Selection of Suitable Set of Decision Rules Using Choquet Integral

Laurent Wendling[1], Jan Rendek[1], and Pascal Matsakis[2]

[1]LORIA - ESIAL, Universit Henri Poincar
54506 Vandœuvre-ls-Nancy, France
[2]CIS Dpt, Universit de Guelph
ON N1G 2W1 Canada
{wendling,rendek}@loria.fr, matsakis@cis.uoguelph.ca

Abstract. An approach to automatically extract pertinent subsets of soft output classifiers, assumed to decision rules, is presented in this paper. They are aggregated into a global decision scheme using the Choquet integral. A selection scheme is defined that discards weak or redundant decision rules, keeping only the most relevant subset. An experimental study, based on real world data attest the interest of our method.

1 Introduction

A pattern recognition system can be roughly decomposed into three successive steps [6,18]. Firstly, the shapes are extracted from their surrounding background. This step, called segmentation, relies heavily on *a priori* knowledge of the documents to be processed. Secondly, a representation is built from the extracted patterns. This representation allows to computationally judge the (dis)similarity between two patterns. It can either be a set of measurements performed on the patterns, forming a vector of features, or a symbolic description of how the pattern can be divided into basic shapes. Thirdly, a decision rule is built, using the representation of the pattern. This rule predicts to which class an observed pattern most likely belongs. It can either be built by introducing some expert knowledge about the patterns, or learned on a representative subset of labeled patterns. Surveys on pattern representation and classification techniques over the last decade fail to conclude whether a set of generic methods performing best on any kind of document, can be found [6,3,11]. Rather, a collection of techniques has been developed to address domain specific issues. For a given application, the choice of a pattern representation or a decision scheme relies at best on the extensive testing of several combinations of techniques. It is not uncommon that a designer chooses by making an educated guess between the techniques he has at hand, or between those he is the most acquainted with. A tempting approach is to combine several decision rules, based on various representations and classification schemes, instead of electing only one. The expected outcome is a more robust final decision, taking advantage of all the decision rules qualities. This approach appears well suited to cope with cases where the

N. da Vitora Lobo et al. (Eds.): SSPR&SPR 2008, LNCS 5342, pp. 947–955, 2008.

available data is too scarce to determine the best method by thorough testing, or to build robust decision rules by learning. Many classifier combination systems have been proposed and compared in the literature [9,16,10,19]. In this paper, an improvement of the aggregation of decision rules using the Choquet integral is proposed. The Choquet integral is part of the aggregation techniques based on fuzzy integrals and have been successfully used as fusion operators in various applications.

2 Background on Choquet Integral Fusion

2.1 Decision Rules Fusion

Let us consider m classes, $\mathcal{C} = \{C_1, \ldots, C_m\}$, and n Decision Rules (DRs) $X = \{D_1, \ldots, D_n\}$. When a new pattern x^o is observed, we wish to find the class it most likely belongs to. Labeling this unknown pattern is a three-steps process. Firstly, for each decision rule j and each class i, we compute ϕ_j^i the degree of confidence in the statement "According to D_j, x^o belongs to the class C_i". Secondly, we combine all these partial confidence degrees into a global confidence degree by choosing a suitable aggregation operator $\mathcal{H}$. Thus, the global confidence degree in the statement "x^o belongs to C_i", noted $\Phi(C_i|x^o)$, is given by:

$$\Phi(C_i|x^o) = \mathcal{H}(\phi_1^i, \ldots, \phi_n^i)$$

Finally, x^o is assigned to the class for which the confidence degree is the highest.

$$label(x^o) = \underset{i=1}{\overset{m}{\arg}}\ max\ \Phi(C_i|x^o)$$

Many aggregation operators were introduced in the literature. If the classification issue implies more than two classes, two learning approaches can be followed. Either each class C_i is paired with his own aggregation operator $\mathcal{H}^i$, or a single global aggregation operator is learned. In the former case, the final decision is slightly modified as the global confidence degree depends on the operator associated with the class.

$$\Phi(C_i|x^o) = \mathcal{H}^i(\phi_1^i, \ldots, \phi_n^i)$$

2.2 Fuzzy Measures and the Choquet Integral

The Choquet integral was first introduced in the capacity theory [2,15]. Let us denote by $X = \{D_1, \ldots, D_n\}$ the set of n decision rules, and $\mathcal{P}$ the power set of X, i.e. the set of all subsets of X.

Definition 1. *A* fuzzy measure *or* capacity, μ, *defined on* X *is a set function* $\mu : \mathcal{P}(X) \longrightarrow [0, 1]$, *verifying the following axioms:*
$\mu(\emptyset) = 0$, $\mu(X) = 1$
$A \subseteq B \Longrightarrow \mu(A) \leqslant \mu(B)$

Fuzzy measures generalize additive measures, by replacing the additivity axiom by a weaker one (monotonicity). Fuzzy measures embed particular cases including probability measure, possibility and necessity measures, or belief and plausibility functions. In our context of decision rules fusion, $\mu(A)$ represents the importance, or the degree of trust in the decision provided by the subset A of DRs. The next step in building a final decision, is to combine the partial confidence degree according to each DR into a global confidence degree, taking those weights into account.

Definition 2. *Let μ be a fuzzy measure on X. The* discrete Choquet *integral of $\boldsymbol{\phi} = [\phi_1, \ldots, \phi_n]^t$ with respect to μ, noted $\mathcal{C}_\mu(\boldsymbol{\phi})$, is defined by:*

$$\mathcal{C}_\mu(\boldsymbol{\phi}) = \sum_{j=1}^{n} \phi_{(j)}[\mu(A_{(j)}) - \mu(A_{(j+1)})]$$

where (.) is a permutation and $A_{(j)} = \{(j), \ldots, (n)\}$ represents the $[j..n]$ associated criteria in increasing order and $A_{(n+1)} = \emptyset$.

2.3 Determining the Fuzzy Measure

There are several methods to determine the most adequate fuzzy measure to be used for a given application and the most straightforward learning approach is based on optimization techniques. The aim is to find the fuzzy measure that minimizes best a criterion on the training set, such has the square error. Considering $(x^k, y^k), k = 1, \ldots, l$, l learning samples where $x^k = [x_1^k, \ldots, x_n^k]^t$ is a n-dimensional vector, and y^k the expected global evaluation of object k, the fuzzy measure can be determined by minimizing [5]: $E^2 = \sum_{k=1}^{l} (\mathcal{C}_\mu(x_1^k, \ldots, x_n^k) - y^k)^2$. This criterion can be put under a quadratic program form and solved by the Lemke method. Nevertheless the method requires at least $n!/[(n/2)!]^2$ learning samples. When little data is available, matrices may be ill-conditioned, causing a bad behavior of the algorithm. To cope with the above problems, "heuristic" algorithms have been developed. To our knowledge, the algorithm providing the best approximation was proposed by Grabisch in [4]. It assumes that in the absence of any information, the most reasonable way to aggregate the partial matching degrees is to compute the arithmetic mean on all the inputs.

2.4 Behavioral Analysis of the Aggregation

The importance index is based on the definition proposed by Shapley in game theory [17]. It is defined for a fuzzy measure μ and a rule i as:

$$\sigma(\mu, i) = \frac{1}{n} \sum_{t=0}^{n-1} \frac{1}{\binom{n-1}{t}} \sum_{\substack{T \subseteq X \setminus i \\ |T|=t}} [\mu(T \cup i) - \mu(T)]$$

It can be interpreted as a average value of the marginal contribution $\mu(T \cup i) - \mu(T)$ of the decision rule i alone in all combinations. The interaction index [14] represents the degree of interaction between two decision rules. If the fuzzy measure is non-additive then some sources interact. The *marginal interaction* between I and j, conditioned to the presence of elements of the combination $T \subseteq X \backslash ij$ is:

$$(\Delta_{ij}\mu)(T) = \mu(T \cup ij) + \mu(T) - \mu(T \cup i) - \mu(T \cup j)$$

Averaging this criterion over all the subsets of $T \subseteq X \backslash ij$ gives the interaction index of sources i and j.

$$I(\mu, ij) = \sum_{T \subseteq X \backslash ij} \frac{(n-t-2)!t!}{(n-1)!} (\Delta_{ij}\mu)(T)$$

A positive interaction index for two DRs i and j means that the importance of one DR is reinforced by the second. A negative interaction index indicates that the sources are antagonist, and their combined use impairs the final decision.

3 Extraction of Decision Rules

3.1 Handling with Learning Error

Lattices (associated to fuzzy measures) are initialized at the arithmetic mean, and are approximated using a training set via a gradient descent. From training pattern, m training samples are created $\Phi^1, \ldots, \Phi^m$, with $\Phi^i = (\phi_1^i, \ldots, \phi_m^i)$ where ϕ_j^i represents the confidence in the fact that the sample belongs to class i, according to DR j. Each of these samples is paired with a target value, i.e. the value an ideal operator would output using this sample as input. For techniques that use a single fuzzy measure no real formula exists and often the following one is used:

$$\mathcal{C}_\mu(\Phi^i) = \begin{cases} 1, \text{ if sample belongs to class } i, \\ 0, \text{ otherwise.} \end{cases}$$

For techniques that use a different fuzzy measure per class, the optimal target value minimizing the quadratic error is known for two classes [4]. Best criterion is defined as follows.

$$E^2 = \sum_i \sum_j |\Psi(\Delta\Phi_{12}(X_k^j)) - 1|^2$$

with X_k^j is the kth training data of class j, Ψ a sigmoid-type function and $\Delta\Phi_{12}$ is given by:

$$\Delta\Phi_{12} = \Phi_{\mu_1}(C_1|X_k^j) - \Phi_{\mu_2}(C_2|X_k^j)$$

The error used in the gradient descent algorithm is given by:

$$e = \Psi(\Delta\Phi_{12}(X_k^j)) - 1)$$

The criterion is modified as follows to process with N-classes:

$$\Delta\Phi_{qr_{r\in X-\{q\}}} = \Phi_{\mu_q}(C_q|X_k^j) - \bigotimes_{\{r\in X-\{q\}\}} \Phi_{\mu_r}(C_r|X_k^j)$$

with q and r two classes. Operator $\bigotimes$ can be a median, *min*... Here the max was kept to move away ambiguous classes and so to favor a discrimate behavior. Median is interesting to preserve acceptable results even if some sources may be contradictory. The sign of $\Delta\Phi_{qr_{r\in X-\{q\}}}$ is studied to set the error to be propagated in the lattice as follows.

$$sign(\Delta\Phi_{qr_{r\in X-\{q\}}}) = \begin{cases} +e = 1./f(\Delta\Phi_{qr_{r\in X-\{q\}}}) \\ -0,\ e = 0 \text{ for } q,\ 1 \text{ for others.} \end{cases}$$

with f an increasing function.

3.2 Weaker Decision Rule Extraction

Once the lattice is learned, the individual performance of each DR is analysed in the produced fuzzy measure. This analysis is performed using the importance and interaction indexes defined in 2.4. The aim is to track the DRs having the weak importance in the final decision, and that positively interact the least with the other rules. Such DRs are assuming to blur the final decision. First low significant rules S_L having an importance index lower than 1 are selected:

$$S_L = \{k \mid n \cdot \sigma(\mu, k) < 1\}$$

The set of rules to be removed MS_L is composed of the rules having an interaction index lower than the mean of the interaction indexes of S_L:

$$MS_L = \{k \mid \sum_{j=1,n} I(\mu, kj) < m\}_{k\in S_L}$$

with the global mean interaction index $m = \frac{1}{|S_L|} \sum_{k\in S_L} \sum_{j=1,n} I(\mu, kj)$

3.3 Extracting Decision Rules

We use the classic fuzzy pattern matching setup. Each class $C_1, \ldots, C_n$ is associated with a fuzzy measure $\mu_1, \ldots, \mu_n$. The evaluation measures are generally classifier dependant [13]. For their computation the classification of training data must be performed to identify the fuzzy measure for each class. The application of our learning algorithm followed by the extraction of the most relevant DRs forms a training epoch is quite similar to global scheme but regarding a set of classes aggregated at las using a *argmax* criterion. As it is not easy to valuate each combinaison of classes. we consider them indepedently. A Greedy feature selection algorithm is used to ensure a continuous extraction of unexpected features per class. At each epoch the weakest descriptor calculated from indices

is removed and so on, while improving a gain (recognition rates). The overall algorithm is described below.

% Initialization %

For EachC_i $\mathcal{C}$

- Learned μ_i
- Extract weakest descriptor

End For Each

% Main %

While Minimization is on the way

- Replace old capacity with new capacity
- Evaluate gain (minimizing cost function)
- Keep the "best" new capacity
- Extract feature for that new capacity

End while

4 Experimental Results

The proposed approach being aimed at situations where little information is available for training, databases having a fair number of categories and a small number of samples have been used . For the experiments, several decision rules are set from different photometric descriptors associated to a basic similarity measure to belong in the same range. A set of nine pattern recognition methods $R = \{R_i\}_{i=1,9}$ is used here, most of them having a low processing time, easy to implement, and invariant to affine transforms such as translation, rotation, or scaling. The descriptors computed on the shapes are: $DC = \{$ ART [8], angular signature [1], GFD [22], Ellipticity,f_0 and f_2-histograms [12], moments of Zernike [7], Yang [21], Radon signature [20]$\}$. B1(9 classesx11samples) and B2(18,12) are well-known sharvit' databases (see figure 1).

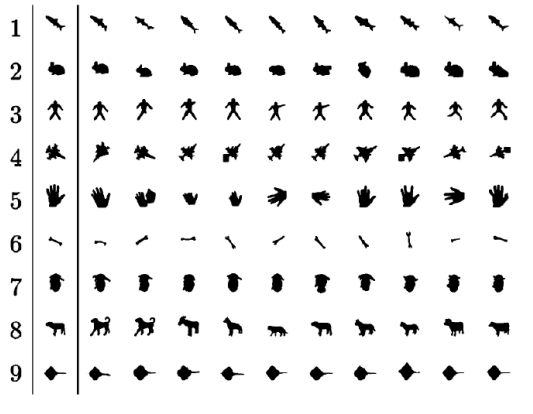

Fig. 1. The first Sharvit' database (B1)

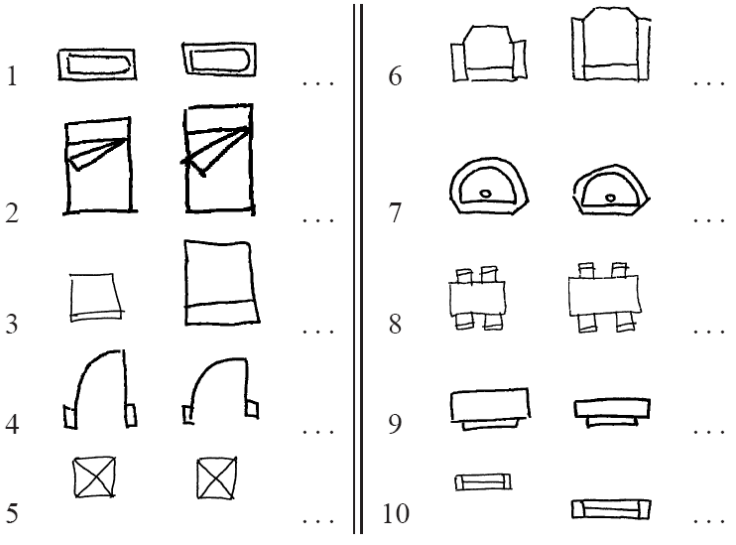

Fig. 2. Samples of the CVC database (B3)

Table 1. Recognition rates reached for each method

	R_1	R_2	R_3	R_4	R_5	R_6	R_7	R_8	R_9	Mi	Ma	Me	Ch
B1	93	59	98	48	74	50	96	85	66	43	99	78	100
B2	88	56	90	48	62	46	86	66	56	59	87	76	93
B3	80	30	38	22	20	46	80	76	56	46	82	78	89

B3(10x300) is a database kindly provided by the CVC Barcelone. Symbols have been drawn by ten people using "anoto" concept. Few samples are provided in figure 2.

An experimental study is carried out from each decision rules R_i, basic aggregation operators as Min, Max, Median and the proposed method based on Choquet integral.

For each test, a crossvalidation was applied (1/3 and 2/3). Table 1 shows the good behavior of the method on these databases. The results reached independently by each decision rule R_i are coherent with the amount of information processed by associated features. Simple aggregation operators: minimum, maximum and median exhibit various behavior depending on the database. In terms of recognition rates, only the maximum achieves better results than the best simple DR on two databases. In comparison to the simple DRs and the simple aggregation operators, our fusion operators based on the Choquet integral consistently achieve better results on each databases, in terms of recognition rates. On each dataset, our recognition method at worse improves a little the recognition rates; it never worsens them. Moreover a decreasing of the number of

decision rules around 50% per class has been obtained. It allows to define a kind of identity map per class consisting in the most suitable decision rules following the application under consideration.

5 Conclusion

The algorithm appears well suited for the situation when no *a priori* knowledge is available about the relevance of a set of decision rules for a given dataset, or when the training set available is to small to build reliable decision rules, or to determine which method is the best. It finds the best consensus between the rules, taking their interaction into account, and discarding the undependable or redundant ones. A way to merge both numerical and expert decision rules in order to improve the recognition process is under consideration.

References

1. Bernier, T., Landry, J.-A.: A new method for representing and matching shapes of natural objects. Pattern Recognition 36(8), 1711–1723 (2003)
2. Choquet, G.: Theory of capacities. Annales de l'Institut Fourier 5, 131–295 (1953)
3. Cordella, L.P., Vento, M.: Symbol recognition in documents: a col of technics? International Journal of Document Analysis and Recognition 3(2), 73–88 (2000)
4. Grabisch, M.: A new algorithm for identifying fuzzy measures and its application to pattern recognition. In: Int. conf. FUZZ'IEEE 1995, pp. 145–150 (1995)
5. Grabisch, M., Nicolas, J.M.: Classification by fuzzy integral - performance and tests. Fuzzy Sets and Systems 65, 255–271 (1994)
6. Jain, A.K., Duin, R.P.W., Mao, J.: Statistical Pattern Recognition: A Review. IEEE Transactions on Pattern Analysis and Machine Intelligence 22(1), 4–37 (2000)
7. Khotanzad, A., Hong, Y.H.: Invariant Image Recognition by Zernike. IEEE Transactions on Pattern Analysis and Machine Intelligence 12(5), 489–497 (1990)
8. Kim, W.-Y., Kim, Y.-S.: A new region-based shape descriptor. In: TR 15-01, Pisa, Italy (1999)
9. Kittler, J., Hatef, M., Duin, R., Matas, J.: On combining classifiers. IEEE Transactions on Pattern Analysis and Machine Intelligence 20(3), 226–239 (1998)
10. Kuncheva, L.I., Whitaker, C.J.: Measures of diversity in classifier ensembles. Machine Learning 51, 181–207 (2003)
11. Lladós, J., Valveny, E., Sánchez, G., Martí, E.: Symbol Recognition: Current Advances and Perspectives. In: Blostein, D., Kwon, Y.-B. (eds.) GREC 2001. LNCS, vol. 2390, pp. 104–127. Springer, Heidelberg (2002)
12. Matsakis, P., Wendling, L.: A New Way to Represent the Relative Position Between Areal Objects. IEEE Transactions on Pattern Analysis and Machine Intelligence 21(7), 634–643 (1999)
13. Mikenina, L., Zimmermann, H.-J.: Improved feature selection and classification by the 2-additive fuzzy measure. Fuzzy Sets and Systems 107, 197–218 (1999)
14. Murofushi, T., Soneda, S.: Techniques for reading fuzzy measures(iii): interaction index. In: Proc. of the 9th Fuzzy Set System, pp. 693–696 (1993)
15. Murofushi, T., Sugeno, M.: A theory of fuzzy measures: representations, the Choquet integral, and null sets. Journal of Math. Anal. Appl. 159, 532–549 (1991)

16. Ruta, D., Gabrys, B.: An overview of classifier fusion methods. Comp. and Information System 7(1–10) (2000)
17. Shapley, L.: A value for n-person games. In: Khun, H., Tucker, A. (eds.) Annals of Mathematics Studies, pp. 307–317. Princeton University Press, Princeton (1953)
18. Smeulders, A.W.M., Worring, M., Santini, S., Gupta, A., Jain, R.: CB Image Retrieval at the End of the Early Years. IEEE Transactions on Pattern Analysis and Machine Intelligence 22(12), 1349–1380 (2000)
19. Stejic, Z., Takama, Y., Hirota, K.: Mathematical aggregation operators in image retrieval: effect on retrieval performance and role in relevance feedback. Signal Processing 85(2), 297–324 (2005)
20. Tabbone, S., Wendling, L.: Binary shape normalization using the Radon transform. In: Nyström, I., Sanniti di Baja, G., Svensson, S. (eds.) DGCI 2003. LNCS, vol. 2886, pp. 184–193. Springer, Heidelberg (2003)
21. Yang, S.: Symbol Recognition via Statistical Integration of Pixel-Level Constraint Histograms: A New Descriptor. IEEE Transactions on Pattern Analysis and Machine Intelligence 27(2), 278–281 (2005)
22. Zhang, D., Lu, G.: Shape-based image retrieval using generic fourier descriptor. Signal Processing: Image Communication 17, 825–848 (2002)