A simple one class classifier with rejection strategy : application to symbol classification

Eugen Barbu*, Clement Chatelain*, Sebastien Adam*, Pierre Heroux*, Eric Trupin*

** LITIS labs. University of Rouen, Avenue de l'université, 76800 Saint Etienne du Rouvray*

E-mail: FirstName.LastName@univ-rouen.fr

Abstract

At the preceding GREC, we have proposed to use a "bag of symbols" formalism (similar to the bag of words approach) for the indexing of a graphical document image database. In this paper, we extend the proposed appraoch through the introduction of a rejection stage in the system. This rejection is based on the use of an original One Class Classifier. Some preliminary results are proposed.

Keywords: pattern recognition, symbol classification, one class classifiers, outliers rejection strategy

1. Introduction

At the preceding GREC [Barbu 2005] , we have proposed to use a "bag of symbols" formalism (similar to the bag of words approach) for the indexing of a graphical document image database. In this contribution, documents are represented and indexed using a set of symbols that are discovered in an unsupervised manner thanks to clustering algorithms and graph mining methods. The first enables to automatically label graph nodes representing connected components while the latter aims at finding frequent subgraphs in the graph describing the whole document. Numerous tests have been led since this proposition. An analysis of the obtained results has shown that the graph mining algorithm tends to integrate outliers in the proposed symbol classes. This problem strongly impacts on the global system performance. In this paper, we propose an outliers rejection strategy which has been developed to tackle this problem. It is based on a simple One-Class classifier and an original rejection strategy to filter outliers. In section 2, the general principles of one-class classifiers are given. Then, in section 3, our proposition is presented. In section 4, the application of the proposed approach to symbol classification is presented and evaluated.

2. One-class classifiers

Nowadays, new statistical learning methods such as Support Vector Machine (SVMs) or Artificial Neural Network (ANNs) provide strong classification performance for various pattern recognition applications. For example, in the field of handwriting recognition, recent benchmarks [Liu 2002] show that a precision accuracy of nearly 99% can be reached on the famous MNIST handwritten digits database, using SVM with RBF kernels. However, in real world applications, another important feature for pattern recognition systems is the rejection ability, i.e. the ability to reject the pattern which do not belong to any of the known classes. Such patterns are frequently denoted as outliers. In the case of a symbol recognition system, it consists in :

- choosing the right symbol class when a symbol appears
- rejecting the pattern if it does not belong to any of the classes of symbols.

The first objective can be performed by finding the best frontiers between the classes, whereas the second objective is obtained by finding either a density estimation of the class, either its boundaries.
One way to perform this second objective is the one-class classifiers. As said in [Tax 2001], the goal of one-class classification is to distinguish between a set of target objects and all other possible objects (per definition considered outliers objects). The advantage of this description is that it does not need outliers samples to train the model.

3. Proposed scheme

In this section we present a simple one-class classifier with an original and robust method to order the elements of a class with respect to the measure of how typical is an object within a class. Using this order, one can fix a rejection threshold in order to discard potential outliers from a class.

One way to obtain typical elements of a class of objects is to keep only the elements close to the center of the class. In order to measure this typicalness of objects, in a general dissimilarity space, a method is to compute the average distance of an element to all other objects of the class [Bunke and Sanfeliu, 1990].

$$T_i = (\sum_{j=1}^{N} d(e_i, e_j))/N$$

where N is the number of objects in a class, e_i is an element of the class, for which the index is computed, and $d(e, e')$ a distance function between objects.

The values $T_{1,} T_{2,}.., T_N$ can be easily calculated and only a percent of this objects can be retained if a filtering is needed. In such a case, The first k elements in the ascending ordered set of typicality values are kept as being closer to class center.

$$T_{j_1} \leq T_{j_2} \leq, \ldots, \leq T_{j_k} \leq T_{j_{k+1}} \leq \ldots \leq T_{j_N}$$

Another possibility to filter a class of objects, is to detect the potential outliers using the following described measure. Given the following notations :

- $kNN^d(o)$ the k-th nearest neighbors of an object o in the learning dataset using a distance function d .
- $FN^d(o)$ represents the farthest neighbor of an object o in the learning using a distance function d .

Then, the following equation can be used to measure if a pattern o_i is an outliers or not :

$$r_i^k = \frac{d(o_i, kNN^d(o_i))}{d(kNN^d(o_i), FN^d(kNN^d(o_i)))}$$

A value r_i^k bigger than 1 stands for an object o_i out of the class, from the point of view of the k-th nearest neighbor of o_i .
The decisions of several neighbors of o_i can be combined in a measure of membership :

$R_i^K = \sum_{k=1}^{K} r_i^k / K$ where K is the number of nearest neighbors used to estimate the membership.

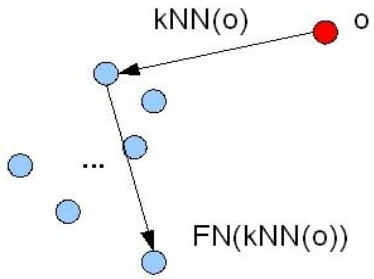


Figure 1. Illustration of the proposed approach principles

In Fig. 2 we can see that this second measure approximates better the class frontiers.

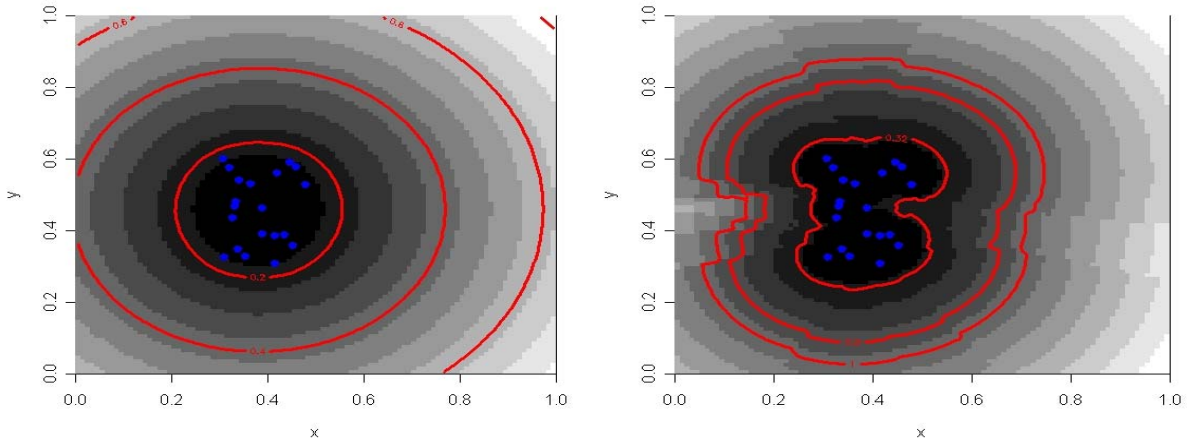


Figure 2. a comparison between the membership functions obtained using T_i and r_k

4. Application to symbol recognition

As said in the introduction, the proposed approach takes place in a graphical document retrieval system which uses a "bag of symbols" formalism to index documents. In this particular application, the aim is to filter outliers symbol which has been found using graph mining techniques. As explained in section 3, the single constraint in order to use the proposed approach is to define a distance measure between two individuals. Consequently, a matching distance between symbol images is needed. As presented in [Barbu 2005], each image can be made of a number of connected components. In the proposed approach each of the connected component is characterized by its surface and Zernike invariants. In order to measure the similarity of two symbol images, we compute the minimum cost matching between the set of connected components of the first image to the connected components of the second, with the possibility to delete connected components [Kriegel and Schönauer, 2003].

The problem is to find a function f that maps a connected component to other or to the delete operation $f:\{c_1, c_2, .., c_w\} \rightarrow \{c_1^{'}, c_2^{'}, ..., c_u^{'}\} \cup \Delta$ and minimize the sum: $\sum_{i=1}^{w} d(c_i, c_{f(i)}^{'})$ where , the distance between connected components is: $d(c_i, c_j^{'}) = L_1(area(c_i), area(c_j^{'})) + L_2(zm(c_i), zm(c_j^{'}))$ with L_1, L_2 euclidean norms.

This type of assignment problems are well studied in the field of combinatorial optimization [Papadimitriou and Steiglitz, 1998]. The Hungarian Method is an algorithm used for weighted assignment with $O(n^3)$ time complexity, where n is the number of connected components in our case.
Figure 3 illustrates the assignment process. On the left, we show the results of the assignment between a symbol and itself. In the middle, we show the results of the assignment between two occurrences of the same symbol. On the right, we show the results of the assignment between two different symbols.

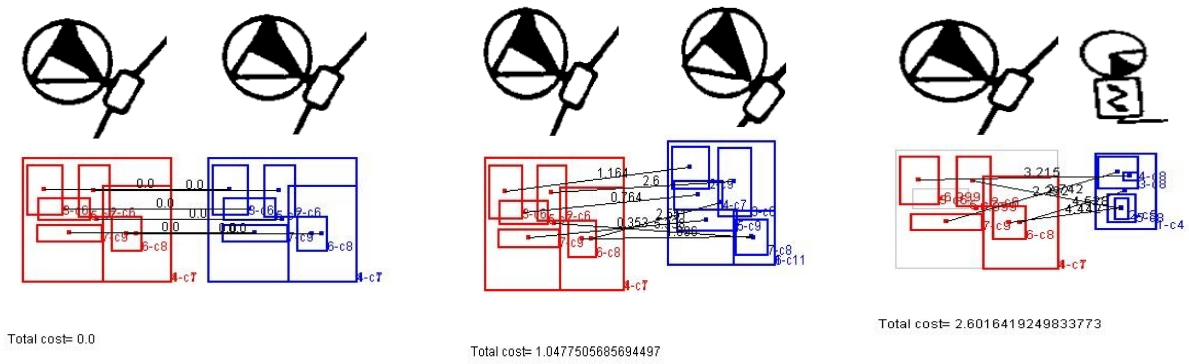
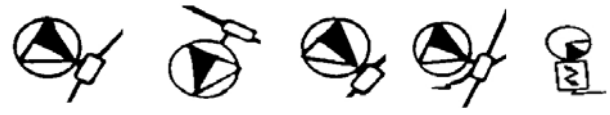


Figure 3. Assignment process



	d1	d2	d3	d4	d5
d1	0.0	1.28	1.04	1.35	2.6
d2	1.28	0.0	1.39	1.68	2.74
d3	1.04	1.39	0.0	1.43	2.58
d4	1.35	1.68	1.43	0.0	2.87
d5	2.6	2.74	2.58	2.87	0.0

Table 1. Assignment results

Table 1 shows another example of assignment results with 5 symbols. Using these results, the proposed approach can be applied to reject the outliers. We obtain :

$$(T_1 = 1.25) < (T_3 = 1.28) < (T_2 = 1.41) < (T_4 = 1.46) < (T_5 = 2.15)$$

$$(R_1^1 = 0.403) = (R_3^1 = 0.403) < (R_2^1 = 0.406) < (R_4^1 = 0.47) < (R_5^1 = 0.89)$$

and from that we can spot *d5* as an outliers in the presented class of symbols.

5. Conclusion

In this paper, we have proposed a simple one class classifier. This one is applied in the context of symbol recognition for document retrieval. Preliminary results have shown promising results. Some tests are led currently on different databases.

Bibliography

[Barbu 2005] Barbu E., Héroux P, Adam S., Trupin E. (2005). Using Bags of symbols for automatic indexing of graphical document image database. GREC'2005, pp 185-193.

[Bunke and Sanfeliu, 1990] Bunke, H. and Sanfeliu, A. (1990). Syntactic and structural pattern recognition : theory and applications. World Scientific.

[Kriegel and Schönauer, 2003] Kriegel, H. P. and Schönauer, S. (2003). Similarity search in structured data. In Data Warehouse and Knowledge Discovery, volume 2737 of Lecture Notes in Computer Science, pages 309 319. Springer-Verlag.

[Liu] Liu, C.L., Sako, H., Fujisawa H. (2002) . Performance evaluation of pattern classifiers for handwritten character recognition. IJDAR, vol. 4, pp 191-204.

[Papadimitriou and Steiglitz, 1998] Papadimitriou, C. H. and Steiglitz, K.(1998). Combinatorial Optimization : Algorithms and Complexity. Dover Pubns.

[Tax&Duin] Tax, D.M.J. , Duin R. P. W. (2001) Combining One-Class Classifiers. MCS, pp 299-308, 2001.