Measuring the Applicability of Self-organization Maps in a Case-Based Reasoning System

A. Fornells, E. Golobardes, J.M. Martorell, J.M. Garrell, E. Bernadó, and N. Macià

Research Group in Intelligent Systems
Enginyeria i Arquitectura La Salle, Ramon Llull University
Quatre Camins 2, 08022 Barcelona, Spain
{afornells,elisabet,jmmarto,josepmg,esterb,nmacia}@salle.url.edu
http://www.salle.url.edu/GRSI

Abstract. Case-Based Reasoning (CBR) systems solve new problems using others which have been previously resolved. The knowledge is composed of a set of cases stored in a case memory, where each one describes a situation in terms of a set of features. Therefore, the size and organization of the case memory influences in the computational time needed to solve new situations. We organize the memory using Self-Organization Maps, which group cases with similar properties into patterns. Thus, CBR is able to do a selective retrieval using only the cases from the most suitable pattern. However, the data complexity may hinder the identification of patterns and it may degrade the accuracy rate. This work analyses the successful application of this approach by doing a previous data complexity characterization. Relationships between the performance and some measures of class separability and the discriminative power of attributes are also found.

Keywords: Statistical and Structural Pattern Recognition, Data Complexity, Neural Networks, Self-Organization Maps, Case-Based Reasoning, Soft Computing.

1 Motivation

Case-Based Reasoning (CBR) [1] is an approach based on solving new problems using others which have been previously solved. The knowledge is represented by a case memory, where each case is defined by a set of features that describe the problem. The way in which CBR works can be summarized in the following steps: (1) it retrieves the most similar cases from the case memory, (2) it adapts them to propose a new solution, (3) it checks if this solution is valid, and finally, (4) it stores the solution according to a learning policy. The CBR performance, in terms of computational time, is related to the size of the case memory because CBR has to explore it in the retrieval phase. Therefore, its organization can help to improve this issue by avoiding the selection of useless cases. There are mainly two organization strategies: (1) The identification of patterns for using only the cases from the best matching patterns [2,3], and; (2) The rejection of cases in

J. Martí et al. (Eds.): IbPRIA 2007, Part II, LNCS 4478, pp. 532–539, 2007.

function of their features' values [4]. However, the use of fewer cases may imply a reduction of the solving capabilities.

The SOMCBR (Self-Organization Map in a Case-Based Reasoning) [5] system is a CBR framework where the case memory has been organized by a Self-Organization Map (SOM) [6]. SOM is a clustering technique that defines patterns by highlighting the most important features of the data. These patterns allow SOMCBR to do a selective retrieval based on using only the cases from the most suitable pattern instead of all the cases. Thus, the computational time is reduced obtaining a meaningful property for real time environments [9]. Nevertheless, the SOMCBR success depends on the existence of reliable data patterns.

The goal of this paper is to show how a previous data complexity [11] analysis can help us to predict the SOMCBR applicability by evaluating the presence of useful data patterns.

The paper is organized as follows. Section 2 explains the previous work on SOMCBR. Section 3 briefly describes the data complexity analysis and proposes a set of metrics as predictors of the SOMCBR applicability. Section 4 summarizes the experiments and the results. Finally, we present the conclusions and further work.

2 Self-organization Map in a Case-Based Reasoning System

SOM is an unsupervised clustering technique from the neural network approach. It defines a topology map, where the cases are grouped in patterns. This ability is used to organize the CBR case memory in the SOMCBR approach [5]. Figure 1 illustrates a case memory organized by a 2-dimensional map of $M \times M$ patterns. The SOM has two layers: (1) The input layer is composed of N neurons, where each neuron represents one of the N-dimensional features of the input case, and; (2) The output layer is composed of $M \times M$ neurons, where each neuron contains a set of similar cases represented by a director vector. Each input neuron is connected to all the output neurons. When a new input case C is introduced in the input layer, each neuron from the output layer computes a degree of similarity between the input case C and its director vector applying a similarity function. In our approach, we use the complementary of the normalized Euclidean distance (see Eq. 1). A value closer to 1 means that the input case C should be similar to the elements from the Xth pattern (M_X). Otherwise, it should be different.

$$similarity(C, M_X) = \left|1 - d(\overline{C}, \overline{M_X})\right| = \left|1 - \sqrt{\frac{\sum_{n:1..N}(C(n) - M_X(n))^2}{N}}\right| \quad (1)$$

The retrieval consists in: (1) Looking for the most similar pattern, and; (2) Comparing with the cases from the selected pattern. Consequently, SOMCBR reduces the computational time because only a subset of the cases are used. Nevertheless, the patterns definition can be compromised due to the data complexity.

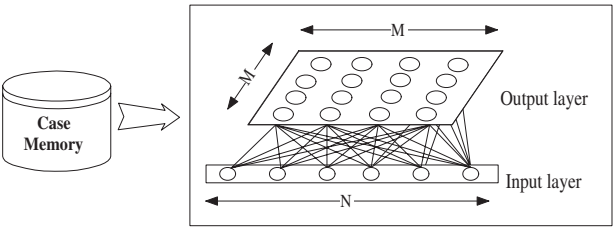

Fig. 1. The case memory is organized by the SOM in order to define $M \times M$ groups of cases with similar properties. This organization allows the CBR system to improve the computational time in the retrieval phase.

3 Data Complexity Measures

The study of data complexity addresses the characterization of the intrinsic complexity of the dataset, and to what extent this complexity is related to the classifier's performance [10]. Although dataset complexity may be related to three main causes (class ambiguity, boundary complexity, and training set sparsity) the previous studies in this matter have been focused on the characterization of boundary complexity, due to the difficulty to determine class ambiguity and the real sparsity of a training set. Ho & Basu [11] proposed a measurement space to identify the different aspects of boundary complexity: the discriminant power of attributes, the separability of classes, and the topology of classes such as the degree of overlap and the geometry of classes distributed as hyperspheres. Based on this previous study, we select those measures that are most relevant to identify meaningful structures in the dataset that could be correlated with SOMCBR clusters. We find that measures related to the separability of classes are the most useful to predict SOMCBR's success. Also measures detecting the degree of class overlap with respect to the feature space are useful to explain SOMCBR's behaviour. Other types of measures given in [11] do not reveal any structure as seen by SOMCBR's clusterization. In the following, we briefly describe these relevant metrics.

Feature efficiency (F3): it defines the efficiency of each feature individually and describes to what extent the feature takes part in the class separability. For each feature, the measure uses a local continuity heuristic which supposes that all the points belonging to the same class are included in the interval between the minimum and maximum value of that feature. Thus, if two instances of opposite classes have the same value for an attribute, there is an overlap and the instances are considered ambiguous for this dimension. The ambiguity is solved removing these instances. The efficiency is then assessed as the ratio of the remaining (non-overlapping) points to all the training points. The measure of feature efficiency is the maximum feature efficiency of all dimensions.

Length of class boundary (N1): it measures the number of training points located near the class boundary. It is based on building a minimum spanning tree

(MST) connecting all training points, using Euclidean distances between each pair of points. Then, the measure computes the number of points of opposite classes that are connected in the MST with respect to the total number of points. N1 is an indicator of class separability and cluster tendency; the higher the measure, the greater the presence of points of different classes on the boundary.

Intra/inter class nearest neighbour distances (N2): it describes the dispersion within classes with respect to the separability of classes. It is based on computing the Euclidean distance of each point with the nearest neighbour within the same class and the nearest neighbour of the opposite class. N2 is the ratio between the average within-class nearest neighbour distances and the average opposite-class nearest neighbour distances. A low value indicates a major degree of clustering and higher separability among different classes.

4 Results and Discussion

4.1 Testbed and Results

Several datasets of different domains and characteristics from the UCI Repository [13] are considered for studying the relation between the data complexity and the SOMCBR applicability. Due to the way in which the complexity measures are implemented [11], the datasets of N-class are split in N datasets of two classes: each class versus all other classes. The name and the number of features and instances are described in table 1.

The experimentation is performed in two parts. First, we compute the data complexity of each normalized dataset for several measures. Next, CBR and SOMCBR are executed applying a 10-fold stratified cross-validation with the following configuration: (1) The retrieve phase uses the Euclidean distance as similarity function; (2) The reuse phase proposes a solution using the most similar case, and; (3) The retain phase does not learn. Additionally, the SOMCBR is tested with 10 random seeds. All these results are also summarized in table 1: N1, N2 and N3 are the complexity measures; %AR and σ are the accuracy rate and its standard deviation for CBR, and for the best configuration of SOMCBR; %R is the reduction in the number of operations between CBR and SOMCBR; p-value is the probability to reject the null hypothesis assuming equal values for %AR of both approaches [14]. Small values of p-value imply a high probability of significant difference between both %AR.

Table 1 is divided (by an horizontal line) in two categories ordered by p-value. **Type 1** represents situations where the computational time is improved and the accuracy rate is at least maintained. On the other hand, **type 2** is produced when the accuracy rate is proportional to the number of cases retrieved and, consequently, the accuracy rate depends on the number of cases used. Therefore, the difference between both types indicates if the SOM is capable or not to splitting the domain in well defined patterns.

Table 1. Summary of the dataset description (number of instances and attributes), the results from CBR and SOMCBR (accuracy rates (%AR) with their standard deviation (σ)) and, the results from the comparison between CBR and SOMCBR (percentage of reduction (%R) and the probability of CBR and SOMCBR being equal (p-value)). The horizontal line divides datasets into type 1 and 2, which are ordered by the p-value.

Dataset			Measures			CBR	SOMCBR	Statistics	
Name	Inst.	Attr.	N1	N2	F3	%AR(σ)	%AR(σ)	%R	p-value
Waveform c1	5000	41	0.24	0.86	0.23	83.2 (1.2)	81.1 (1.2)	89.2	0.00
Vehicle c1	846	19	0.12	0.42	0.46	93.4 (2.4)	87.5 (4.7)	86.9	0.00
Vehicle c4	846	19	0.09	0.54	0.22	96.0 (4.2)	89.2 (3.6)	87.7	0.00
Balance c2	625	5	0.20	0.62	0.00	87.0 (3.1)	81.8 (4.1)	89.7	0.00
Waveform c2	5000	41	0.27	0.90	0.15	80.2 (1.4)	78.8 (1.6)	89.7	0.01
Pim	768	9	0.44	0.84	0.01	71.3 (3.4)	69.9 (3.4)	87.9	0.03
Wpbc	198	34	0.42	0.91	0.18	73.7 (7.1)	73.2 (9.2)	82.5	0.03
Waveform c3	5000	41	0.23	0.85	0.24	83.6 (1.8)	82.7 (1.6)	89.3	0.03
Balance c3	625	5	0.20	0.62	0.00	86.9 (3.7)	82.3 (6.5)	89.5	0.04
Tao	1888	3	0.07	0.16	0.36	95.4 (1.3)	94.9 (1.6)	81.8	0.06
Wdbc	569	31	0.07	0.56	0.52	95.1 (3.2)	95.3 (2.7)	80.2	0.09
Wbcd	699	10	0.06	0.34	0.12	95.3 (2.2)	94.6 (2.6)	86.9	0.09
Vehicle c3	846	19	0.37	0.74	0.06	73.9 (4.1)	73.4 (4.5)	82.5	0.11
Vehicle c2	846	19	0.37	0.71	0.04	75.3 (3.4)	75.4 (2.9)	81.9	0.11
Bpa	345	7	0.58	0.91	0.03	62.9 (6.0)	63.2 (5.1)	52.6	0.17
Heart-Statlog	270	14	0.37	0.67	0.01	74.1 (6.4)	76.3 (8.3)	87.1	0.19
Balance c1	625	5	0.21	0.65	0.00	83.7 (2.2)	86.1 (4.9)	89.0	0.21
Wisconsin	699	10	0.06	0.33	0.12	96.1 (2.0)	96.9 (2.4)	84.5	0.33
Ionosphere	351	35	0.23	0.63	0.19	86.9 (4.1)	88.1 (3.6)	64.0	0.41
Iris c2	150	5	0.01	0.10	1.00	100.0 (0.0)	100.0 (0.0)	56.3	0.00
Thyroids c2	215	6	0.06	0.23	0.81	98.1 (3.3)	97.2 (4.0)	52.8	0.01
Thyroids c1	215	6	0.05	0.23	0.85	98.1 (3.3)	96.3 (4.3)	51.4	0.02
Iris c1	150	5	0.09	0.17	0.75	95.3 (4.3)	93.3 (5.9)	60.7	0.04
Wine c1	178	14	0.05	0.43	0.72	98.3 (3.7)	97.2 (5.1)	68.6	0.05
Wine c2	178	14	0.07	0.49	0.76	97.2 (4.3)	97.2 (4.3)	67.9	0.05
Thyroids c3	215	6	0.10	0.31	0.67	97.2 (4.0)	95.8 (4.4)	54.2	0.08
Iris c3	150	5	0.10	0.21	0.56	94.7 (5.8)	93.3 (6.6)	60.9	0.08
Wine c3	178	14	0.12	0.57	0.58	94.9 (5.2)	95.5 (4.9)	65.3	0.09

4.2 Relationship Between Data Complexity and SOMCBR

We establish a classification where the datasets are divided into the two types previously explained. Regarding type 2 datasets, the computational time does not improve in a great percentage (%$R < 70\%$) and the probability defined by the p-value is small (p-value $< 10\%$). Figure 2(a) shows the relationship between the p-value and the percentage of reduction of the computational time, %R, for all datasets tested. The perpendicular lines delimit the region of type 2 datasets.

Even so, the goal is to find a representation on the complexity space to distinguish between types 1 and 2. Thus, this should indicate *a priori* the applicability of SOMCBR according to the defined threshold values of p-value and %R. The

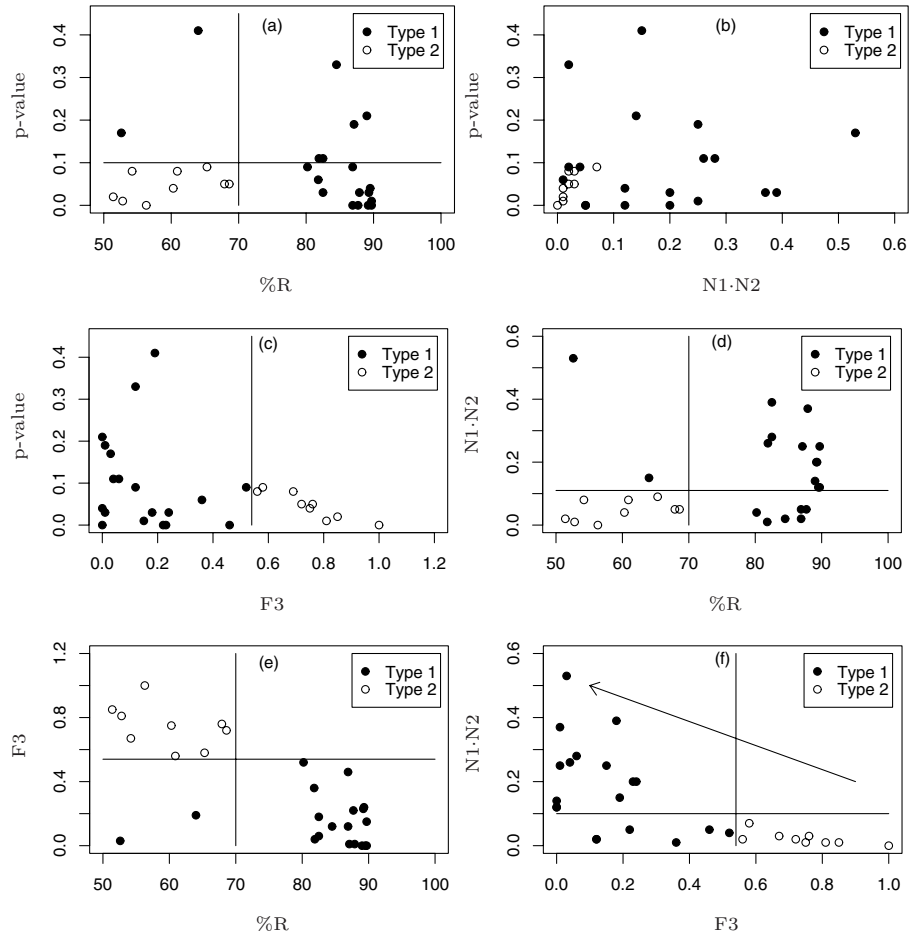

Fig. 2. The charts show how the combination of p-value, $\%R$, and complexity measures are useful tools for distinguishing the behaviour of type 1 and 2. The chart (f) allow us to predict the SOMCBR applicability.

F3, N1, and N2 complexity measures present interesting properties to distinguish between both types. Because the N1 and N2 separability measures have a similar behaviour, we can work with their product in order to promote extreme behaviours, especially in the case of datasets with low values.

Figures 2(b, c, d, e) depict the complexity measures (F3 and N1·N2) with the previously defined p-values and $\%R$. In Figure 2(b), we observe that all datasets of type 2 are near to the origin, with low values of N1·N2 and p-value, but there are some overlaps with the location of type 1 datasets. On the other hand, Figure 2(c) shows that type 2 problems are separated from type 1 with respect to F3. Moreover, type 2 problems are mainly related to high values of F3. Figures 2(d) and 2(e) show similar results, where we also plot the complexity measures and the

percentage of reduction $\%R$. On both figures, the values of F3 and N1·N2 define separated regions in two types.

These figures suggest some tendencies: (1) Datasets with high values of $\%R$ appear in regions with low values of F3. (2) Datasets with high values of F3 have very low values of p-value. (3) Low values of $\%R$ are slightly correlated with low values of N1 and N2 product.

Furthermore, Figure 2(f) represents a complexity space on N1·N2 and F3, where there are four possible situations. We can see how these measures settle ranges for all datasets belonging to type 2: high values (> 0.55) of F3 measure and very low values (< 0.1) of the mentioned N1·N2. A high value of F3 means a high separability of classes because the attributes are not overlapped. A low value of N1·N2 implies high linear separability. The arrow indicates the sense of data complexity. Thus, SOMCBR is recommendable for the rest of the complexity space represented in figure 2(f), that is, for complex domains.

Therefore, the *a priori* discrimination between type 1 and 2 with complexity measures allows us to obtain patterns of good performance of the SOMCBR without having to apply it.

5 Conclusions and Further Research

SOMCBR is a CBR characterized by the organization of the case memory by means of a SOM, which is responsible for grouping the cases into patterns. These patterns allow the retrieval phase to reduce its computational time because it only uses the cases associated with the most similar pattern instead of using the whole case memory. However, the solving capabilities can be compromised if the patterns are not well defined. This can happen in complex and noisy domains.

This paper is a first step in trying to relate the data topology and the SOMCBR application using complexity measures. These measures are based on estimating the problem hardness through the geometrical data structure. By the study of some graphical representations of the complexity space, we can conclude that the F3 measure and the product of N1 and N2 measures are useful to determine when the SOMCBR should be used, namely, for complex domains. Therefore, these complexity measures help us to predict *a priori* the performance of the SOMCBR without applying it.

Further work involves two issues. First, extending the analysis of the effects of other complexity measures and more datasets. Second, studying others ways of retrieving cases from SOMCBR in order to avoid losing useful cases if clusters are not well defined.

Acknowledgements. We would like to thank the Spanish Government for the support under grants TIN2006-15140-C03-03, TIN2005-08386-C05-04 and the *Generalitat de Catalunya* for the support under grants 2005SGR-302 and 2006FIC-0043. Also, we would like to thank *Enginyeria i Arquitectura La Salle* of Ramon Llull University for the support to our research group.

References

1. Aamodt, A., Plaza, E.: Case-based reasoning: Foundations issues, methodological variations, and system approaches. IA Communications 7, 39–59 (1994)
2. Wess, S., Althoff, K.D., Derwand, G.: Using k-d Trees to Improve the Retrieval Step in Case-Based Reasoning. In: Wess, S., Richter, M., Althoff, K.-D. (eds.) Selected papers from the First European Workshop on Topics in Case-Based Reasoning. LNCS, vol. 837, pp. 167–181. Springer, Heidelberg (1994)
3. Lenz, M., Burkhard, H.D., Brückner, S.: Applying Case Retrieval Nets to Diagnostic Tasks in Technical Domains. In: Smith, I., Faltings, B.V. (eds.) Advances in Case-Based Reasoning. LNCS, vol. 1168, pp. 219–233. Springer, Heidelberg (1996)
4. Vernet, D., Golobardes, E.: An Unsupervised Learning Approach for Case-Based Classifier Systems. Expert Update. The Specialist Group on Artificial Intelligence 6(2), 37–42 (2003)
5. Fornells, A., Golobardes, E., Vernet, D., Corral, G.: Unsupervised case memory organization: Analysing computational time and soft computing capabilities. In: Roth-Berghofer, T.R., Göker, M.H., Güvenir, H.A. (eds.) ECCBR 2006. LNCS (LNAI), vol. 4106, pp. 241–255. Springer, Heidelberg (2006)
6. Kohonen, T.: Self-Organization and Associative Memory. Series in Information Sciences, vol. 8. Springer, Heidelberg (1989)
7. Oja, M., Kaski, S., Kohonen, T.: Bibliography of Self-Organizing Map (SOM) Papers: 1998-2001. Neural Computing Surveys 3, 1–156 `http://www.cis.hut.fi/research/refs/` (2003)
8. Kaski, S., Kangas, J., Kohonen, T.: Bibliography of Self-Organizing Map (SOM) Papers: 1981-1997. Neural Computing Surveys 1, 102–350 `http://www.cis.hut.fi/research/refs/` (1998)
9. Fornells, A., Golobardes, E., Vilasís, X., Martí, J.: Integration of strategies based on relevance feedback into a tool for retrieval of mammographic images. In: Corchado, E., Yin, H., Botti, V., Fyfe, C. (eds.) IDEAL 2006. LNCS, vol. 4224, pp. 116–124. Springer, Heidelberg (2006)
10. Basu, M., Ho, T.K.: Data Complexity in Pattern Recognition. Springer, Heidelberg (2006)
11. Ho, T.K., Basu, M.: Complexity measures of supervised classification problems. IEEE Transaction on Pattern Analysis and Machine Intelligence 3(24), 289–300 (2002)
12. Bernadó-Mansilla, E., Ho, T.K.: Domain of competence of XCS classifier system in complexity measurement space. IEEE Transaction Evolutionary Computation 1(9), 82–104 (2005)
13. Blake, C.L., Merz, C.J.: UCI repository of machine learning databases (1998)
14. Sheskin, D.J.: Handbook of Parametric and Nonparametric Statistical Procedures. CRC Press, Boca Raton (1997)