Semi-Supervised Vector Quantization for proximity data

Xibin Zhu Frank-Michael Schleif
Barbara Hammer

CITEC - Centre of Excellence,
Bielefeld University, 33615 Bielefeld, Germany

Abstract. Semi-supervised learning (SSL) is focused on learning from labeled *and* unlabeled data by incorporating structural and statistical information of the available unlabeled data. The amount of data is dramatically increasing, but few of them are fully labeled, due to cost and time constraints. This is even more challenging for non-vectorial, proximity data, given by pairwise proximity values. Only few methods provide SSL for this data, limited to positive-semi-definite (psd) data. They also lack interpretable models, which is a relevant aspect in life-sciences where most of these data are found. This paper provides a prototype based SSL approach for proximity data.

1 Introduction

Large data sets are more and more common but the annotation of these data is often time consuming and costly. Accordingly, many of them are only partially labeled. Semi-supervised learning (SSL, [1, 2]) integrates the structural and statistical knowledge of unlabeled data in the training process of a classifier, rather to ignore unlabeled points as it is still very common. A variety of methods on SSL has been published [1]. They all focus on vectorial data sets, often binary-class data, or in case of kernel-approaches like the Semi-Supervised SVM (S3VM) (see e.g. [2]) on psd kernel matrices.

Another very relevant source of partially labeled data, are proximity data, which however are not much considered in the literature as an SSL problem. Proximity, (dis-)similarity or relational data sets, are based on pairwise comparisons of objects providing score-values of the proximity of the objects. A vector space is not necessarily available for such data and the score-functions need not to be metric. In fact, many domain specific proximity measures are of this type, with the most prominent example of alignment algorithms, like the Smith-Waterman algorithm for sequence data [3]. Multiple methods for relational learning were published (see work based on [4]) but view for SSL problems. In this paper we propose a sparse SSL algorithm, directly applicable on non-psd proximity multi-class data.

We review relational supervised prototype learning (RPC) as introduced by the authors earlier and a specific model employing conformal prediction as proposed recently in [5] called C-RPC. Thereafter we introduce an extension of C-RPC to semi-supervised learning (SC-RPC). We show the effectiveness on real life data for known vectorial data sets and biomedical dissimilarity data. Finally we summarize our results and discuss potential extensions.

2 Preliminaries about dissimilarity data

Let $\mathbf{v}_j \in \mathbb{V}$ be a set of objects defined in some data space, with $|\mathbb{V}| = N$. We assume, there exists a dissimilarity measure such that $D \in \mathbb{R}^{N \times N}$ is a dissimilarity matrix measuring the pairwise dissimilarities $D_{ij} = d(\mathbf{v}_i, \mathbf{v}_j)$ between all pairs $(\mathbf{v}_i, \mathbf{v}_j) \in \mathbb{V}$. Any reasonable (possibly non-metric) distance measure is sufficient. We assume zero diagonal $d(\mathbf{v}_i, \mathbf{v}_i) = 0$ for all i and symmetry $d(\mathbf{v}_i, \mathbf{v}_j) = d(\mathbf{v}_j, \mathbf{v}_i)$ for all i, j.

3 Relational prototype based learning

We assume a training set is given where data point $\mathbf{v}_j$ is labeled $\mathbf{l}_j \in \mathbb{L}, |\mathbb{L}| = L$. The objective is to learn a classifier f such that $f(\mathbf{v}_k) = \mathbf{l}_k$ for any given data point. Thereby, $\mathbf{v}_k$ is represented implicitly by a vector of known dissimilarities with respect to $W \subseteq \mathbb{V}$. In [6] the authors proposed a relational prototype classifier (RPC) used as the basic method in this article.

Classification takes place by means of k prototypes $\mathbf{w}_j$ in the pseudo-Euclidean space, which are priorly labeled. Typically, a winner takes all rule is assumed, i.e. a data point is mapped to the label assigned to the prototype which is closest to the data in pseudo-Euclidean space using Eq. (1). For relational data classification, the key assumption is to restrict prototype positions to linear combinations of data points of the form $\mathbf{w}_j = \sum_i \alpha_{ji} \mathbf{v}_i$ with $\sum_i \alpha_{ji} = 1$. Then dissimilarities can be computed implicitly by means of

$$d(\mathbf{v}_i, \mathbf{w}_j) = [D \cdot \alpha_j]_i - \frac{1}{2} \cdot \alpha_j^t D \alpha_j \tag{1}$$

where $\alpha_j = (\alpha_{j1}, \ldots, \alpha_{jn})$ refers to the vector of coefficients describing the prototype $\mathbf{w}_j$ implicitly, as shown in [7]. The original underlying cost function [8] was adapted in [6] for relational learning and becomes:

$$E_{\mathrm{RPC}} = \sum_i \Phi \left(\frac{[D\alpha^+]_i - \frac{1}{2} \cdot (\alpha^+)^t D \alpha^+ - [D\alpha^-]_i + \frac{1}{2} \cdot (\alpha^-)^t D \alpha^-}{[D\alpha^+]_i - \frac{1}{2} \cdot (\alpha^+)^t D \alpha^+ + [D\alpha^-]_i - \frac{1}{2} \cdot (\alpha^-)^t D \alpha^-} \right), \tag{2}$$

where the closest correct and wrong prototypes are referred to, $\mathbf{w}^+$ and $\mathbf{w}^-$, respectively, corresponding to the coefficients α^+ and $\alpha-$, respectively and $\Phi(x) = (1 + \exp(-x))^{-1}$. A simple stochastic gradient descent leads to adaptation rules for the coefficients α^+ and α^- [6]. After every adaptation step, normalization takes place to guarantee $\sum_i \alpha_{ji} = 1$. The prototypes are initialized as random vectors corresponding to random values α_{ij} which sum to one. Out-of-sample extension of the classification to new data is done as shown in [7].

3.1 Semi-supervised conformal prediction for RPC (SC-RPC)

Let (T1) denote the labeled training data with $\mathbf{z}_i = (\mathbf{v}_i, \mathbf{l}_i) \in \mathbb{Z} = \mathbb{V} \times \mathbb{L}$. Furthermore let $\mathbf{v}_{N+1}$ be a new data point with unknown label . The *conformal prediction* computes for given training data $(\mathbf{z}_i)_{i=1,\ldots,N}$, an observed data point $\mathbf{v}_{N+1}$, and a chosen error rate ϵ an $(1-\epsilon)$-*prediction region* $\Gamma^\epsilon(\mathbf{z}_1, \ldots, \mathbf{z}_l, \mathbf{v}_{N+1}) \subseteq \mathbb{L}$ consisting of a number of possible label assignments.

3.1.1 Computation of the prediction region

To compute the conformal prediction region, a non conformity measure is fixed $A(\mathcal{D}, \mathbf{z})$. It is used to calculate a non conformity value μ that estimates how an observation $\mathbf{z}$ fits to given representative data $\mathcal{D}=\{\mathbf{z}_1, \ldots, \mathbf{z}_N\}$. The conformal algorithm for classification is as follows: given a nonconformity measure A, significance level ϵ, examples $\mathbf{z}_1, \ldots, \mathbf{z}_N$, object $\mathbf{v}_{N+1}$ and label $\mathbf{l}$, it is decided whether $\mathbf{l}$ is contained in $\Gamma^\epsilon(\mathbf{z}_1, \ldots, \mathbf{z}_N, \mathbf{v}_{N+1})$:

$$
\begin{aligned}
&\text{set } \mathbf{z}_{N+1} := (\mathbf{x}_{N+1}, \mathbf{l}) \\
&\text{for } i = 1, \ldots, N+1 \text{ set} \\
&\qquad \mu_i := A(\{\mathbf{z}_1, \ldots, \mathbf{z}_{N+1}\} \backslash \{\mathbf{z}_i\}, \mathbf{z}_i) \\
&\text{set } r_l := \frac{|\{i = 1, \ldots, N+1 \mid \mu_i \geq \mu_{N+1}\}|}{N+1} \qquad \text{include } \mathbf{l} \text{ if } r_l > \epsilon
\end{aligned}
$$

Given $\mathbf{z} = (\mathbf{x}_i, \mathbf{l})$ and a trained relational prototype model W, we choose

$$
\mu_i := \frac{d^+(\mathbf{x_i})}{d^-(\mathbf{x_i})} \tag{3}
$$

with $d^+(\mathbf{x}_i)$ being the distance between $\mathbf{x}_i$ and the closest prototype labeled $\mathbf{l}$, and $d^-(\mathbf{x}_i)$ being the distance between $\mathbf{x}_i$ and the closest prototype labeled differently than $\mathbf{l}$ where distances are computed according to Eq. (1)[1]

3.1.2 Confidence and credibility

The prediction region $\Gamma^\epsilon(\mathbf{z}_1, \ldots, \mathbf{z_N}, \mathbf{v}_{N+1})$ is the core of conformal prediction. For a given error rate ϵ it contains the possible labels of $\mathbb{L}$ that ensure low error ϵ. But how can we use it for prediction?

Suppose we use a meaningful non conformity measure A. If the value ϵ is approaching 0, a conformal prediction with almost no errors is required, which can only be satisfied if the prediction region contains all possible labels. If we raise ϵ we allow errors to occur and as a benefit the conformal prediction algorithm excludes unlikely labels from our prediction region, increasing its information content. In detail those $\mathbf{l}$ are discarded for which the r-value is less or equal ϵ. Hence only a few $\mathbf{z}_i$ are as non conformal as $\mathbf{z}_{N+1} = (\mathbf{v}_{N+1}, \mathbf{l})$. This is a strong indicator that $\mathbf{z}_{N+1}$ does not belong to the distribution $\mathbb{Z}$ and so $\mathbf{l}$ seems not to be the right label. If one further raises ϵ only those $\mathbf{l}$ remain in the conformal region that can produce a high r-value meaning that the corresponding $\mathbf{z}_{N+1}$ is rated as very typical by A.

So one can trade error rate against information content. The most useful prediction is those containing exactly one label. Therefore, given an input $\mathbf{l}_i$ two error rates are of particular interest, ϵ_1^i being the smallest ϵ and ϵ_2^i being the greatest ϵ so that $|\Gamma^\epsilon(\mathcal{D}, \mathbf{v}_i)| = 1$. ϵ_2^i is the r-value of the best and ϵ_1^i is the r-value of the second best label. Thus, typically, a conformal predictor outputs the label $\mathbf{l}$ which describes the prediction region for such choices ϵ, i.e. $\Gamma^\epsilon = \{\mathbf{l}\}$, and the classification is accompanied by the two measures

[1] See [5] for a detailed discussion about this non-conformity measure and its validity.

$$\text{confidence} : 1 - \epsilon_1^i = 1 - r_{y_{2\text{nd}}} \qquad \text{credibility} \ : \epsilon_2^i = r_{y_{1\text{st}}} \tag{4}$$

Confidence says something about being sure that the second best label and all worse ones are wrong. Credibility says something about to be sure that the best label is right respectively that the data point is (a)typical and not an outlier.

3.1.3 Model complexity and SSL in C-RPC

We use the additional information provided by a conformal relational prototype classifier to automatically adapt the complexity of the model, i.e. the number of prototypes. We assume that a larger amount of the data is unlabeled which we denote as T2, the training set is denoted as T1 to train the model, while T2 is used to estimate the suitability of the current model by means of conformal prediction. For this subset, we compute μ-values according to (3). This provides point estimates for confidence and credibility of the classifier. We collect the set of points $\mathcal{B}$ with low credibility and/or confidence. A low confidence is given if $(1 - \epsilon_1^i) \leq (1 - \frac{1}{L})$ and a low credibility is observed for $\epsilon_2^i \leq \frac{1}{L}$. Hence we define

$$\mathcal{B} = \left\{ \mathbf{v}_i \in T2 : \left(1 - \epsilon_1^i\right) \leq \left(1 - \frac{1}{L}\right) \vee \epsilon_2^i \leq \frac{1}{L} \right\} \tag{5}$$

If $|\mathcal{B}|$ is large, in our case we take the boundary ≥ 5, the complexity of the classifier is not yet sufficient. Hence, this parameter controls the sparsity of the model. We found by some independent experiments on simulated data, that $|\mathcal{B}| = 5$ is a good compromise between too dense $|\mathcal{B}| =\leq 5$ or to sparse models $|\mathcal{B}| \gg 5$. A new prototype is created and set to the representative data point (median) in $\mathcal{B}$, it is labeled according to the label of the nearest neighbor from set T1. The detailed, commented algorithm is shown in Alg. 1.

The C-RPC and SC-RPC approach scale quadratic in the number of training examples and the size of prototype representations scales in $\mathcal{O}(N)$. Kernel approaches need to calculate a valid kernel from the dissimilarities with $\mathcal{O}(N^2) - \mathcal{O}(N^3)$ and show often very dense models. Also kNN scales in $\mathcal{O}(N^2)$ regarding the runtime. So all standard methods are quite costly whereas our approach provides *sparse,interpretable* models which can be trained in reasonable time and keep good generalization and query time for the test set, permitting pointwise measures of confidence.

4 Experiments

We evaluate SC-RPC (Alg-1) and C-RPC (Alg-2) on a larger range of tasks including, vectorial SSL benchmark data sets[2], i.e. Digit1, USPS, G241c, G241n, COIL and five well known UCI data sets[3]. Dissimilarity matrices D have been generated by using the squared-Euclidean distance. Further we evaluate our

[2] `http://www.kyb.tuebingen.mpg.de/ssl-book`

[3] `http://archive.ics.uci.edu/ml/datasets.html`

Algorithm 1 Semi-Supervised C-RPC (SC-RPC)

init: $credi\_threshold := \frac{1}{L}$, $confi\_threshold := 1 - \frac{1}{L}$; $W := \emptyset$;
$\mathcal{B} := \emptyset$; $improve := 1\%$;
$max\_itr := 100$ ▷ maximal total iterations
$max\_ctn\_best := 10$ ▷ maximal iterations for a result as winner
$W :=$ train $T1$ by RPC ; $W\_Best = W$;
$acc :=$ evaluation of W; ▷ accuracy w.r.t. $T1$
$A\_T1 := \{\mu_i, \forall i \in T1\}$,$A\_T2 := \{\mu_i, \forall i \in T2\}$ ▷ μ-values of $T1, T2$: eq. (3)
generate $\mathcal{B}$ ▷ eq. (5)
while $|\mathcal{B}| \geq 5$ & $itr < max\_itr$ & $ctn\_best \leq max\_ctn\_best$ **do**
  $W := W \bigcup \{$new prototype(s) from $\mathcal{B}\}$
  $W :=$ train $T1$ by RPC given W; ▷ training with given prototypes
  $acc\_new :=$ evaluation of W; ▷ new accuracy
  $A\_T1 := \{\mu_i, \forall i \in T1\}$; $A\_T2 := \{\mu_i, \forall i \in T2\}$ ▷ μ-values acc to eq. (3)
  $Confi := \{1 - \epsilon_1^i, \forall i \in T2\}$; $Credi := \{\epsilon_2^i, \forall i \in T2\}$; ▷ eq. (4)
  generate $\mathcal{B}$,
  if $acc\_new - acc \geq improve$ **then**
    $W\_Best = W$; $acc = acc\_new$; $ctn\_best = 0$;
  else $ctn\_best = ctn\_best + 1$;
  end if
end while
return W_Best;

data on two relational data sets, where no direct vector embedding exists and the data are given as (dis-)similarities. The *SwissProt* data set (SWISS) consists of $5,791$ samples of protein sequences in 10 classes taken as a subset from the popular SwissProt database of protein sequences [9] (release 37). The 10 most common classes such as Globin, Cytochrome b, etc. provided by the Prosite labeling where taken. Sequences are compared using Smith-Waterman alignment [3]. The *Copenhagen Chromosomes* data (CHROMO) constitute a benchmark from cytogenetics [10]. 4,200 human chromosomes from 21 classes, represented by grey-valued images and encoded as strings measuring the thickness of their silhouettes. These strings can directly be compared using the edit distance with insertion/deletion costs 4.5 [10]. We randomly select 100 examples of the data to be used as labeled examples, and use the remaining data as unlabeled data. The experiments are repeated for 30 times and the average test-set accuracy (on the unlabeled data) and standard deviation are recorded. Both algorithms have been initialized with 1 prototype per class, selected randomly from the labeled data set. The results are shown in Table 1. In all but three cases, semisupervised learning improves the result.

5 Conclusions

We proposed an extension of C-RPC for semi-supervised learning. It is a natural multi-class semi-supervised learner for vectorial and non-vectorial data sets. Our experiments show that the approach shows in general superior results compared to standard C-RPC learning based on the labeled data alone. In future work

	Diabetes	German	Haberman	House	WDBC
Alg-1	**71.00** (2.6)	69.7 (0.70)	**73.3** (2.6)	**90.4** (1.7)	**93.3** (1.3)
Alg-2	69.10 (3.5)	**70.0** (0.52)	71.7 (4.4)	89.17 (1.07)	92.2 (1.5)
	Digit1	USPS	G241c	G241n	COIL
Alg-1	**93.21** (2.6)	80.00 (0.01)	**73.1** (5.3)	69.70 (4.7)	**64.7** (9.0)
Alg-2	82.42 (9.6)	79.95 (0.27)	72.31 (5.13)	69.89 (3.65)	56.75 (4.63)
	SwissProt	CHROMO			
Alg-1	**84.04**(2.51)	**80.70**(3.11)			
Alg-2	82.57(3.51)	79.44(2.62)			

Table 1: Classification results for different vectorial and non-vectorial data.

we will explore SC-RPC for non-i.i.d. labeled data and approach large scale problems using techniques discussed in [6][4]

References

[1] O. Chapelle, B. Schölkopf, and A. Zien, editors. *Semi-Supervised Learning*. MIT Press, Cambridge, MA, 2006.

[2] Xiaojin Zhu and Andrew B. Goldberg. Introduction to semi-supervised learning. *Synthesis Lect. on Artif. Int. and Mach. Learning*, 3(1):1–130, 2009.

[3] Dan Gusfield. *Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology*. Cambridge University Press, 1997.

[4] E. Pekalska and R. Duin. *The dissimilarity representation for pattern recognition*. World Scientific, 2005.

[5] F.-M. Schleif, X. Zhu, and B. Hammer. A conformal classifier for dissimilarity data. In *Proceedings of AIAI 2012*, page press, 2012.

[6] A. Gisbrecht, B. Mokbel, F.-M. Schleif, X. Zhu, and B. Hammer. Linear time relational prototype based learning. *J. of Neural Sys.*, page press, 2012.

[7] Barbara Hammer and Alexander Hasenfuss. Topographic mapping of large dissimilarity data sets. *Neural Computation*, 22(9):2229–2284, 2010.

[8] Atsushi Sato and Keiji Yamada. Generalized learning vector quantization. In David S. Touretzky, Michael Mozer, and Michael E. Hasselmo, editors, *NIPS*, pages 423–429. MIT Press, 1995.

[9] B. Boeckmann, A. Bairoch, R. Apweiler, M.-C. Blatter, A. Estreicher, E. Gasteiger, M.J. Martin, K. Michoud, C. O'Donovan, I. Phan, S. Pilbout, and M. Schneider. The swiss-prot protein knowledgebase and its supplement trembl in 2003,. *Nucleic Acids Research*, 31:365–370.

[10] M. Neuhaus and H. Bunke. Edit distance based kernel functions for structural pattern classification. *Pattern Recognition*, 39(10):1852–1863, 2006.

[4] **Acknowledgments**: F.-M. Schleif was supported by the "German Sc. Found. (DFG)" (HA-2719/4-1). Financial support from the Cluster of Excellence 277 Cognitive Interaction Technology funded by the German Excellence Initiative is gratefully acknowledged.