques. The existing features are:

  – The type of the tissue (dense, fatty and fatty glandular);

  – The position of the breast: left or right.

  The transactions are of the form [Image ID, Class Label, F1; F2; :::; Fn] where F1:::Fn are n features extracted for a given image. The type of tissue is an important feature to be added to the feature database, being well known the fact that for some types of tissue the recognition is more difficult than for others. Method with these features incorporated could increase the accuracy rate

- This project is a part of an important Data-Mining project We can show in this report that Association rule mining does help us in reducing the load on the experts to manually go through these images We intend to build an automated system, which would to a large extent automatically detect association rules from these images. The end-system would independently for most of the prediction process.

- We need a systematic approach to determine an optimal similarity threshold for support & confidence or at least a close one. A very high threshold means only perfect matches are accepted. Finding the right similarity threshold for each image type looks like an interesting problem. Right now it is provided by the user but it can be changed to be tuned by the algorithm itself.

4. CONCLUSION

In this paper, an new method for association rule mining is proposed. The main features of this method are that it only scans the transaction database once, it does not produce candidate itemsets, and it adopts the Boolean vector "relational calculus" to discover frequent itemsets. In addition, it stores all transaction data in binary form, so it needs less memory space and can be applied to mining large databases.

BIBLIOGRAPHY

[1] R. Agrawal, T. Imielinski, and A. Swami, "Mining Association Rules between Sets of Items in Large Databases", In Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, Pages 207–216, Washington, DC, May 26-28 1993.

[2] Agrawal, R., Imielinski, T., & Swami, A. (1993), "Mining Association Rules between Sets of Items in Large Databases", Proceedings of the ACM SICMOD Conference on Management of Data, pp. 207-216, Washington, D.C.

[3] Han, J., Pei, J., & Yin, Y (2000), "Mining Frequent Patterns Candidate Generation" In Proc. 2000 ACM-SIGMOD Int. Management of Data (SIGMOD'00), Dallas, TX.

[4] Berzal, F., Blanco, I., Sánchez, D. and Vila, M.A. "Measuring the Accuracy and Importance of Association Rules: A New Framework" Intelligent Data Analysis, 6:221-235, 2002.

[5] David A. Clausi, "An Analysis of Co-occurrence Texture Statistics as a Function of Gray Level Quantization", Can. J. Remote Sensing, 28, No. 1, pp. 45-62, 2002.

[6] Bodon, F. "A Fast Apriori Implementation", In Proc. IEEE ICDM Workshop on Frequent Item set Mining Implementations, 2003.

[7] Brijs, T. Vanhoof, K. and Wets, G., "Defining Interestingness for Association Rules", In Int. Journal of Information Theories and Applications, 10:4, 2003.

[8] Tung, A., Lu, H., Han, J., & Feng, L. (2003), "Efficient Mining of Intertransaction Association Rules", IEEE Transaction on Knowledge and Data Engineering, 15(1), 43-56.

[9] Xu, Z. & Zhang, S. (2003), "An Optimization Algorithm Base on Apriori for Association Rules", ComputerEngineering, 29(19), 83-84.

[10] 4th European Conference of the International Federation for Medical and Biological Engineering ECIFMBE 2008 23–27 November 2008 Antwerp, Belgium, 10.1007/978-3-540-89208-3_144, Jos Vander Sloten, Pascal Verdonck, Marc Nyssen and Jens Haueisen.

[11] Rabi Narayan Panda, Dr. Bijay Ketan Panigrahi, Dr. Manas Ranjan Patro, "Feature Extraction for Classification of Micro calcifications and Mass Lesions in Mammograms", IJCSNS International Journal of Computer Science and Network Security, 9, No.5, May 2009.

[12] J. Han, M. Kamber (2001), Data Mining, Morgan Kaufmann Publishers, San Francisco, CA.

[13] R. C. Gonzalez and R. E..Woods, "Digital Image Processing", Second Edition 2002.

[14] R. C. Gonzalez, "Digital Image Processing using Matlab" Pearson Publication, 2005.

[15] "Image Processing The Fundamentals" Maria Petrou University of SurreN Guildford, UK. Panagiota Bosdogianni Technical Universify of Crete, Chania, Greece John Wiley & Sons, LTD.