Proceedings of the International Multiconference on
Computer Science and Information Technology pp. 85–91

ISSN 1896-7094

Image Classification with Customized Associative Classifiers*

Łukasz Kobyliński and Krzysztof Walczak

Institute of Computer Science, Warsaw University of Technology
ul. Nowowiejska 15/19, 00-665 Warszawa, Poland

Abstract. In recent years the concept of utilizing association rules for classification emerged. This approach proved to often be more efficient and accurate than traditional techniques. In this paper we extend existing associative classifier building algorithms and apply them to the problem of image classification. We describe a set of photographs with features calculated on the basis of their color and texture characteristics and experiment with different types of rules, which make use of the information about the existence of a particular feature on an image, its occurrence count and spatial proximity to accurately classify the images. We suggest using association rules more closely tied to the nature of image data and compare the results of classification with simple rules, taking into consideration only the existence of a particular feature on an image.

Keywords: image, classification, class association rules, associative classifiers

1 Introduction

The large volume of multimedia data that is being collected every day, particularly in professional fields, such as satellite and aerial imagery in land planning, agriculture or forestry, but also by anyone owning a digital camera, brings us to the problem of extracting meaningful information from such collections of raw image data without the need of human intervention. Classification is an important part of any knowledge-retrieval system and especially significant in these applications, where images are the main source of information in the decision making process.

In this paper we propose an application of extended Class Association Rules (CARs) to image classification. The approach is suitable for analyzing large sets of photographs, as it derives from data mining techniques developed with such databases in mind. As we have shown earlier by experimenting with a similar method [4], it gives particularly good results in classification of images obtained from remote-sensing methods, but may also be used for analysis of general sets of photographs as well. We extend the concept of association rules with recurrent items for classification with information specific to image analysis, such as the number of occurrences of a particular feature on the image or the maximum size of a region with uniform feature characteristics. A modified version of the CBA algorithm is used to mine such rules from a training set of images, described by their color and texture features. The rules have the form of an implication between a limited number of features with appropriate weights and a category label. They are pruned and used to create a classifier suitable for efficient classification of unseen examples.

The employed method of classification is a two-stage process, in which the classifier is built on the basis of a training set and used to associate category labels with previously unseen examples of images. At first a symbolic representation of the images is created to enable the use of data mining methods, by calculating their color and texture features and clustering them into a structure of a dictionary of representative values. The classifier is created on the basis of a reduced set of discovered rules. New photographs are processed in exactly the same way as the training ones, without the extra dictionary building and rule mining steps. The existing dictionary is used to label particular blocks of the images with identifiers of the dictionary entries and rules from the classifier are applied to classify the photographs into categories.

The rest of the paper is organized as follows: Section 2 presents previous work related to the subject of CARs and association rule mining in image databases. In Section 3 we give the details of our approach to image representation, which is then used in classification process. In Section 4 we describe the concept of associative classification and propose extended association rule mining and classifier building algorithms. Section 5 presents experimental results of image classification and Section 6 closes with a conclusion and discussion on possible enhancements.

* The research has been partially supported by grant No 3 T11C 002 29 received from Polish Ministry of Education and Science.

2 Previous Work

While the concept of mining association rules for classification was first proposed in [3], the first classifier building algorithm CBA was introduced in [6], followed by CMAR [5] and ARC [12]. The idea of including recurrent items in association rules was presented as a modification of the Apriori algorithm in [13] and the FP-growth algorithm in [8]. Finally, the possibility of incorporating recurrent items into CARs was presented in [9] by a modification of the ARC-BC algorithm.

Recent data mining applications to image databases considered the classification of mammograms [1], mining association rules between regions of paintings [2] or features of aerial images, including their spatial relationship [10].

A similar image representation approach to the method presented here was proposed in [14], where the authors compared different representational models, which extract image features from its individual blocks.

In [4] we have shown that achieving good classification results of aerial photographs is possible even with simple association rules, having a class label in the consequent. Here we elaborate on this subject and compare the results with a classifier more closely related to the problem of image classification.

3 Image Representation

A preliminary step of creating a symbolic representation of the source images is required before applying any data mining methods to the database. The images are thus firstly normalized by bringing them to a common resolution and performing histogram equalization. Secondly, they are divided into a grid of 32×32 pixels blocks. Lastly, color and texture features of each of the blocks are calculated, to be used in further processing. This initial procedure may be performed before the actual classification process, for example while adding a new photograph to the database.

An additional step of creating a dictionary of typical feature values is necessary before training a new classifier. This is performed by clustering the values to find a chosen number of group centroids, which then become the elements of a dictionary. Individual blocks of the images are then labeled with identifiers of the most similar entries present in the dictionary. The representation of a particular image consists of a list of all identifiers associated with its blocks.

3.1 Calculating Color Features

Color features are represented by a histogram calculated in the HSV color space, with the H channel quantized to 18 values and S and V channels to 3 values each. In effect the representation takes the form of a 162-element vector of real values between 0 and 1. Histogram intersection measure is used to compare two feature vectors h and g:

$$d_I(h,g) = 1 - \sum_{i=0}^{N-1} \min(h[i], g[i]). \tag{1}$$

3.2 Calculating Texture Features

A statistical approach presented in [7], which utilizes Gabor filtering is used to represent the important information about the texture visible on the photographs. The feature vector consists of mean and standard deviation values calculated from images resulting from filtering the original pixels with a bank of Gabor functions. These filters are scaled and rotated versions of the base function, which is given by the formula:

$$g(x,y) = \left(\frac{1}{2\pi\sigma_x\sigma_y}\right) exp\left[-\frac{1}{2}\left(\frac{x^2}{\sigma_x^2} + \frac{y^2}{\sigma_y^2}\right) + 2\pi jWx\right]. \tag{2}$$

Six different orientations and four scales of the base function are used to filter every photograph. The images resulting from consecutive filterings take thus the form of:

$$W_{mn}(x,y) = |I(x,y) * g_{mn}(x,y)|, \tag{3}$$

where $*$ denotes spatial convolution, m filter orientation and n scale. The final feature vector consisting of mean and standard deviation values takes the form of:

$$\overline{f} = [\mu_{00}\sigma_{00} \cdots \mu_{M-1N-1}\sigma_{M-1N-1}]. \tag{4}$$

Comparing two feature vectors $\overline{f}^{(i)}$ and $\overline{f}^{(j)}$ is accomplished by the distance measure given below:

$$d(i,j) = \sum_{m}\sum_{n} d_{mn}(i,j), \tag{5}$$

$$d_{mn}(i,j) = \left|\frac{\mu_{mn}^{(i)} - \mu_{mn}^{(j)}}{\alpha(\mu_{mn})}\right| + \left|\frac{\sigma_{mn}^{(i)} - \sigma_{mn}^{(j)}}{\alpha(\sigma_{mn})}\right|, \tag{6}$$

where $\alpha(\mu_{mn})$ and $\alpha(\sigma_{mn})$ are the standard deviations of the values over entire database.

3.3 Creating the Feature Dictionary

The dictionary consists of the most typical color and texture features of the individual blocks of photographs in the training set. It is created by clustering corresponding feature values into a chosen number of groups. The clustering is performed using a k-Means algorithm with the histogram intersection measure for comparing color features and Gabor feature distance for comparing texture features.

Centroids resulting from the clustering operation become the elements of the dictionary and are labeled with consecutive natural numbers. These identifiers are then used to describe blocks of the images in the database. During the classification phase the previously created dictionary is queried with color and texture feature values and responds with the labels of the most similar entries.

4 Associative Classification

Associative classifiers are a recent, two-stage approach to classification, in which a set of association rules between the attribute values and category labels is first discovered and then a compact classifier is created by selecting the most important rules for classification.

4.1 Association Rules for Classification

Formally, an association rule used for classification is an implication of the form of $X \rightarrow c$, where itemset X is a non-empty subset of all possible items in the database, $X \subseteq I$, $I = \{i_1, i_2, \ldots, i_n\}$, and c is a class identifier, $c \in \{c_1, c_2, \ldots, c_n\}$. Let thus a rule itemset be a pair $< X, c >$, containing the itemset X and a class label c. The rules are discovered in a training set of transactions D_t. Each transaction is a triple of the form $< tid, Y, c >$, containing a transaction identifier tid, itemset $Y \subseteq I$ and a class label c.

In our approach we discover the most interesting association rules between the images in the training set, described by dictionary entries, and their category labels. This is a slight modification of the classic association rule mining problem, as the implication consequent is always limited to a class label. The aim of the mining is then to discover the rules that are a subset of general association rules and have the following form:

$$R_c : \{color_1, \ldots, color_n,\ texture_1, \ldots, texture_m\} \Rightarrow class\ label \tag{7}$$

We adapt the existing methods of association rule mining to create a classifier suitable for categorization of image data. Direct application of any rule mining algorithm to a transactional database containing images represented by feature values in their particular locations would result in a large number of irrelevant associations. Following this observation, we consider only the existence, occurrence count and spatial proximity of features to create rules that are sufficiently general to classify previously unseen examples.

The initial set of discovered rules is usually very large, so it is necessary to limit the number of associations by specifying the minimum support and confidence values and employing various pruning techniques. We use the CBA approach proposed in [6] to mine the rules along with frequent itemsets and then apply a pruning strategy to limit their number.

4.2 Considering Occurrence Count

Extending association rules to include the information about item occurrence count in multimedia applications was first proposed in [13]. We use this general idea to mine classification rules with recurrent items and apply a selected number of such associations to the problem of image classification. A slight modification of calculating the support of such rules is necessary, as a single transaction may increase the support of an itemset by more than one. The support of an itemset X may thus be calculated as [8]:

$$\mathrm{supp}(X) = \sum_{k=0}^{|D|} \frac{\phi(X, t_k)}{|D|}, \tag{8}$$

where ϕ is a function that returns the ratio by which a transaction t_k of a database D supports itemset X and is defined as:

$$\phi(X, t) = \min\left(\frac{\alpha_j}{\beta_j}\right), j = 1 \ldots n, \tag{9}$$

$t_k = \{\alpha_1 i_1, \alpha_2 i_2, \ldots, \alpha_n i_n\}$, $X = \{\beta_1 i_1, \beta_2 i_2, \ldots, \beta_n i_n\}$, $\alpha_i \neq 0, \beta_i \neq 0$.

The support of a rule with recurrent items is calculated similarly as when considering simple association rules, by counting the support of a set consisting of both the rule's antecedent and consequent. The definition of confidence also remains unchanged and may be calculated as $\mathrm{supp}(X \cup Y)/\mathrm{supp}(X)$. The definition of a frequent itemset may be extended by including an additional condition of maximum support $\mathrm{supp}(X) < \Sigma$, apart from its minimum value $\mathrm{supp}(X) > \sigma$, which helps to minimize the number of uninteresting rules.

A modified version of the CBA algorithm, presented as Algorithm 1, is used to mine all possible rules or only the rules having a certain maximum number of items in the antecedent.

Algorithm 1 CBA-RG with recurrent items

Input D_t (training set), σ (min. support), Σ (max support), δ (min. confidence)
Output CAR (class association rules with recurrent items)

1: $F_1 \leftarrow$ {frequent 1 rule itemsets}
2: $M \leftarrow$ {maximum occurrence of frequent 1 itemsets in D_t}
3: $CAR_1 \leftarrow \{f \in F_1 \mid \mathrm{supp}_{rule}(f) < \Sigma \wedge \mathrm{conf}(f) > \delta\}$
4: $k \leftarrow 1$
5: **while** $F_k \neq \emptyset$ **do**
6:   $C_{k+1} \leftarrow (F_k \otimes F_k) \cup \{f \in F_k \oplus x \in F_1 \mid count(x, f) < M[x]\}$
7:   **for all** $t \in D$ **do**
8:     **for all** $c \in C_{k+1}$ **do**
9:       $\mathrm{supp}_X(c) = \mathrm{supp}_X(c) + \phi(c, t)$
10:       $\mathrm{supp}_{rule}(c) = \mathrm{supp}_{rule}(c) + \phi(c, t) \mid class(c) = class(t)$
11:     **end for**
12:   **end for**
13:   $F_{k+1} \leftarrow \{c \in C_{k+1} \mid \mathrm{supp}_{rule}(c) > \sigma\}$
14:   $CAR_{k+1} \leftarrow \{f \in F_{k+1} \mid \mathrm{supp}_{rule}(f) < \Sigma \wedge \mathrm{conf}(f) > \delta\}$
15: **end while**
16: **return** $\bigcup_k CAR_k$

In the lines 1–3 a first pass over the database is made to find all sufficiently frequent itemsets, which can be used to build rules with a single value in the antecedent. The maximum number of occurrences of every item in the transactions of the database is also counted, similarly as in the MaxOccur algorithm [13], to limit the number of item recurrences while generating candidates. Line 6 generates candidates using the Apriori method and includes another occurrence of an existing item, whenever existing count is below the maximum value. The *count* function returns the item's current number of occurrences in an itemset, while the $\otimes$ and $\oplus$ symbols denote itemset merging and item concatenation operations respectively. Lines 7–12 are used to independently calculate the support of each rule and the support of its antecedent. These values are then used to calculate confidence of the rules in line 14.

4.3 Considering Spatial Proximity

Apart from the association rules between the number of particular features present on the images and their category labels we also consider rules that include the information about spatial proximity of the features. While mining for the association rules we check for spatial relationships between the recurring features and only then include them multiple times in the rules, when they form a single area on the image.

It is possible to mine such rules without any change to the abovementioned algorithm by slightly changing the representation of the images. For every element of the dictionary each transaction is scanned to find the largest area covered by a single feature. The original number of occurrences of every item is then reduced to that maximum value before the association rule mining algorithm is applied. Table 1 illustrates the difference between both approaches to image representation.

Table 1. An example of image representation when considering spatial proximity of features

B_1, T_1	B_1, T_2	B_2, T_1	B_1, T_2
B_2, T_1	B_1, T_2	B_2, T_2	B_1, T_1
B_1, T_2	B_2, T_1	B_1, T_1	B_1, T_2
B_2, T_2	B_2, T_1	B_1, T_1	B_2, T_2

$\rightarrow$

Direct representation:
$9B_1$, $7B_2$, $8T_1$, $8T_2$
Considering spatial proximity of features:
$5B_1$, $3B_2$, $4T_1$, $3T_2$

4.4 Building the Classifier

Having found all the rules with minimum and maximum support, as well as the minimum confidence, we come to the problem of creating a classifier which will then be used to associate category labels with previously unseen images. The final classifier is created by first sorting the rules according to their confidence and support in descending order and the number of items in their antecedents in ascending order. Next, for every rule in the sorted list, all elements of the training set matching that rule are found and removed from further processing. A rule is then added to the classifier if it matches at least one element of the set. At each step of the iteration a default class of the classifier is selected that minimizes the error of classification of the remaining data. Lastly, when the rule or data set is empty, the final classifier is reduced to the first number of rules that decrease the general error rate of classification.

4.5 Classification

Classification is performed by applying the first matching rule from the classifier to a given image described by dictionary entries. A default class label is then given to an image for which there are no matching rules. An image matches a rule when it contains each of the items of the rule's antecedent with at least the same occurrence count.

Table 2 shows an example representation of a few photographs without considering spatial relationships between the features, a possible classifier content and classification result. The classifier was created using the CBA approach to limit the number of rules. Dictionary entries are identified by B_i (color) and T_i (texture) labels. The dictionary size was 8 entries for color and 8 for texture in this example. The first two images are matched by the first and second rule of the classifier respectively and are associated with C_1 category label. The last two images are classified using the $default\_class$ value, as they remain unmatched by any rule.

Table 2. An example image representation, a classifier content and the classification results

I	Image Features
I_1	$7B_2, 38B_3, 51B_4, 88T_1, 7T_2, 1T_3$
I_2	$2B_1, 3B_2, 23B_3, 68B_4, 65T_2, 15T_3, 6T_4$
I_3	$23B_1, 72B_2, 1B_3, 4T_1, 57T_2, 21T_3, 14T_4$
I_4	$48B_1, 14B_2, 34B_3, 1T_1, 60T_2, 24T_3, 11T_4$

Rules
$2B_3, 1B_4, 1T_1 \Rightarrow C_1$
$1B_2, 3B_3, 1B_4 \Rightarrow C_1$
$default\_class = C_0$

I	Class
I_1	C_1
I_2	C_1
I_3	C_0
I_4	C_0

5 Experimental Results

We have verified the results of image classification of the proposed method on the test dataset made available by the authors of the SIMPLIcity CBIR system [11]. We have chosen 400 photographs, having a resolution of 384 × 256 pixels and associated with four different category labels, namely *buses*, *flowers*, *horses* and *mountains*. The accuracy of both the approach described above and the method of classification with simple class association rules proposed earlier in [4] has been compared by performing classification of the same set of images belonging to two different categories.

The results of the experiments are presented in Table 3. For each dictionary size k (the number of different color and different texture entries) classification accuracy for both the methods is shown. The first experiment, described as Exp. 1, has considered classification between *horse* and *flower*, the second *bus* and *mountain* and the third between *bus* and *horse* sets of photographs. We have used ten-fold cross-validation to reduce the influence of any random factors. The rules have been mined with minimum support of 0.01, minimum confidence 0.50, maximum support 1.00 and the antecedent length limited to 5 items.

Figure 1 presents the relationship between the dictionary size, chosen method of classification and the number of rules found and included in the classifier. It is clear that including recurrent items when discovering the rules significantly increases their number. Considering spatial proximity of features helps to reduce the number of found associations, as well as the number of rules in the classifier.

The presented results prove that extending class association rules with recurrence of items and information about spatial proximity of features may improve classification accuracy of photographs, represented according to our approach. Most of the results obtained using the newly proposed method were of a higher accuracy than the one presented earlier. Other authors have proved that associative classifiers are better than C4.5 and similar methods (experiments have been conducted on a set of 34 benchmark problems from UCI machine learning repository), so we did not perform another comparison of the classifiers by themselves.

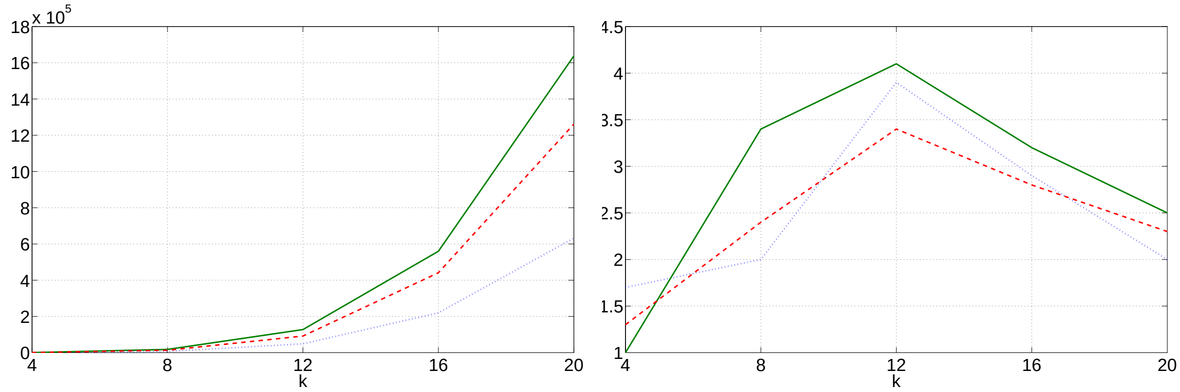
While the association rules with recurrent objects may be thought of as a generalization of simple rules with binary information about item existence, the problem of selecting the most effective ones for classification remains. That is why not every experiment turned out to give better results when considering the extended rules. There are cases for which the used method of rule selection produces better results with simple class association rules. Considering spatial proximity of the features present on an image does not seem to further increase the classification accuracy considerably, but helps to limit the number of discovered associations.

Table 3. Classification accuracy of the four test datasets

k	Simple rules			With recurrence			With spatial proximity		
	Exp. 1	Exp. 2	Exp. 3	Exp. 1	Exp. 2	Exp. 3	Exp. 1	Exp. 2	Exp. 3
4	93.02	87.73	93.26	94.12	91.74	94.24	93.35	89.86	93.86
8	90,70	94.48	99.44	93.48	93.27	98.96	91.32	94.31	99.17
12	96,51	95.70	96.07	97.53	96.48	97.45	95.85	96.35	96.24
16	95,93	95.70	97.75	95.26	95.19	96.89	95.32	96.18	97.89
20	94.77	93.25	98.32	94.34	94.24	98.43	94.80	94.11	98.39

6 Conclusions

In this paper we have proposed an extension of associative classifiers with recurrent items and experimented with the application of association rules to classification of photographs. We have used class association rules with recurrent items and considered the spatial proximity of the features of a particular image to accurately classify a set of photographs. We have applied this method to a dataset containing photographs associated with four different categories and presented results of their classification. The described approach has proved to perform better than the previously tested classifier utilizing only simple rules with no item occurrence information. Associative classification of images is a promising area of research, as many different approaches to image representation and association rule mining and pruning may be proposed to improve accuracy of the process.



(a) (b)

Fig. 1. Influence of the dictionary size on the number of rules. Solid line corresponds with the classifier with recurrent items, dashed line with the classifier considering spatial proximity of features and the dotted line with the classifier utilizing simple rules. (**a**) The number of discovered rules in respect to the size of the dictionary. (**b**) The number of rules in the classifier in respect to the size of the dictionary.

References

1. Antonie M.-L., Zaïane O. R., Coman A.: *Associative classifiers for medical images,* Revised Papers from MDM/KDD and PAKDD/KDMCD, (2002) 68–83.
2. Ardizzone E., Daurel T., Maniscalco U., Rigotti C.: *Extraction of association rules between low-level descriptors and semantic descriptors in an image database,* In Proc. 1st Int. Workshop on Multimedia Data and Document Eng., (2001).
3. Bayardo R.: *Brute-force mining of high-confidence classification rules,* In Proc. 3rd Int. Conf. on Knowledge Discovery and Data Mining, (1997) 123–126.
4. Kobyliński Ł., Walczak K.: *Data mining approach to classification of archaeological aerial photographs,* In Proc. Int. IIS: Intelligent Information Processing and Web Mining Conf. Advances in Soft Computing, Springer, (2006), 479–487.
5. Li W., Han J., Pei J.: *CMAR: Accurate and efficient classification based on multiple class-association rules,* IEEE International Conference on Data Mining, (2001).
6. Liu B., Hsu W., Ma Y.: *Integrating classification and association rule mining,* In Proc. 4th Int. Conf. on Knowledge Discovery and Data Mining, (1998) 80–86.
7. Manjunath B. S., Ma W.: *Texture features for browsing and retrieval of image data,* IEEE Trans. on Patt. Anal. and Machine Intell., **18**, (1996) 837–842.
8. Ong K.-L., Ng W.-K., Lim E.-P.: *Mining multi-level rules with recurrent items using FP'-Tree,* In Proc. 3rd ICICS, (2001).
9. Rak R., Stach W., Zaïane, Antonie M-L.: *Considering re-occurring features in associative classifiers,* In Proc. 9th Pacific-Asia Conf. on Knowledge Discovery and Data Mining, (2005) 240–248.
10. Tešić J., Newsam S., Manjunath B. S.: *Mining image datasets using perceptual association rules,* In Proc. SIAM International Conference on Data Mining, 6th Workshop on Mining Scientific and Engineering Datasets, (2003) 71–77.
11. Wang J. Z., Li J., Wiederhold G.: *SIMPLIcity: Semantics-sensitive integrated matching for picture libraries,* IEEE Trans. on Patt. Anal. and Machine Intell., **23**, (2001) 947–963.
12. Zaïane O. R., Antonie M.-L.: *Classifying text documents by associating terms with text categories,* In Proc. 13th Australasian Database Conference, (2002) 215–222.
13. Zaïane O. R., Han J., Zhu H.: *Mining recurrent items in multimedia with progressive resolution refinement,* In Proc. 16th Int. Conf. on Data Eng. (2000), 461–470.
14. Zhu L., Rao A., Zhang A.: *Theory of keyblock-based image retrieval,* ACM Trans. Inf. Syst., **20**, (2002) 224–257.