ESANN'2008 proceedings, European Symposium on
Artificial Neural Networks - Advances in Computational Intelligence and Learning
Bruges (Belgium), 23-25 April 2008, d-side publi., ISBN 2-930307-08-0.

SOM based clustering with instance-level constraints

Fazia Bellal[1], Khalid Benabdeslem[1] and Alexandre Aussem[1]

1- University of Lyon1, LIESP
8, Avenue Niels Bohr, 69622 Villeurbanne cedex, France

Abstract.
This paper describes a new topological map dedicated to clustering under instance-level constraints. In general, traditional clustering is used in an unsupervised manner. However, in some cases, background information about the problem domain is available or imposed in the form of constraints, in addition to data instances. In this context, we modify the popular SOM algorithm to take these constraints into account during the construction of the topology. We present experiments on synthetic known databases with artificial constraints. We then apply the new method to a real problem of clustering melanoma data in health domain.

1 Introduction

Knowledge extraction by data mining often involves unsupervised clustering processes so as to optimize the organizing of data sets in regard of their similarities. However, these algorithms only access to variables which describe each data but they do not deal with any other kind of given information. Nevertheless, taking *a priori* knowledge into account in such algorithms, if there exists, is an important problem concerning at the same time the expression, structuration and formalisation of knowledge in view of its integration in automatic clustering processes. The first work on this area was presented in [6] based on the modification of the COBWEB algorithm proposed in [3]. Furthermore, the same authors have proposed an another approach which integrates constraints in the K-means algorithm [7]. Their COP-k-means algorithm attempts to find a set partition that minimizes the vector quantization error and satisfies all constraints at the same time. Moreover, in [9], an exploration of the use of instance and cluster-level constraints was performed with agglomerative hierarchical clustering. In [2] it is shown improvements in quality and computational complexity of clustering by using constraints in a graph b-coloring clustering algorithm. The authors empirically illustrate the benefits from using constraints to improve cluster purity and average distortion. Alternatively, in this paper, we propose a new version of Kohonen's algorithm based on the control of neurons respecting or violating some given constraints on patterns during the construction of the topological map. Thereby, we will compare our method to all cited constrained-clustering based methods.

ESANN'2008 proceedings, European Symposium on
Artificial Neural Networks - Advances in Computational Intelligence and Learning
Bruges (Belgium), 23-25 April 2008, d-side publi., ISBN 2-930307-08-0.

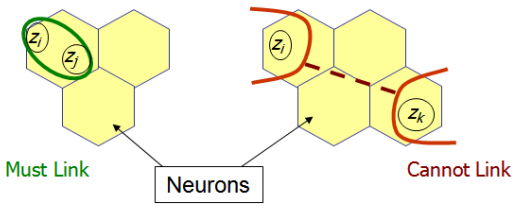

Fig. 1: Must-Link and Cannot-Link constraints

2 Constrained Topological Map

2.1 Notations

Let us assume that a set of N samples or patterns $z_i = (z_{i1}, \ldots, z_{id}) \in \mathbb{R}^d$ is given where d is the number of variables. The learning database formed by these samples is denoted by $A = \{z_i \in \mathbb{R}^d,\ i = 1, \ldots, N\}$. We also suppose that we have an undirected graph $G = (V, E)$ where V is a set of vertices and E a set of edges. This graph represents a neural network by considering that each vertex $v \in V$ is a unit or neuron. The number of neurons in V is denoted by C. The distance $\delta(c, r)$ between two neurons $c, r \in V$ is the length of the minimal path linking c and r in G. In order to take into account the influence of a neuron r on a neuron c, we introduce a symmetric kernel function $\mathcal{K} : \mathbb{R} \to \mathbb{R}_+$ satisfying

$$\lim_{|x| \to \infty} \mathcal{K}(x) = 0, \tag{1}$$

so that the mutual influence between r and c is $\mathcal{K}(\delta(c, r))$. The neighborhood set around neuron c is denoted $Neigbor(c)$. It contains all nodes up to a certain radius T in the grid from node c. For each neuron $c \in V$, a reference vector $w_c^t = (w_{c1}^t, \ldots, w_{cd}^t) \in \mathbb{R}^d$ is defined where $0 \leq t \leq N_{iter}$, N_{iter} being the maximum number of iterations in Kohonen's algorithm [4].

2.2 Generation of constraints

In some situations, much unlabeled data and a little labeled data are available in the dataset. Thereby, we express this knowledge in the form of two types of constraints (see fig. 1) :

- ***Must-Link* constraints**, denoted by $Con_{=}(z_i, z_j)$ indicate that two patterns z_i and z_j must belong to the same neuron,
- ***Cannot-Link* constraints**, denoted by $Con_{\neq}(z_i, z_j)$ indicate that two patterns z_i and z_j cannot belong to the same neuron.

It appears that *Must-Link* constraints are transitive, i.e., for any patterns $(z_i, z_j, z_k) \in V$, if $Con_{=}(z_i, z_j)$ and $Con_{=}(z_j, z_k)$ are true, then it follows that $Con_{=}(z_i, z_k)$ must also be satisfied. Similarly, if $Con_{=}(z_i, z_j)$, $Con_{=}(z_k, z_l)$ and $Con_{\neq}(z_i, z_k)$ are true, then $Con_{\neq}(z_i, z_l)$ and $Con_{\neq}(z_j, z_k)$ are also satisfied.

ESANN'2008 proceedings, European Symposium on
Artificial Neural Networks - Advances in Computational Intelligence and Learning
Bruges (Belgium), 23-25 April 2008, d-side publi., ISBN 2-930307-08-0.

Algorithm 1 RespectCon$(A, V, i, Con_{=}, Con_{\neq})$

$V_i^{viol} = \emptyset$
for all $c \in V$ **do**
  for all z_j such that $Con_{=}(z_i, z_j)$ **do**
    if $z_j \notin c$ **then** $V_i^{viol} = V_i^{viol} \cup \{c\}$
  end for
  for all z_k such that $Con_{\neq}(z_i, z_k)$ **do**
    if $z_k \in c$ **then** $V_i^{viol} = V_i^{viol} \cup \{c, Neigbor(c)\}$
  end for
end for
$V_i^{resp} = V \backslash V_i^{viol}$
return V_i^{resp}

We define $V_i^{resp} \subset V$ as the set of all neurons in V satisfying constraints on a given pattern $z_i \in A$. In the same way, $V_{viol} \subset V$ is the set of all neurons in violation with $z_i \in A$. In the case of *Cannot-Link* constraints, the idea is to clamp neurons which are located in the neighborhood set around the neuron violating these constraints. The computation of V_i^{resp} is performed by removing from V neurons which violate constraints on z_i as described in algorithm 1. Patterns z_i which over-constrain the problem are omitted.

The computation of sets V_i^{resp} for all patterns $z_i \in A$ is the key of our approach since it will allow the integration of constraints in Kohonen's algorithm.

2.3 The constrained algorithm

Let us assume that $t > 0$ and that a pattern z_i has been randomly chosen in A. Contrary to Kohonen's approach [4], the best-matching unit c^* is now sought in the subset $V_i^{resp} \subset V$ computed from algorithm 1. In other words, we have:

$$c^* = \arg \min_{c \in V_i^{resp}} \|z_i - w_c^t\|^2. \tag{2}$$

Moreover, only neurons influenced by constraints on z_i may be corrected after the computation of c^*. In other words, the stochastic gradient based correction scheme is now restricted to V_i^{resp}:

$$w_c^t = w_c^{t-1} - \mu^t \mathcal{K}^T(\delta(c, c^*))(w_c^{t-1} - z_i), \quad \forall c \in V_i^{resp}, \tag{3}$$

where μ^t is the learning rate which is a decreasing function of time. These modifications lead to a new algorithm called *CrTM*.

3 Experimental results

3.1 Evaluation metrics

For the evaluation of the efficiency of the CrTM algorithm, we propose to use the Rand index [5]. This index measures the correspondance between two partitions

Algorithm 2 CrTM(A,V,$Con_{=}$,$Con_{\neq}$,N_{iter})

1: Set $t = 0$ and initialize vectors w_c^t to random values for all $c \in V$.
2: **for** $t = 1, \ldots, N_{iter}$ **do**
3: Select a random pattern $z_i \in A$
4: Compute $V_i^{resp} = RespectCon(A, V, i, Con_{=}, Con{\neq})$
5: Find $c^* \in V_i^{resp}$ as defined in (2).
6: Update reference vectors using (3).
7: **end for**

P_1 and P_2 of a data set A. In our case, P_1 is the correct partition produced by labels of predefined classes and P_2 is the partition obtained from the $CrTM$ algorithm. Each partition is regarded as a set of $N(N-1)/2$ pairs of decisions. For each pair of points (z_i, z_j), P_i assigns them to the same class or to two different classes. Assuming a is the number of decisions where z_i belongs to the same class as z_j in P_1 and P_2 and b is the number of decisions where z_i and z_j do not belong to the same class in P_1 and P_2, we obtain $(a+b)$ correct decisions and the *overall accuracy* between P_1 and P_2 is:

$$Acc = Rand(P_1, P_2) = \frac{a+b}{N(N-1)/2}. \quad (4)$$

We also show that knowledge brought by constraints may even improve the performance of classifiers on patterns which are not constrained. Then, we compute aside *overall accuracy*, the one on a *held-out test set* which is a subset of data set composed of instances that are not directly or transitively affected by the constraints. This represents a real learning performance measure since such a *Held-Out* improvement reveals if the algorithm managed to learn constraints and to generalize this type of knowledge so as to influence the classification of unconstrained patterns.

3.2 Results obtained with artificial constraints

We propose to evaluate the CrTM algorithm on two labeled databases issued from the UCI bank [1]: *Tic-tac-toe* and *Heart disease*. To generate artificial constraints, we randomly select two instances from the data set and check their labels. If they have the same label, we create a *Must-Link* constraint, otherwise, a *Cannot-Link* constraint. We compare our algorithm using results obtained on these databases to three other constrained methods. We use a random initialization, a gaussian neighborhood function and default values for the learning parameters. The database *Tic-tac-toe* contains 100 patterns described by 9 qualitative variables and 2 output classes. Without any constraint, the overall accuracy obtained with Kohonen's algorithm is 66.6%. Defining 500 random constraints, we obtain 96.7% for overall accuracy and 91.4% for held-out as illustrated in figure 2. With the same number of constraints, the results obtained with COP-COPWEB, COP-Kmeans and COP-b-coloring are 49%, 56%

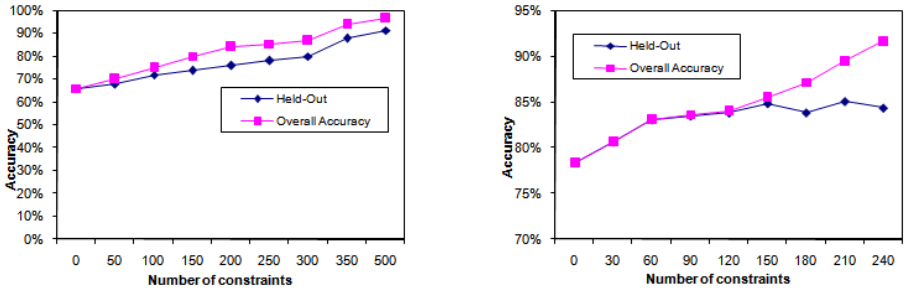

Fig. 2: Results obtained with the CrTM algorithm on *Tic-tac-toe* (left) and *Heart disease* (right)

and 82% respectively [6, 7, 2]. Thus, the CrTM algorithm yield a significant improvement. The database *Heart disease* contains 303 patterns described by 5 numerical variables and 8 qualitative variables. The instances were also classified into two classes. Without any constraint, the overall accuracy for Kohonen's algorithm is 78%. Using 240 random constraints, the overall accuracy for CrTM rises to 91% (see fig. 2). Using 150 random constraints, held-out reaches 85%. In comparison, the overall accuracy for COP-b-coloring is 50% without constraints and the held out is 66% with 500 constraints. On the one hand, we observe that constraints always improve the quality of the partition, and on the other hand we note that our algorithm gives a better clustering than others with a smaller number of constraints.

3.3 Application to real data : Melanoma cancer

Melanoma is a malignant tumor of melanocytes which are found predominantly in skin but also in the bowel and the eye. The melanoma base used in this study, was kindly supplied by *VISOON* company, Lyon, France. It contains 226 pictures of melanoma. Each picture is described by the "ABCD rule"[8] which is a widely accepted tool to promote early detection of melanoma: **A**symmetry, **B**order irregularity, **C**olor variegation, and **D**iameter greater than 6 mm. Doctors labelled some picture of data, that is to say these patterns are associated with an output class concerning the disease statement (either healthy or with heart-disease). Therefore, we generate real constraints imposed by experts. The overall accuracy obtained with Kohonen's algorithm is 78.2% without constraints. Adding 210 constraints to the problem, the overall accuracy for CrTM is 84%, and the held-out is 81.2% as illustrated in fig.3.

4 Conclusion

In this paper we proposed a general method for incorporating background knowledge during the construction process of self-organizing maps. We transform background knowledge, from partially labeled data or a priori knowledge on the

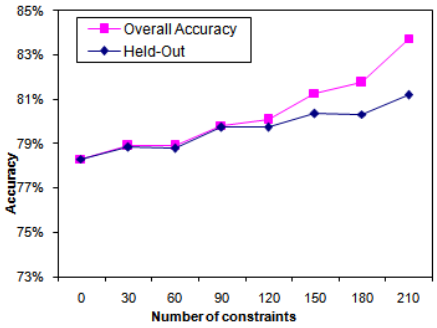

Fig. 3: Performance of CrTM on melanoma data

domain of real data sets, in the form of instance-level constraints. The propose algorithm deals with binary (between two pair of instances) and deterministic (belonging or not to the same cluster) constraints. Our experimental results indicate that only a small amount of constraints is necessary to significantly improve the quality of the clustering. Furthermore, the empirical study we have carried out tends to demonstrate that prior information may be applied in a real domain. The future of our work will concern the extension of our method to more complex constraints such as probabilistic or conditional constraints and the optimization of the CrTM algorithm using a hierarchical clustering.

References

[1] C. Blake, C. Merz, Uci repository of machine learning databases. Technical Report, University of California, 1998.

[2] H. Elghazel, K. Benabdelslem, A. Dussauchoy, Constrained graph b-coloring based clustering approach, (DaWaK), *LNCS N°4654*, pages 262-271, Regensburg (Germany) 2007.

[3] D. Fisher, Knowledge acquisition via incremental conceptual clustering, *Machine Learning*, 2:139-172, Kluwer Academic Publishers, 1987.

[4] T. Kohonen, *Self-Organizing Maps*, Springer, Berlin, 1994.

[5] W. M. Rand, Objective criteria for the evaluation of clustering method, *Journal of the American Statistical Association*, 66:846-850, 1971.

[6] K. Wagstaff and C. Cardie, Clustering with instance-level constraints, *ICML*, pages 1103-1110, 2000.

[7] K. Wagstaff, C. Cardie, S. Rogers and S. Schrödl, Constrained k-means clustering with backgroung knowledge , *ICML*, pages 577-584, 2001.

[8] Stolz, W. Riemann, A .Abcd rule of dermatoscopy:A new practical method for early recognition of malignant melanoma. *Eur J Dermatol 4*, pages 521-527, 1994.

[9] I. Davidson, S.S. Ravi Agglomerative hierarchical clustering with constraints:theorical and empirical results *Proceedings PKDD 2005*, 3721:59-70, 2005.