Applying Associative Classifier PGN for Digitised Cultural Heritage Resource Discovery

Krassimira Ivanova[1], Iliya Mitov[1], Peter L. Stanchev[1,2]
Milena Dobreva[3], Koen Vanhoof[4] and Benoit Depaire[4]

[1] Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, Sofia, Bulgaria
kivanova@math.bas.bg, mitov@mail.bg, stanchev@math.bas.bg
[2] Kettering University, Flint, USA
[3] Computer and Information Sciences Dept., University of Strathclyde, Glasgow, UK
milena.dobreva@strath.ac.uk
[4] Hasselt University, Hasselt, Belgium
koen.vanhoof@uhasselt.be, benoit.depaire@uhasselt.be

Abstract. Resource discovery is one of the key services in digitised cultural heritage collections. It requires intelligent mining in heterogeneous digital content as well as capabilities in large scale performance; this explains the recent advances in classification methods. Associative classifiers are convenient data mining tools used in the field of cultural heritage, by applying their possibilities to taking into account the specific combinations of the attribute values. Usually, the associative classifiers prioritize the support over the confidence. The proposed classifier PGN questions this common approach and focuses on confidence first by retaining only 100% confidence rules. The classification tasks in the field of cultural heritage usually deal with data sets with many class labels. This variety is caused by the richness of accumulated culture during the centuries. Comparisons of classifier PGN with other classifiers, such as OneR, JRip and J48, show the competitiveness of PGN in recognizing multi-class datasets on collections of masterpieces from different West and East European Fine Art authors and movements.

Keywords: Data Mining, Associative Classifier, Metadata Extraction, Cultural Heritage

1 Introduction

Every touch to artworks builds a bridge between cultures, times and individual personalities. Numerous art and architectural masterpieces have been created over the centuries, and are scattered all over the world. For most people the direct touch to these treasures is impeded by various obstacles. On the other hand, the access to masterpieces is a necessary but not sufficient condition in understanding them because this is a learning process which includes not only the artefact itself but also the context of its creation.

Nowadays online search engines and digital collections have significantly increased the possibilities to consult both the artefacts and their cultural context. Such collections present the colourfulness of art history as well as relevant metadata; provide additional information on purely technical details as well as on more abstract levels ranging from details on artefacts' creation to personal biographical details on their creators. The access to digitised art helps the users to understand the original messages in the masterpieces. However, the unprecedented growth of digital collections and resources requires the development of image retrieval techniques which would aid resource discovery for efficient and high-quality large scale retrieval tasks.

The use of metadata can significantly improve the quality of resource discovery. Metadata help search engines and people to distinguish between relevant from non-relevant objects in the process of resource discovery. However, the human creation of all metadata, especially those describing the content of an object, is a typical bottleneck in the development of digital collections. Addressing this challenge attracts more research in automatic metadata generation. The proposed approaches can be categorized into two major subcategories: *harvesting* and *mining* (*extraction*) of metadata [1].

Harvesting of metadata is the process of automatic extraction of predefined fields. The collection process relies on metadata produced by humans or semi-automatic processes, with appropriate application software. Examples of harvesting are the processes assuring interoperability of metadata from various systems and platforms [2] and extraction of metadata from non-cooperating digital libraries [3]. *Extraction* of metadata occurs when an algorithm automatically extracts metadata from the content of the resource. Sources for the extraction of metadata can be grouped mainly in: *content analysis*, *context analysis*, *usage*, and *composite structure* [4].

Data mining is a part of the overall process of Knowledge Discovery in Databases [5].While knowledge discovery is defined as the process of seeking new knowledge about an application domain [6]. Data mining is the process of analyzing a large set of raw data in order to extract hidden information which can be predicted. It developed into a discipline, which is at the confluence of artificial intelligence, data bases, statistics, natural language processing, and machine learning. Data mining addresses several aspects, the main being: classification, clustering, association and regularities. In addition to the analysis of data from many different dimensions or sides, a key further process is summarizing the relationships identified [7].

Data mining methods are divided mainly in two main types: *verification-oriented* (the system verifies user's hypothesis); and *discovery-oriented* (the system finds new rules and patterns autonomously) [8]. Most of the *discovery-oriented techniques* are based on inductive learning [9], where a model is constructed explicitly or implicitly by generalizing from a sufficient number of training examples. The underlying assumption of the inductive approach is that the trained model is applicable to future unseen examples. The discovery methods branch into *description methods* versus *prediction methods*.

Description-oriented data mining methods focus on understanding the way the underlying data operates. The main approaches exploited are *clustering* (the process of grouping the data, with high similarity within the group, using different kinds of distance measures) *link analysis* (the process of uncovering relationships among data,

such as finding matches in data for known pattern of interest, identifying anomalies where known patterns are violated, or discovering new patterns of interest [10]) and *summarization* (the process of data reducing on the base of *extraction* or *abstraction*).

Prediction-oriented methods aim to build a behavioural model that can get new and undiscovered samples and are able to predict values of one or more variables related to the sample. Two main branches exist: *classification* and *estimation*. These two forms of data analysis are used to extract models describing significant data classes or to predict future data trends. The main difference between classification and estimation is that classification maps the input space into predefined classes, while estimation models the input space into a real-valued domain.

Classification models predict discrete, unordered labels. The classification is the problem of identifying the group to which the query belong, where the identity of the group is unknown, on the basis of a training set of data containing instances whose group is known. There are several big groups of classifiers: Bayesian Methods, Support Vector Machines, Decision Trees, Decision Rules, Class Association Rules, Lazy Learners, Neural Networks, and Genetic Algorithms.

When we have the task to select a method for a particular domain, obviously there is a rich choice. Confronted with such a wide range of options, for the task of access to digitised art images we decided to concentrate on decision trees, decision rules and class association rules

Our attention is focused mainly on the associative classifiers, which generate a set of association rules from a given training set. Various associative classifiers exist, such as the very first one *CBA* [11], *CMAR* [12], *ARC-AC* and *ARC-BC* [13], *CPAR* [14], *CorClass* [15], *ACRI* [16], *TFPC* [17], *HARMONY* [18], *MCAR* [19], *2SARC1* and *2SARC2* [20], *CACA* [21], *ARUBAS* [22], etc.

Usually, the generation of association rules from a training set is guided by the support and confidence metrics. Many associative classifiers set a minimum support level and use the confidence metric to rank the remaining association rules. This approach, with a primary focus on support and confidence as the second criterion, will reject 100% confidence rules if the support is too low.

For the purposes of the experiments presented in this paper we used the associative classifier PGN [23]. It is based on different methodological approach of the standard one, which prioritizes support over confidence. Contrary, PGN focuses on confidence first by retaining only 100% confidence rules. Our assumption was that such approach would be particularly useful in the case of multi-class datasets, which is the case of our test collection.

This paper is organized as follows: Section 2 makes a brief overview of the proposed associative classifier PGN; Section 3 presents experimental results and comparison of PGN with other classifiers, such as OneR, JRip and J48, showing the competitiveness of the used approach in PGN for recognizing multi-class datasets on the example of a collection of masterpieces from different West and East European Fine Art authors and movements. Finally, in the conclusion steps for future development are highlighted.

2 Associative Classifier PGN

Here we present a summary of the main steps of the algorithm of the associative classifier PGN; it is described in more details in [24].

2.1 Learning

The training process consists of *generalization* (the process of associative rule mining), following by *pruning* (the process of clearing exceptions between classes and lightening the pattern set). For each class, a separate set of association rules is generated.

The generalization consists of two phases:

1. Adding instances to the sub-set in the pattern set, correspondingly to their class-labels.
2. Creating all possible intersection patterns between patterns within the class.

In the pruning step some patterns are removed from the pattern set:

1. Deleting all contradictory patterns as well as general patterns that have exception patterns in some other class. This step tries to supply the maximum confidence of the resulting rules.
2. Removing more concrete patterns within the classes. This step ensures compactness of the pattern set that can be used in the recognition stage.

As a result in the pattern set remain only patterns that are general for the class that they belong to and their bodies are not subsets of the bodies of patterns in other classes.

2.2 Classification

The record to be recognized is given by the values of its attributes $Q = (?\,|\, a_1, a_2, ..., a_n)$. Some of the features may be omitted.

To classify new instances with the pruned rule set, the definition for the size of an association rule must be introduced first. The association rule size corresponds to the number of non-class attributes which have a non-missing value: $|P| = |\{a_i \,|\, 1 \le i \le n-1, a_i \ne "-"\}|$. The intersection percentage between a pattern P and a query Q is defined as $IP(P,Q) = \frac{|P \cap Q|}{|P|}$.

To classify a new instance, the intersection percentage between the test case and every rule is calculated. This allows for two different scenarios:

- when the maximum intersection percentage occurs only in one class (for only one single rule or for different rules but in the same class), this class becomes the predicted class for the new instance;
- when the maximum intersection percentage occurs multiple times for rules from different classes, the supports of these rules are summed per class. The

class with the highest aggregated support becomes the predicted class for the new instance.

Note that this classification scheme also uses association rules which do not cover the test case perfectly for classification purposes.

The experiments made in [24] demonstrated that PGN shows very good results in comparison with classifiers with similar classification models, such as J48 (representative of Decision Trees) and JRip (representative for Decision Rules) [25], usually receiving bigger accuracy. The possibility to take into account the combinations between attributes leads to significantly outperforming of OneR [26], which chooses the most informative single attribute for each class-label and bases the rule on this attribute alone. PGN shows very good behaviour especially in the case of multi-class datasets [24].

3 Classification Results on the Example of a Digital European Fine Art Collection

For this study, we made an experiment over a dataset that included visual features, extracted by 600 paintings of 19 artists from different movements of West-European fine arts and Eastern Medieval Culture [27]. The pictures were obtained from different web-museums sources using ArtCyclopedia as an entry point to museum-quality fine art on the Internet (Table 1).

Table 1. List of the artists, which paintings were used in experiments, grouped by movements

Movement	Artist
Icons (60)	Icons (60)
Renaissance (90)	Botticelli (30); Michelangelo (30); Raphael (30)
Baroque (90)	Caravaggio (30); Rembrandt (30); Rubens (30)
Romanticism (90)	Friedrich (30); Goya (30); Turner (30)
Impressionism (90)	Monet (30); Pissarro (30); Sisley (30)
Cubism (90)	Braque (30); Gris (30); Leger (30)
Modern Art (90)	Klimt (30); Miro (30); Mucha (30)

The visual features were constructed as follows: The pixels in the images are converted into the HSL color model. The quantization of Hue is made to 13 bins, $ih = -1,...,NH-1$, $NH = 12$, where one value is used for achromatic colors ($ih = -1$) and twelve hues are used for fundamental colors ($ih = 0,...,NH-1$). The quantization function is non-linear with respect to taking into account the misplacement of artists' color wheel and Hue definition in HSL color space. The quantization intervals are given in Figure 1. The saturation and lightness are linearly quantized into NS-bins ($is = 0,...,NS-1$), respectively NL-bins ($il = 0,...,NL-1$). We have used $NS = 10$ and $NL = 10$.

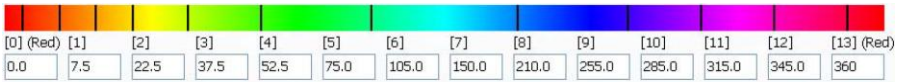

Fig.1. Quantization of Hue

The visual features are used to classify movements and artists styles. We made three-fold cross validation using the datasets that contains hue values, saturation values, luminance values separately and all three together. We analyzed the results of OneR, JRip, J48, and PGN, comparing average accuracies and confusion matrices.

Table 2 and Figure 2 show the accuracies by different classifiers by distribution of hue, saturation, luminance separately and all three together.

As expected the accuracies obtained by the classifiers based on one colour component are similar and we have an increase in accuracy by combining the components. The table shows however a curiosity. As we can see, examining all attributes together does not increase the accuracy of the OneR classifier for movements. In the three fold-cases for "HSL" dataset OneR choose "v0" attribute as most appropriate, but not "s7" or "s8" as in the case of "Saturation" dataset, it leads to decreasing of overall accuracy in HSL dataset than in simpler one "Saturation" dataset.

Table 2. Accuracies for visual features; movements as class label

Database	OneR	JRip	J48	PGN
Hue	27.83	34.00	39.00	42.83
saturation	34.83	33.00	35.33	36.50
luminance	30.67	35.00	38.50	45.83
HSL	33.50	49.00	47.00	63.17

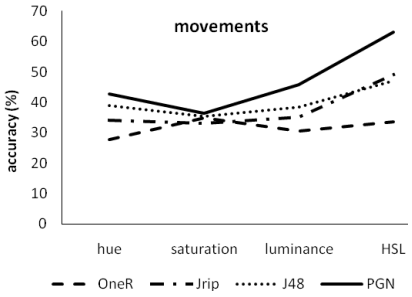

Fig.2. The accuracies of different classifiers by hue, saturation, luminance separately and all three together

As we can see PGN shows the best accuracies from examined models for all datasets. Additionally PGN shows the best possibilities to explore specific combinations of attribute values; it achieves the biggest increase of accuracy by examining all three characteristics together.

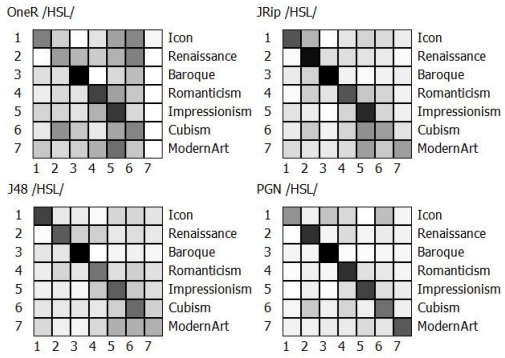

Fig.3. Confusion matrices for HSL features, movements as class labels

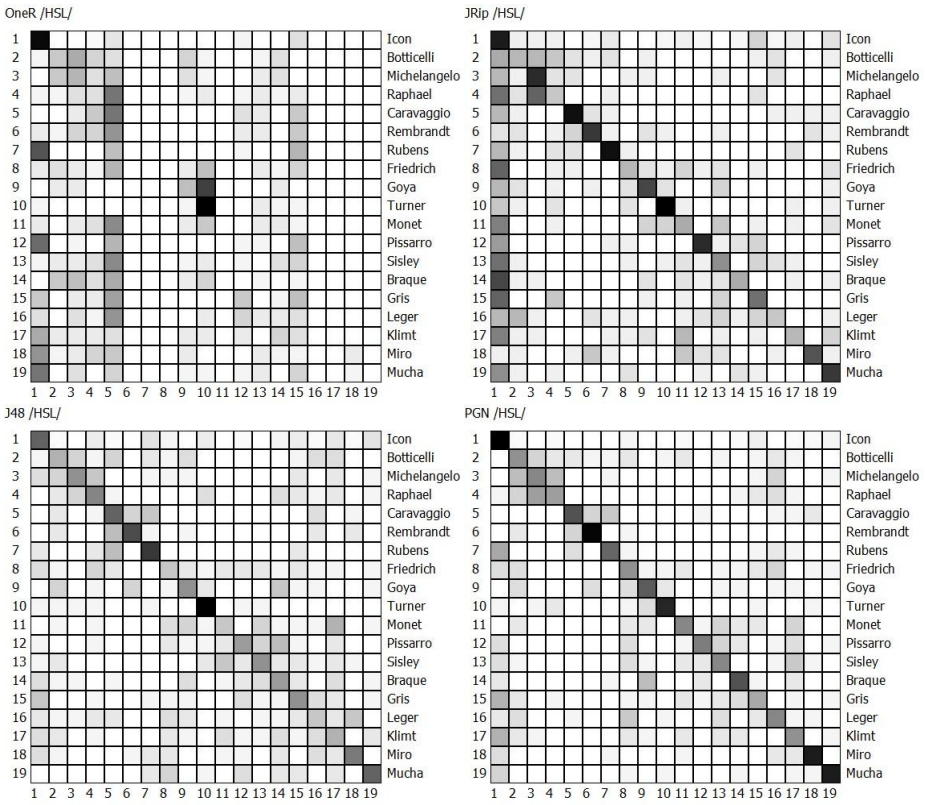

Fig.4. Confusion matrices for HSL features, artists as class labels

Figure 3 and Figure 4 show the confusion matrices for movements and for artists' names respectively. In the visualization of confusion matrices, the darker a square is, the bigger is the percentage of images following into corresponded square.

Analyzing the movements results three patterns immediately get attention. First the Baroque movement is the easiest to predict, OneR fails to predict Modern Art, PGN is the only classifier with a smooth consistent black/gray downwards diagonal. The first pattern repeats patterns seen in the descriptive analysis. It seems that Modern Art pictures cannot be characterized with one visual attribute. The characteristic PGN rules can better discriminate than J48 rules especially between the movements Romanticism, Impressionism, Cubism and Modern Art. Let's mention again the specifics of the PGN against other classifiers. All other classifiers take into account in one on other manner the support, controversially to PGN, which focuses primarily on the confidence of the association rules and only in a later stage on the support of the rules.

Analysing the artist results the three mentioned patterns are confirmed and two new ones are seen: the presence of vertical lines (dark or light) and the presence of "movement" squares.

It is clear that based on visual characteristics OneR is not able to classify the different artist paintings. JRip predicts almost 25% of the paintings as Icon (the vertical line in the JRip confusion matrix).

The datasets that we use here are specific because all artists are represented with equal numbers of paintings, and all selected movements contain also fixed number of artists, i.e. the distributions are equal. The exception is Icons, which are twice more than each artist and two-thirds than the movements. Because of this, we can see for the precision of Icons the tendencies of losing percentages for movements and enforcing ones for artists for OneR, Jrip and J48 – here and in consequent analyses.

The grey squares show some common tendencies of recognizing or misplacing the class labels. For instance, it is interesting that the Renaissance painters Botticelli, Michelangelo and Raphael are not recognized correctly but are misclassified mainly within their own group. Icons, Michelangelo, Caravaggio, Rembrandt, Rubens, Turner, Pissarro, Miro and Mucha are easier to classify.

4 Conclusion and Future Work

The growing number of digitised cultural heritage collections brought to a radically new level the access of users to art collections. Accessibility, however is hindered by the very large volume of available resources which calls for new approaches in resource discovery building on methods for content based image analysis; this would enhance search using not only available metadata but also user preferences related to the image content.

In this paper, we succinctly presented a vast range of methods for content based retrieval, concentration on the associative classifiers, which generate a set of association rules from a given training set – an approach which is particularly suited for art images where training sets are easy to construct. We used the classifier PGN over a dataset that included visual features, extracted by 600 paintings of 19 artists

from different movements of West-European fine arts and Eastern Medieval Culture. The results of the experiments confirm our expectations that the proposed approach to prioritize confidence over the support has its reason and leads to outperforming PGN against other rule-based classifiers especially in the case of multi-class datasets.

This result is quite interesting having in mind that PGN uses an approach which questions the traditional method employed by associative classifiers which prioritize support over confidence. PGN gives priority to confidence retaining only 100% confidence rules. In a task which includes multiple classes this new approach shows an advantage; evaluation of approaches and classifiers and coming with clear criteria which tools work best for specific cases of information retrieval is one of the areas where definitely more work will follow in the future years.

We believe that this approach can be successfully implemented in the resource discovery as a part of access functions in established digital libraries, repositories and aggregators and this way to increase the possibilities of such storages for ease access.

Acknowledgments. This work was supported in part by Hasselt University under the Project R-1875 "Search in Art Image Collections Based on Color Semantics", the Project R-1876 "Intelligent Systems' Memory Structuring Using Multidimensional Numbered Information Spaces", and by the Bulgarian National Science Fund under the Project D002-308 "Automated Metadata Generating for e-Documents Specifications and Standards".

References

1. Greenberg, J.: Metadata extraction and harvesting: A comparison of two automatic metadata generation applications. Journal of Internet Cataloging, 6(4), 59–82 (2004)
2. Martines, F., Morale, F.: Investigation of metadata applications at Palermo astronomical observatory". Library and Information Services in Astronomy IV, July 2-5 (2002)
3. Shi, R., Maly, F., Zubair, M.: Automatic metadata discovery from non-cooperative digital libraries. In Proc. of IADIS Int. Conf. on e-Society 2003, Lisbon, Portugal (2003)
4. Cardinaels, K., Meire, M., Duval, E.: Automating metadata generation: the simple indexing interface. In Proc. of 14th Int. Conf. on WWW, Chiba, Japan, ACM, NY, 548-556 (2005)
5. Fayyad, U., Piatetsky-Shapiro, G., Smyth, P.: From data mining to knowledge discovery: an overview. In Advances in Knowledge Discovery and Data Mining. American Association for AI, Menlo Park, CA, USA, 1-34, (1996)
6. Klosgen, W, Zytkow, J.: Knowledge discovery in databases terminology. In Advances in Knowledge Discovery and Data Mining. AAAI Press, 573-592 (1996)
7. Kouamou, G.: A software architecture for data mining environment. Ch.13 in New Fundamental Technologies in Data Mining, InTech Publ., 241-258 (2011)
8. Maimon, O., Rokach, L.: Decomposition Methodology for Knowledge Discovery and Data Mining. Vol. 61 of Series in Machine Perception and Artificial Intelligence, WSP (2005)
9. Mitchell, T.: Machine Learning, McGraw-Hill (1997)
10. Berry, P., Harrison, I., Lowrance, J., Rodriguez, A., Ruspini, E., Thomere, J., Wolverton, M.: Link Analysis Workbench. Technical Report for Air Force Res. Lab. IFOIPA (2004)
11. Liu, B., Hsu, W., Ma, Y.: Integrating classification and association rule mining. In Knowledge Discovery and Data Mining, 80-86 (1998)
12. Li, W., Han, J., Pei, J.: CMAR: Accurate and efficient classification based on multiple class-association rules. In: Proc. of the IEEE ICDM, 369-376 (2001)

13. Zaiane, O., Antonie, M.-L.: Classifying text documents by associating terms with text categories. J. Australian Computer Science Communications, 24(2), 215-222 (2002)
14. Yin, X., Han, J.: CPAR: Classification based on predictive association rules. In SIAM Int. Conf. on Data Mining, 331-335 (2003)
15. Zimmermann, A., De Raedt, L.: CorClass: Correlated association rule mining for classification. In Discovery Science, LNCS, Vol. 3245, 60-72 (2004)
16. Rak, R., Stach, W., Zaiane, O., Antonie M.-L.: Considering re-occurring features in associative classifiers. In Advances in Knowledge Discovery and Data Mining, LNCS, Vol. 3518, 240-248 (2005)
17. Coenen, F., Leng, P.: Obtaining best parameter values for accurate classification. In Proc. IEEE ICDM, 597-600 (2005)
18. Wang, J., Karypis, G.: HARMONY: Eficiently mining the best rules for classification. In Proc. of SDM, 205-216 (2005)
19. Thabtah, F., Cowling, P., Peng, Y.: MCAR: multi-class classification based on association rule. In Proc. of the IEEE ACS, 33-33 (2005)
20. Antonie, M.-L., Zaiane, O., Holte, R.: Learning to use a learned model: A two-stage approach to classification. In Proc. of IEEE, 33-42, (2006)
21. Tang, Z., Liao, Q.: A new class based associative classification algorithm. Int. Journal of Applied Mathematics, 36(2), 15-19 (2007)
22. Depaire, B., Vanhoof, K., Wets, G.: ARUBAS: an association rule based similarity framework for associative classifiers. In Proc. of IEEE ICDM Workshops, 692-699 (2008)
23. Mitov, I., Ivanova, K., Markov, K., Velychko, V., Vanhoof, K., and Stanchev, P.: PaGaNe – a classification machine learning system based on the multidimensional numbered information spaces. In WSPS on CEIS, No. 2, 279-286 (2009)
24. Mitov, I.: Class Association Rule Mining Using Multi-Dimensional Numbered Information Spaces. PhD Thesis, Hasselt University, Belgium (2011)
25. Witten, I., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques. 2nd Edition, Morgan Kaufmann, San Francisco (2005)
26. Holte, R.: Very simple classification rules perform well on most commonly used datasets. Machine Learning, Vol. 11, 63-91 (1993)
27. Ivanova, K., Stanchev, P., Vanhoof, K.: Automatic tagging of art images with color harmonies and contrasts characteristics in art image collections. Int. Journal on Advances in Software, 3 (3&4), 474-484 (2010)