Pattern Recognition Letters
journal homepage: www.elsevier.com

Adaptive Conformal Semi-Supervised Vector Quantization for Dissimilarity Data

Xibin Zhu[a,**], Frank-Michael Schleif[b], Barbara Hammer[a]

[a] *Bielefeld University, Center of Excellence, Inspiration 1, 33619 Bielefeld, Germany*
[b] *School of Computer Science, University of Birmingham, Birmingham B15 2TT, UK*

ARTICLE INFO

Article history:

Semi-Supervised Learning
Proximity Data
Dissimilarity Data
Conformal Prediction
Generalized Learning Vector Quantization

ABSTRACT

Most existing semi-supervised learning (SSL) algorithms focus on vectorial data given in Euclidean space or representations by means of valid kernel matrices. A lot of real life data, especially in bioinformatics domain, are non–metric given in the form of (dis-)similarities. Those data are not widely addressed in the SSL domain. In this paper we extend a prototype-based classifier for dissimilarity data to semi-supervised tasks employing *conformal prediction* providing point-wise confidence measures about the classification. By means of the confidence values a so-called 'secure region' of unlabeled data can be identified and further used to improve the trained model based on labeled data while adapting the model complexity to 'cover' a so-called 'insecure region' of labeled data. This way an intuitive semi-supervised multi-class classification scheme results which can (i) directly deal with arbitrary symmetric dissimilarity matrices, (ii) which offers intuitive classification by means of sparse prototypical class representatives, and (iii) which adapts model complexity supported by a confidence measure. In the experiments we show its effectiveness on simulated dissimilarity data and compare it with state-of-the-art methods on benchmarks from SSL domain and real-life non-vectorial data sets.

© 2014 Elsevier Ltd. All rights reserved.

1. Introduction

Big data is getting more and more challenging regarding storage and analysis requirements. Due to the sheer amount of data, only few of these data are completely labeled, and labeling of all data is indeed very costly and time consuming. Accordingly many data sets, in life sciences for example, are only partially labeled. Techniques of data mining, visualization, and machine learning are necessary to help people to analyze those data. Especially semi-supervised learning (SSL) techniques are widely used for this setting. The idea of semi-supervised learning is to learn the model not only from the labeled training data, but to also incorporate structural and statistical information in additionally available unlabeled data. A variety of SSL methods has been published (Chapelle et al., 2006; Zhu and Goldberg, 2009). Most of them focus on vectorial data given in Euclidean space or representations by means of positive semi-definite (psd) kernel matrices.

A lot of real world data, like biological sequences, are non-vectorial, often non-Euclidean and given in the form of pairwise proximities, which are based on pairwise comparisons of objects providing some score-value of the (dis-)similarity of the objects. Those data are also referred to as *proximity* or *relational data*. An underlying vector space is not necessarily available and there is no guarantee of metric conditions. Examples of those proximity or (dis-)similarity measures are edit distance based measures for strings or images (Haasdonk and Bahlmann, 2004) or popular similarity measures in bioinformatics such as scores obtained by the Smith-

**Corresponding author
e-mail: xzhu@techfak.uni-bielefeld.de (Xibin Zhu), schleify@cs.bham.ac.uk (Frank-Michael Schleif), bhammer@techfak.uni-bielefeld.de (Barbara Hammer)

Waterman, FASTA, or blast algorithm (Gusfield, 1997).

Methods based on similarity data with partial label information, where the similarities are defined on a metric space, as discussed in (Pekalska and Duin, 2005), can be effectively handled by semi-supervised extensions of kernel methods or other recently proposed, effective strategies (Subramanya and Bilmes, 2011; Tanha et al., 2014). However, in case of non-metric (dis-)similarity data without an explicit underlying vector representation and without requesting a metric space only few methods have been proposed so far in the literature of SSL (Rajadell et al., 2011; Trosset et al., 2008), and kernel techniques can be applied using some costly, potentially degenerating, transformations on the proximity data only (Pekalska et al., 2004).

First, we take a glance at SSL methods. One way to categorize SSL methods is to divide the field into generative models, low-density separation methods, and graph-based techniques typically used for a classification objective. A recent introduction to SSL is given in (Zhu and Goldberg, 2009). In generative models, the most basic technique is given by *self-training*. A classifier is first trained on the labeled instances and is then applied to unlabeled instances. Usually, some subset of those newly labeled instances are then used together with the original labeled data, to retrain the model. The major advantages of self-training are its simplicity and the fact that it is a wrapper method. It can 'wrap' the learner without changing its inner workings. In this paper we adopt this approach.

In (Suzuki et al., 2007) a more advanced approach was proposed. It employs expectation maximization (EM) to estimate parameters also on unsupervised data within a semi-supervised learning problem. In graph-based methods, the nodes of a graph represent labeled and unlabeled data, while some weights are assigned to its edges, which represent the similarities of two nodes. Now one may assume that similar points share common labels, which can be propagated according to some heuristics as shown in (Zhu and Goldberg, 2009). In this way labels are propagated from labeled data through the unlabeled data region. Different variations of this principle have been proposed, recently also for prototype based learning methods (Cruz-Barbosa and Vellido, 2010; Amis and Carpenter, 2010) and on large scale problems (Mantrach et al., 2011).

In low-density separation methods, probably the most popular semi-supervised learner is the *transductive Support Vector Machine* (TSVM) or variants thereof as the recently proposed S4VM (Li and Zhou, 2011). The semi-supervised SVM (S3VM) aims at approaching one optimal low-density separator employing unlabeled data, whereas *Safe* S3VM (S4VM) tries to exploit multiple candidate low-density separators simultaneously to reduce the risk of identifying a poor separator with unlabeled data. Besides, multi-kernel approaches have been recently analyzed for S3VM to incorporate additional meta-knowledge in the semi-supervised optimization (Tian et al., 2012). While most of these methods are defined for two-class problems, employing e.g. one-vs-rest wrappers for the multi-class case, native multi-class semi-supervised learning are analyzed less intensively. A multi-class S3VM approach was proposed in (Xu and Schuurmans, 2005), using a boosting strategy in (Song et al., 2011) and employing sparse Newton-optimization (Gieseke et al., 2012). Another recently published multi-class boosting technique in (Tanha et al., 2014) introduces a cost function based on empirical error of labeled data and similarity between labeled and unlabeled data. However, to solve the cost function as a convex problem the employed similarity metric has to be a valid kernel, i.e. positive semi-definite. Moreover, probabilistic models for semi-supervised learning based on nearest neighbor classifiers have been proposed recently (Ghosh, 2012) which allow multi-class learning. Some of these approaches are transductive like (Ghosh, 2012) and out of sample extensions are not naturally available limiting the applicability of the approaches for novel data in practice. A more theoretical analysis of SSL concepts was recently given in (Singh et al., 2008), discussing theoretical properties of semi-supervised learning and cases where SSL significantly improves the model compared to standard supervised learning, ignoring unlabeled data.

In contrast with the black box property of SVM and its semi-supervised variants, prototype-based methods enjoy a wide popularity in various application domains (Grbovic and Vucetic, 2013; Ortiz et al., 2013; Ortiz-Bayliss et al., 2012; Bacciu and Starita, 2009; Lee and Cho, 2006) due to their intuitive and simple behavior: they represent their decision in terms of typical representatives (referred to as prototypes) in the input space and classification is based on the distance of data to these prototypes. Prototypes can be directly inspected by domain experts in the field in the same way as data points. Popular supervised techniques include standard learning vector quantization (LVQ) and extensions to more powerful settings such as variants based on cost functions such as generalized LVQ (GLVQ) or robust soft LVQ (RSLVQ) (Sato and Yamada, 1995; Seo and Obermayer, 2003), just to name a few. A recently published prototype-based method extends the ability of GLVQ such that it can directly deal with dissimilarity data (Hammer et al., 2013), which we will use for semi-supervised problems.

In this paper we adopt the self-training approach with the prototype-based classifier proposed in (Hammer et al., 2013) for semi-supervised tasks employing the conformal prediction technique (Vovk et al., 2005; Shafer and Vovk, 2008), which provides a confidence measure of the classification. Using the confidence values a so-called *secure region* of unlabeled data can be identified during self-training and used in the retraining. This can potentially enhance the performance of the training, and at the same time conformal prediction estimates a so-called *insecure region* of labeled data helping to adapt the model complexity.

This paper is organized as follows. First we give a short review of the prototype-based technique for dissimilarity learning which we will use in the sequel in section 2. Subsequently, in section 3, we briefly introduce the concept of conformal prediction. Thereafter we show how to combine both techniques in the self-training approach for semi-supervised learning in section 4. Then we show the effectiveness of our technique on simulated data, compare it to state-of-the-art methods on SSL

benchmarks, and show results for biomedical dissimilarity data in section 5. Finally we summarize our results and discuss potential extensions.

2. Prototype-based relational learning

The basic idea of LVQ is to model data distribution(s) by positioning prototypes in the data space as accurately as possible. Assume data are given as vectors: $\mathbf{x}_i \in \mathbb{R}^d,\ i = 1, \ldots, N$ with label $l_i \in \mathbb{L} = \{1, \ldots, L\}$. LVQ is characterized by m prototypes $\mathbf{w}_j \in \mathbb{R}^d$ in the same space with priorly defined labels $c(\mathbf{w}_j) \in \mathbb{L}$. Besides classic heuristically motivated methods, one of the well-known cost function based learning vector quantization techniques is Generalized LVQ (GLVQ) from (Sato and Yamada, 1995).

Training of GLVQ aims at finding the positions of the prototypes while also taking the generalization ability into account, using the cost function

$$E_{GLVQ} = \sum_{i=1}^{N} \Phi\left(\frac{d(\mathbf{x}_i, \mathbf{w}^+(\mathbf{x}_i)) - d(\mathbf{x}_i, \mathbf{w}^-(\mathbf{x}_i))}{d(\mathbf{x}_i, \mathbf{w}^+(\mathbf{x}_i)) + d(\mathbf{x}_i, \mathbf{w}^-(\mathbf{x}_i))}\right) \tag{1}$$

where $\mathbf{w}^+(\mathbf{x}_i)$ is the closest prototype with the same label as $\mathbf{x}_i$ and $\mathbf{w}^-(\mathbf{x}_i)$ is the closest prototype with a different label than $\mathbf{x}_i$. $d(\cdot,\cdot)$ is the squared Euclidean distance. Φ is a monotonically increasing function, e.g. $\Phi(x) = (1+\exp(-x))^{-1}$. GLVQ tries to minimize the cost function (1) by means of a stochastic gradient descent, leading to Hebbian learning rules of prototypes, i.e. the closest prototype with the same label is attracted to $\mathbf{x}_i$ while the one with different label is pushed away from $\mathbf{x}_i$. Classification takes place by a so-called "winner takes all" principle: $\mathbf{x} \mapsto c(\mathbf{w}_j)$ where $d(\mathbf{x}, \mathbf{w}_j)$ is minimum, i.e. a new data point is labeled by the closest prototype.

GLVQ models have excellent generalization ability (Hammer et al., 2005; Biehl et al., 2006), however, they severely depend on the underlying metric, which is usually chosen as Euclidean metric. Thus, if data are inherently non-Euclidean, for example given in a form of a dedicated non-Euclidean dissimilarity measures such as dynamic time warping for time series, or alignment for symbolic strings (Gusfield, 1997), etc., it can not be applied. Recent research has extended GLVQ to directly deal with dissimilarity data (Hammer et al., 2013), which we will discuss in the following.

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 $(\mathbf{v}_i, \mathbf{v}_j) \in \mathbb{V} \times \mathbb{V}$. Any reasonable (possibly non-metric) distance measure is sufficient. Additionally, 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 $\{i, j\}$. Thereby, $\mathbf{v}_k$ is represented implicitly by a vector of known dissimilarities with respect to all $\mathbf{v}_j \in \mathbb{V}$. A training set is given where data point $\mathbf{v}_j$ is labeled by $\mathbf{l}_j \in \mathbb{L}$. As detailed in (Pekalska and Duin, 2005), dissimilarity data can always be embedded in pseudo-euclidean space in such a way that $d(\mathbf{v}_i, \mathbf{v}_j)$ is induced by a symmetric (but possibly not positive semi-definite) bilinear form.

For dissimilarity data classification, the key assumption is to restrict prototype positions to linear combinations of data points of the form

$$\mathbf{w}_j = \sum_i \gamma_{ji}\mathbf{v}_i \text{ with } \sum_i \gamma_{ji} = 1 \tag{2}$$

in the pseudo-Euclidean space. Then dissimilarities between data points and prototypes can be computed implicitly by means of

$$d(\mathbf{v}_i, \mathbf{w}_j) = [D \cdot \gamma_j]_i - \frac{1}{2} \cdot \gamma_j^t D \gamma_j \tag{3}$$

where $\gamma_j = (\gamma_{j1}, \ldots, \gamma_{jn})$ refers to the vector of coefficients describing prototype $\mathbf{w}_j$.

Thus, the cost function of GLVQ (1) can be transferred to the relational setting. The corresponding cost function of *Relational Generalized Learning Vector Quantization* (RGLVQ) becomes:

$$E_{\mathrm{RGLVQ}} = \sum_i \Phi\left(\frac{[D\gamma^+]_i - \frac{1}{2}\cdot(\gamma^+)^t D\gamma^+ - [D\gamma^-]_i + \frac{1}{2}\cdot(\gamma^-)^t D\gamma^-}{[D\gamma^+]_i - \frac{1}{2}\cdot(\gamma^+)^t D\gamma^+ + [D\gamma^-]_i - \frac{1}{2}\cdot(\gamma^-)^t D\gamma^-}\right), \tag{4}$$

where the closest correct and wrong prototypes are referred to, $\mathbf{w}^+$ and $\mathbf{w}^-$, respectively, corresponding to the coefficients γ^+ and γ^-, respectively. A simple stochastic gradient descent leads to adaptation rules for the coefficients γ^+ and γ^- in RGLVQ: component k of these vectors is adapted as

$$\begin{aligned}
\Delta\gamma_k^+ &\sim -\Phi'(\mu(\mathbf{v}_i)) \cdot \mu^+(\mathbf{v}_i) \cdot \frac{\partial\left([D\gamma^+]_i - \frac{1}{2}\cdot(\gamma^+)^t D\gamma^+\right)}{\partial\gamma_k^+} \\
\Delta\gamma_k^- &\sim \Phi'(\mu(\mathbf{v}_i)) \cdot \mu^-(\mathbf{v}_i) \cdot \frac{\partial\left([D\gamma^-]_i - \frac{1}{2}\cdot(\gamma^-)^t D\gamma^-\right)}{\partial\gamma_k^-}
\end{aligned}$$

with

$$\begin{aligned}
\mu(\mathbf{v}_i) &= \frac{d(\mathbf{v}_i, \mathbf{w}^+) - d(\mathbf{v}_i, \mathbf{w}^-)}{d(\mathbf{v}_i, \mathbf{w}^+) + d(\mathbf{v}_i, \mathbf{w}^-)} \\
\mu^+(\mathbf{v}_i) &= \frac{2 \cdot d(\mathbf{v}_i, \mathbf{w}^-)}{(d(\mathbf{v}_i, \mathbf{w}^+) + d(\mathbf{v}_i, \mathbf{w}^-))^2} \\
\mu^-(\mathbf{v}_i) &= \frac{2 \cdot d(\mathbf{v}_i, \mathbf{w}^+)}{(d(\mathbf{v}_i, \mathbf{w}^+) + d(\mathbf{v}_i, \mathbf{w}^-))^2}
\end{aligned}$$

The partial derivative yields $\frac{\partial\left([D\gamma_j]_i - \frac{1}{2}\cdot\gamma_j^t D\gamma_j\right)}{\partial\gamma_{jk}} = d_{ik} - \sum_l d_{lk}\gamma_{jl}$. After every adaptation step, normalization takes place to guarantee $\sum_i \gamma_{ji} = 1$. In this way, a learning algorithm which adapts prototypes in a supervised manner is given for general dissimilarity data, whereby prototypes are implicitly embedded in pseudo-Euclidean space.

The prototypes are initialized as random vectors corresponding to random values γ_{ij} which sum to one. It is possible to take class information into account by setting all γ_{ij} to zero which do not correspond to the class of the prototype. Out-of-sample extension of the classification to new data is possible based on the following observation: For a novel data point $\mathbf{v}$ characterized by its pairwise dissimilarities $D(\mathbf{v})$ to the data used for training, the dissimilarity of $\mathbf{v}$ to a prototype γ_j is

$$d(\mathbf{v}, \mathbf{w}_j) = D(\mathbf{v})^t \cdot \gamma_j - \frac{1}{2} \cdot \gamma_j^t D \gamma_j, \tag{5}$$

i.e. the data point is assigned to the label of the closest prototype. More details about the generalization ability can also be found in (Hammer et al., 2013).

2.1. Limitations

RGLVQ models work very effectively as shown in (Hammer et al., 2013), but they have two major limitations. They are crisp classifiers, where the classification function predicts only the class label but without any additional information about the confidence of the prediction. Especially in the life science some kind of reliability measure, similar to statistical *p*- or *q*-values would be beneficial. Only few attempts exist to give reliability estimates for these methods (see e.g. (Cordella et al., 1999; de Stefano et al., 2000)). The second drawback is that the complexity of the model in terms of the number of prototypes needs to be specified a priori.

In this contribution, we propose to use conformal prediction to enhance classification results with a level of confidence, and to automatically grow a model with suitable model complexity. Reliability, sometimes also referred to as confidence, has been the subject of a theory called *conformal prediction* as introduced in (Proedrou et al., 2002; Vovk et al., 2005). In the next section we will briefly introduce the concept of conformal prediction.

3. Conformal prediction

Conformal prediction is a statistical method assessing each classification decision by providing two measures: *credibility* and *confidence*. Thereby, this technique can be accompanied by a formal stability analysis as provided in (Vovk et al., 2005). For more details see (Shafer and Vovk, 2008) which is a recent tutorial on the topic.

We follow the general approach of conformal prediction as reviewed in (Vovk et al., 2005; Shafer and Vovk, 2008). Denote the labeled training data $\mathbf{z}_i = (\mathbf{v}_i, \mathbf{l}_i) \in \mathbb{Z} = \mathbb{V} \times \mathbb{L}$. Furthermore let $\mathbf{v}_{N+1}$ be a new data point with unknown label $\mathbf{l}_{N+1}$, i.e. $\mathbf{z}_{N+1} := (\mathbf{v}_{N+1}, \mathbf{l}_{N+1})$. 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, the *conformal prediction* computes 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. The applied method ensures that if the data $\mathbf{z}_i$ are *exchangeable*[1] then

$$P(\mathbf{l}_{N+1} \notin \Gamma^\epsilon(\mathbf{z}_1, \ldots, \mathbf{z}_N, \mathbf{v}_{N+1})) \leq \epsilon \tag{6}$$

holds asymptotically for $N \to \infty$ for each distribution of $\mathbb{Z}$. One says that the predictor is *asymptotically valid*. It is important to mention, that the probability is unconditional, such that if we repeat the process of drawing samples $\mathbf{v}_{N+1}$ and generating Γ^ϵ a number of n times we will find with respect to statistical fluctuations that in less than $\epsilon \cdot n$ cases the real label $\mathbf{l}_{N+1}$ is not under the predicted labels of Γ^ϵ.

[1] *exchangeability* is a weaker condition than data being i.i.d. which is readily applicable to the online set ting as well, for example (Vovk et al., 2005)

Algorithm 1 Conformal Prediction (CP)

function CP($\mathcal{D}$, $\mathbf{v}_{N+1}$, ϵ)
  for all $\mathbf{l} \in \mathbb{L}$ **do**
    $\mathbf{z}_{N+1} := (\mathbf{v}_{N+1}, \mathbf{l})$
    for $i = 1, \ldots, N+1$ **do**
      $\mathcal{D}_i := \{\mathbf{z}_1, \ldots, \mathbf{z}_{N+1}\} \backslash \{\mathbf{z}_i\}$
      $\alpha_i^{\mathbf{l}} := A(\mathcal{D}_i, \mathbf{z}_i)$ ▷ non conformity of $\mathbf{z}_i$ against $\mathcal{D}_i$
    end for
    $p_{N+1}^{\mathbf{l}} := \frac{|\{i=1,\ldots,N+1 \,|\, \alpha_i^{\mathbf{l}} \geq \alpha_{N+1}^{\mathbf{l}}\}|}{N+1}$
  end for
  return $\Gamma^\epsilon := \{\mathbf{l} : p_{N+1}^{\mathbf{l}} > \epsilon\}$
end function

3.1. Computation of prediction region

To compute the conformal prediction region Γ^ϵ, a *non-conformity measure* is fixed $A(\mathcal{D}, \mathbf{z})$. It is used to calculate a non-conformity value α that estimates how an observation $\mathbf{z}$ fits to given representative data $\mathcal{D}$=$\{\mathbf{z}_1, \ldots, \mathbf{z}_N\}$, we will give an example in section 3.1.1. In theory, any measure could be used, providing a nontrivial result for suitable choices only. Given a non-conformity measure A, significance level ϵ, examples $\mathbf{z}_1$, $\ldots$, $\mathbf{z}_N$, object $\mathbf{v}_{N+1}$ and a possible label $\mathbf{l}$, it is decided whether $\mathbf{l}$ is contained in $\Gamma^\epsilon(\mathbf{z}_1, \ldots, \mathbf{z}_N, \mathbf{v}_{N+1})$ according to algorithm 1.

However, this method would entail high computational costs, especially for large data set, because this procedure has to be done for all leave-one-out multi-sets for each of the test objects with all possible labels $(\mathbf{v}_{N+1}, \mathbf{l})$. To get rid of this problem, some extensions of conformal prediction have been published, i.e. *Inductive Conformal Prediction* (ICP) (Papadopoulos et al., 2002; Vovk, 2012a) and *Cross Conformal Prediction*(CCP) (Vovk, 2012b). Inductive conformal prediction divides the training data into two subsets: *proper training set* and *calibration set*. The model is trained on the proper training set and then used to calculate the non-conformity values of the calibration set. For new data points, classification takes place only based on the non-conformity of the calibration set. As pointed out by (Vovk, 2012a) the size of the calibration set should be reasonably large to cover the data statistic. Although ICP is computationally more efficient, since the training process only has to be done once, it is predictively less efficient in comparison to the original conformal prediction, in which the training set serves as proper training set and also as calibration set. To avoid this problem another approach, cross-conformal prediction has been proposed, which combines cross-validation with inductive conformal prediction. During the cross-validation process (by taking one fold as calibration set and the remaining folds as proper training set) the data statistic of the whole training set is accumulatively considered, finally the non-conformity of each calibration is merged to classify new data, see (Vovk, 2012b) for more details.

In this work we focus on semi-supervised problems, hence the size of the training set (i.e. labeled data) is usually not large such that we can not use ICP or CCP for our purpose. We decided to modify the original conformal prediction in a different way: we do not match the model exactly against each data set $\mathcal{D}_i$ but instead use the whole training data (i.e. $\mathcal{D}$, excluding $\mathbf{z}_{N+1}$). In this way learning must be performed only once on $\mathcal{D}$. This procedure is motivated by two facts: (1) since we intend

to use prototype-based method to train the model, the positions of prototypes depend on the whole data distribution and are in general not widely affected by a single data point, (2) the information loss will be small if the number of training data is reasonably large, so that adding $\mathbf{z}_i$ but leaving out $\mathbf{z}_{N+1}$ will not affect the learning results. Before we go into more details about the proposed method, we will first discuss a key point of conformal prediction, the non-conformity measure.

3.1.1. Non-Conformity Measure

As explained above, the non-conformity measure $A(\mathcal{D}, \mathbf{z})$ should evaluate whether a test example $\mathbf{z}$ fits representative data $\mathcal{D}$. It is the part of the method that can incorporate detailed knowledge about the data distribution. Nevertheless one can use any real valued function[2], but maybe with negative impact on the prediction efficiency.

For given $\mathbf{z} = (\mathbf{x}, \mathbf{l})$ and a trained relational GLVQ model, we choose as non-conformity measure

$$\alpha_{\mathbf{x}}^{\mathbf{l}} := \frac{d^+(\mathbf{x})}{d^-(\mathbf{x})} \tag{7}$$

with $d^+(\mathbf{x})$ being the distance between $\mathbf{x}$ and the closest prototype labeled $\mathbf{l}$, and $d^-(\mathbf{x})$ being the distance between $\mathbf{x}$ and the closest prototype labeled differently than $\mathbf{l}$ where distances are computed according to Eq. (3). We expect that values $\alpha_x^{\mathbf{l}}$ are small for data $\mathbf{z}$ for which the prediction has high confidence, but it is large if the label does not comply with data.

3.2. Confidence and credibility

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

Suppose we use a meaningful non-conformity measure A, e.g. eq. (7). 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 p^l-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 data distribution $\mathbb{Z}$ and so $\mathbf{l}$ does not seem to be the right label. If one further raises ϵ only those $\mathbf{l}$ remain in the conformal region that can produce a high p^l-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 $\mathbf{v}_i$ two error rates are of particular interest, ϵ_1^i being the smallest ϵ and ϵ_2^i being the largest ϵ so that $|\Gamma^\epsilon(\mathcal{D}, \mathbf{v}_i)| = 1$. ϵ_2^i is the p-value of the best and ϵ_1^i is the p-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

$$\text{confidence} : cf_i := 1 - \epsilon_1^i = 1 - p^{\mathbf{l}_{\text{2nd}}} \tag{8}$$

$$\text{credibility} : cr_i := \epsilon_2^i = p^{\mathbf{l}_{\text{1st}}} \tag{9}$$

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 typical and not an outlier. An example is shown in Figure 1: the data consist of two well-separated clusters. The data points around the centers (e.g. in the dashed circles) have higher credibility and higher confidence than the data farther from the centers. The data points that are a bit farther from the centers but not outliers (e.g. in the dashed ellipses) have higher credibility but lower confidence (because they are nearer to the other cluster than the data around the centers). Furthermore there are two types of outliers: (i) the data points are far away from the centers but nearer to the other cluster than other data points in the same cluster, so they have lower credibility and lower confidence. (ii) the data points are far away from the centers and even farther away from the other cluster than other data points in the same cluster, so they have lower credibility and higher confidence.

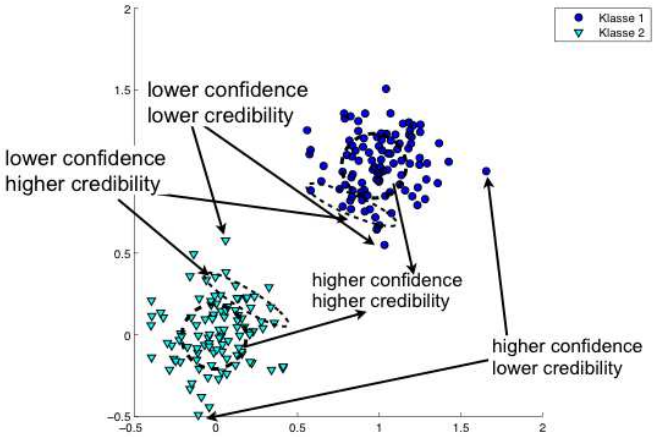


Fig. 1: An example about confidence and credibility

The non-conformity measure has a direct impact on the efficiency of the prediction region. A good, informative measure will exclude wrong labels for small error rates and will reject typical data only for large error rates, meaning that $\epsilon_2^i - \epsilon_1^i$ is large for typical data $\mathbf{v}_i$. That means, that a good measure can give useful information already for small error rate ϵ_1^i and on the other hand one would have to face up a high average error rate ϵ_2^i to exclude the right label from the prediction region.

We would like to point out that the concept of conformal prediction permits pointwise measures of confidence which change if the training data is adapted, also if the decision boundaries remain the same. This means, that similar as in classical statistics, more densely populated training regions permit a better confidence in a decision. Due to the definition of conformal predic-

[2] *Any* measurable function on $\mathbb{Z}^{(*)} \times \mathbb{Z}$ taking values in the extended real line is a non conformity measure.

Algorithm 2 Self training

1: T_{lab}:= labeled data, T_{unlab}:= unlabeled data
2: **repeat**
3: Train model f based on T_{lab} using supervised learning
4: Apply f to T_{unlab}
5: remove a subsect S from T_{unlab} and add $\{(x, f(x))|x \in S\}$ to T_{lab}

tion, this is automatically achieved also in online scenarios.

4. Semi-supervised conformal relational GLVQ

RGLVQ opens a way to directly deal with dissimilarity data. As mentioned in section 2.1 it has two major limitations: (i) It is a crisp classifier without any additional information about the confidence of the prediction and (ii) the number of prototypes has to be defined in advance. In the supervised case, these problems have been already addressed by (Schleif et al., 2014) in which the concept of inductive conformal prediction is integrated into a sparse prototype-based classifier for dissimilarity learning problems resulting a sparse prototypical representation of data. In this work we focus on semi-supervised case and by extending our previous work (Zhu et al., 2013) we propose a prototype-based conformal classifier with self-adaptation of model complexity based on the data with high confidence and high credibility values provided by conformal prediction.

First, we denote T_{lab} as labeled data and T_{unlab} as unlabeled data. Generally, in semi-supervised learning unlabeled data are used to improve the trained model based on labeled data in some way. *Self-training* (Zhu and Goldberg, 2009) is a very simple approach, which takes iteratively a part of the unlabeled data with predicted labels as new training data into the retraining process to optimize the model, as shown in Algorithm 2. After the first training of model f on labeled data, the model f is then used to predict the labels of unlabeled data. A subset S of the unlabeled data together with their predicted labels are selected and added to the labeled data, which builds a new larger set of labeled data. The model f is retrained on the new unlabeled data, and the procedure is repeated. As pointed out by (Zhu and Goldberg, 2009), the key assumption of self-training is that the predictions, at least the high confidence ones, tend to be correct. S should consist of the unlabeled data with the most confident predictions.

In this work we combine conformal prediction with self-training to find the most confident unlabeled data (see Algorithm 4). We first train the model on labeled data (T_{lab}) using RGLVQ, based on the model we proceed with the conformal prediction step (line 20-26): For T_{lab} and T_{unlab}, we compute non-conformity values (α) according to (7) (line 21-22). Based on these non-conformity values a p-value is estimated for each possible label and each unlabeled point from T_{unlab} (line 23-24). For classification using the conformal classifier, the label of a unlabeled item will be finally predicted as the label with the largest p-value. This refers to the label set provided by the conformal predictor which contains only one label. More complex schemes, by analyzing for example label sets with more than one label would be possible as well, but are not further considered here. The confidence value (cf_i) is given as one minus the second largest p-value (eq. (8)) and the credibility (cr_i) is the largest p-value of this item (eq. (9)) (line 25-26) (for more detail see section 3.2).

Data used for self-training

In oder to identify unlabeled items with high confidence predictions we define a measure cc as the product of confidence and credibility values: For a given data point $\mathbf{v}_i \in T_{\text{unlab}}$,

$$cc_i := cf_i \cdot cr_i \tag{10}$$

A high cc-value of a unlabeled item indicates that with high probability its predicted label (that with the highest p-value) is the true underlying label. For self-training the unlabeled data with predicted labels of high probability can be taken into the next retraining. The region which consists of these unlabeled items is referred to as '*secure region*' (denoted as $\mathcal{SR}$). To identify $\mathcal{SR}$ we take a fraction (prc) of the top cc-values of the unlabeled data[3].

Adaptation of model complexity

On the other hand we also collect a set of points of the "labeled" data (i.e. original labeled items and the items with high cc-values labeled by previous iterations) with low credibility and confidence values, which builds a so-called '*insecure region*' ($\mathcal{ISR}$) of the training data,

$$\mathcal{ISR} := \{\mathbf{v}_i \in T_{\text{lab}} : cf_i \leq \zeta_1 \vee cr_i \leq \zeta_2\}. \tag{11}$$

A low confidence value is given if the confidence value cf_i or the credibility cr_i below a user defined threshold ζ_i or ζ_2, respectively. Defined values for ζ_1 or ζ_2 can be derived from the quantiles of confidence/credibilty values as observed in the data.

The $\mathcal{ISR}$ will be represented by a new prototype as the median of $\mathcal{ISR}$. This step automatically adapts the complexity of the model, i.e. the number of prototypes. In the next retraining this new prototype will be also trained on the new training data.

During the self-training process the training set T_{lab} is iteratively augmented by adding the secure region of the unlabeled data $\mathcal{SR}$ to itself while the unlabeled data T_{unlab} is shrunk by discarding the secure region. The performance of the retraining is evaluated based on the original labeled data only. The method terminates if the improvement of the performance is not significant (less than 1%) after a certain number of iterations ($win_{\text{max\_itr}}$) or the maximal number of iterations are reached (max_{itr}) or the insecure region ($\mathcal{ISR}$) is too small or the unlabeled set T_{unlab} is empty, i.e. all unlabeled data have been considered in the retraining. The proposed method is referred to as *Secure Semi-Supervised Conformal RGLVQ* (S3-C-RGLVQ).

5. Experiments

We evaluate S3-C-RGLVQ on a large range of tasks. First, we demonstrate its performance for two artificial data sets:

[3] prc is customizable and in our experiments we set prc = 5% which is a good compromise between learning performance and efficiency.

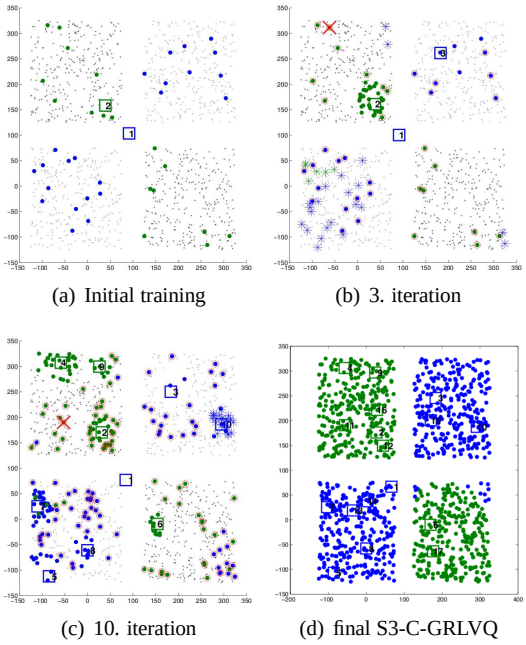
Algorithm 3 secure semi-supervised conformal RGLVQ

init: W := randomly initialized, $W_{\text{new}} := \emptyset$, $W_{\text{best}} := W$, $\mathcal{ISR} := \emptyset$; $\mathcal{SR} := \emptyset$ ▷ W: randomly initialized prototypes, W_{new}: new prototype chosen from insecure region,W_{best} : best prototype identified by retraining process
T_{lab} := labeled data; T_{unlab} := unlabeled data
$improve = 1\%$ ▷ threshold of improvement: default 1%
$EvalSet := T_{lab}$ ▷ Evaluation set, i.e. labeled data
$itr = 0$ ▷ iteration counter
$ctn_{\text{best}} = 0$ ▷ counter for best result
$max_{\text{itr}} = 100$ ▷ maximal total iterations
$win_{\text{max\_itr}} = 10$ ▷ maximal iterations for a result as winner
$acc_{\text{best}} = 0$
repeat ▷ self-training process
  $W := W \bigcup W_{\text{new}}$ ▷ see description around eq. (11)
  $T_{\text{lab}} := T_{\text{lab}} \cup \mathcal{SR}$, $T_{\text{unlab}} := T_{\text{unlab}} \backslash \mathcal{SR}$
  W := train T_{lab} by RGLVQ given W ▷ retraining with given prototypes
  acc := evaluation of W on $EvalSet$;
  if $acc - acc_{\text{best}} \geq improve$ **then**
    $W_{\text{best}} = W$, $acc_{\text{best}} = acc$, $ctn_{\text{best}} = 0$
  else
    $ctn_{\text{best}} = ctn_{\text{best}} + 1$
  end if
  ▷ conformal prediction step
  $A_{T_{\text{lab}}} := \{\alpha_i, \forall i \in T_{\text{lab}}\}$ ▷ α-values of T_{lab} w.r.t. W: eq. (7)
  $A^{\mathbb{L}}_{T_{\text{unlab}}} := \{\alpha_i^l, \forall i \in T_{\text{unlab}}, \forall l \in \mathbb{L}\}$ ▷ α-values of T_{unlab} for all possible labels w.r.t. W: eq. (7)
  $P_{T_{\text{lab}}} := \{p_i, \forall i \in T_{\text{lab}}\}$ ▷ p-values of T_{lab}
  $P^{\mathbb{L}}_{T_{\text{unlab}}} := \{p_i^l, \forall i \in T_{\text{unlab}}, \forall l \in \mathbb{L}\}$ ▷ p-values of T_{unlab} for all possible labels based on $A_{T_{\text{lab}}}$ and $A^{\mathbb{L}}_{T_{\text{unlab}}}$
  $CF_{T_{\text{lab}}} := \{cf_i, \forall i \in T_{\text{lab}}\}$; $CR_{T_{\text{lab}}} := \{cr_i, \forall i \in T_{\text{lab}}\}$;
  $CF_{T_{\text{unlab}}} := \{cf_i, \forall i \in T_{\text{unlab}}\}$; $CR_{T_{\text{unlab}}} := \{cr_i, \forall i \in T_{\text{unlab}}\}$; ▷ confidence/credibility of $T_{\text{lab}}/T_{\text{unlab}}$ by means of $P_{T_{\text{lab}}}/P^{\mathbb{L}}_{T_{\text{unlab}}}$: eq. (8),(9)
  generate $\mathcal{ISR}$ of T_{lab} based on $CF_{T_{\text{lab}}}$ and $CR_{T_{\text{lab}}}$ ▷ eq. (11)
  generate $\mathcal{SR}$ of T_{unlab} based on $CF_{T_{\text{unlab}}}$ and $CR_{T_{\text{unlab}}}$ ▷ eq. (10) and $prc = 5\%$
  generate W_{new} from $\mathcal{SR}$
  $itr = itr + 1$
until $|\mathcal{ISR}| < 1\% \cdot |T_{unlab}|$ or $itr = max_{itr}$ or $ctn_{\text{best}} = win_{\text{max\_itr}}$ or $T_{\text{unlab}} = \emptyset$
return W_{best};

checkerboard data and banana-shaped data, with known vector representation to show the ability of dealing with partially labeled data, especially non i.i.d labeled data. Then we compare S3-C-RGLVQ with state-of-the-art semi-supervised SVMs on SSL binary-class benchmarks. For vectorial data the dissimilarity matrices D are obtained using the squared-Euclidean distance. Additionally, five real life non-vectorial multi-class data sets from the bioinformatics domain are used to compare with original RGLVQ (trained only on labeled data). For all experiments, prototypes are randomly initialized based on labeled data and one prototype per class.

Artificial data sets: The checkerboard data set consists of two classes with 1200 data points, in two dimensions and $2 \cdot 2$ clusters. We randomly select about 3% as labeled data and the remaining data as unlabeled data. RGLVQ can learn these data only if the prototypes are initialized near the centers of the multi-modal distributions, provided a sufficient number of prototypes. The S3-C-RGLVQ on the other hand automatically adapts its model complexity according to the introduced scheme, leading to an effective model with minimum initialization of one prototype per class only. As an example, Figure 2 shows some intermediate results up to convergence. We randomly initialized two prototypes only on labeled data. Figure 2(a) shows that after the initial training two prototypes are located in the center of the labeled data. Obviously, in this case one prototype per each class is not sufficient to model the whole data space. In Figure 2(b) after the conformal prediction process, the secure region of unlabeled data and the insecure region of labeled data can be identified. To 'cover' the insecure region a new prototype (marked by red cross) is added thereinto. Moreover, there are some unlabeled data misclassified by CP, which will be taken into the current retraining process. The reason thereof is that due to the smaller number of prototypes at the early stage which are not well distributed into the multi-modal clusters, a reasonable number of points with relatively lower confidence/credibility values (i.e. lower cc-value) exists, which is a natural consequence, because by chance 50% got the correct label. By a larger value of the parameter 'prc' some of these points can be considered in the next training. In this case those points can also be considered as outliers. Due to the fact that the prototype-based method is very stable against outliers, i.e. the positions of prototypes depend on the whole data distribution and are not widely affected by a single point, the movement of the prototypes is mainly dominated by the correctly classified points and the labeled data. As shown in Figure 2(d), once the algorithm converges, those points can be correctly assigned to their closest prototypes. Fig. 2(c) shows also the intermediate result in the 10th iteration with more prototypes.

Another simulated data set consists of two banana-shaped data clouds indicating two classes. Each banana consists of 300 two dimensional data points, see Figure 3. We randomly select non i.i.d. a small fraction (ca. 5%) of each banana as labeled data, the remaining as unlabeled data. The dissimilarity matrix D thereof is obtained by Euclidean distance again. With the same setting for checkerboard data we start with one prototype per class and train the initial model on the labeled data as



(a) Initial training (b) 3. iteration

(c) 10. iteration (d) final S3-C-GRLVQ

Fig. 2: (a) The data consists of green/blue labeled data and gray unlabeled data. Two prototypes are trained on labeled data and marked with squares. (b) After the initial training, by means of CP the secure region $\mathcal{SR}$ can be found which consists of the unlabeled data marked by stars, as well as the insecure region $\mathcal{ISR}$ which contains labeled data marked by red circles. The new prototype taken from $\mathcal{ISR}$ is marked with a big red cross. (c) During the self-training process additional prototypes are step–wise created. (d) the final result until convergence

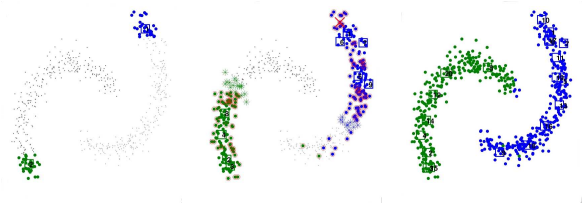
shown in Fig. 3(a). The number of prototypes increased stepwise during the retraining process by adding new prototype in the insecure region, while by means of secure region the unlabeled data are iteratively considered. Thereby at the end the data manifold can be well studied.

UCI two-class data sets: Furthermore, we evaluate the proposed method on different widely used benchmarks for semi-supervised learning from the UCI repository[4] and compare it with the best semi-supervised SVM with RBF-kernel taken from (Li and Zhou, 2011)[5]. To keep the same experimental setting, 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 12 times and the average test-set accuracy (on the unlabeled data) and standard deviation are reported in Table 1. Except voting data, the proposed method provides comparable results for all remaining data sets.

Real life multi-class data sets: Moreover, we also evaluate the methods on five real life relational data sets from the bioinformatics domain, where no direct vector embedding exists and the data are given as (dis-)similarities.

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

[5] In this paper the authors made a comprehensive comparison between different semi-supervised SVMs, e.g. TSVM, S3VM, S4SVM, etc. with linear and rbf kernels. For our experiments we pick the best result of rbf-kernel among them as reference for each data.



(a) two bananas (b) 10. iteration (c) final model of S3-C-RGLVQ

Fig. 3: (a) The data consist of green/blue labeled data and gray unlabeled data. Two initial prototypes are trained on labeled data and marked with squares. (b) The secure region $\mathcal{SR}$ consists of the unlabeled data marked by stars and the insecure region $\mathcal{ISR}$ contains labeled data rounded by red circles. The new prototype taken from $\mathcal{ISR}$ is marked with a big red cross. During the self-training process additional prototypes are created. (c) the final result of S3-C-RGLVQ

Table 1: Classification accuracy (% ± std) of UCI Benchmarks for two classes problems for SSL

two-class UCI data	Semi-RLVQ	Semi-SVM^{best}(rbf)
diabetes	70.17 ± 2.32	70.3 ± 2.1
german	71.61 ± 1.14	71.0 ± 1.1
haberman	**73.30 ± 5.02**	68.3 ± 2.8
voting	89.20 ± 0.89	**92.6 ± 1.6**
wdbc	92.34 ± 1.19	**93.6 ± 1.7**
austrailian	**83.22 ± 1.51**	81.8 ± 1.9
breast-cancer	**96.20 ± 0.51**	95.5 ± 1.0

- The *SwissProt* data set consists of 5,791 samples of protein sequences in 10 classes taken as a subset from the popular SwissProt database of protein sequences (Boeckmann B, 2003) (release 37). These sequences are compared using the Smith-Waterman algorithm (Gusfield, 1997) .

- The *Copenhagen Chromosomes* data constitute a benchmark from cytogenetics. 4,200 human chromosomes from 21 classes are represented by grey-valued images. These are transferred to strings measuring the thickness of their silhouettes. These strings can be directly compared using the edit distance based on the differences of the numbers and insertion/deletion costs 4.5 (Neuhaus and Bunke, 2006).

- The *Sonatas* data set contains complex symbolic data similar to Mokbel et al. (2009). It is comprised of pairwise dissimilarities between 1,068 sonatas from the classical period (by Beethoven, Mozart and Haydn) and the baroque era (by Scarlatti and Bach). The musical pieces were given in the MIDI file format, taken from the online MIDI collection *Kunst der Fuge*[6]. Their mutual dissimilarities were measured with the normalized compression distance (NCD), see (Cilibrasi and Vitányi, 2005). The musical pieces are classified according to their composer.

[6] `http://www.kunstderfuge.com`

- The Zongker digit dissimilarity data (2000 samples in 10 classes) from (Duin, 2012) is based on deformable template matching. The dissimilarity measure was computed between 2000 handwritten NIST digits in 10 classes, with 200 entries each, as a result of an iterative optimization of the non-linear deformation of the grid (Jain and Zongker, 1997).

- The Vibrio data set consists of 1,100 samples of vibrio bacteria populations characterized by mass spectra. The full data set consists of 49 classes of vibrio-sub-species. The mass spectra are preprocessed with a standard workflow using the BioTyper software (Maier et al., 2006). The Vibrio similarity matrix S has a maximum score of 3. The corresponding dissimilarity matrix is obtained as $D = 3 - S$.

These data sets constitute typical examples of non-Euclidean data which occur in complex systems, such as medical image analysis, mass spectrometry, and symbolic domains. In all cases, dedicated preprocessing steps and dissimilarity measures for structures are used. The dissimilarity measures are inherently non-Euclidean and cannot be embedded isometrically in a Euclidean vector space.

We use the same experimental setting as for the UCI data, i.e. we randomly select 100 examples as labeled data, the remaining as unlabeled data (with 10 repeats), prototypes are initialized based on labeled data and one prototype per class. For comparison, we report the results of RGLVQ trained only on labeled data to tackle another problem for SSL, i.e. the degeneration issue as discussed by (Singh et al., 2008; Li and Zhou, 2011; Zhu and Goldberg, 2009). In order to keep the comparisons fair the number of prototypes for each class for RGLVQ is set to the number of prototypes for each class of the final S3-C-RGLVQ model. The mean classification accuracies are reported in Table 2.

In all cases but one, a better classification accuracy can be obtained using conformal prediction compared to original RGLVQ only based on labeled data without consideration of additional information about unlabeled data. The chromosome is a perfectly balanced data set, it leads to the fact that the initial model based only on the labeled data is almost perfectly trained by RGLVQ, so that the potential to improve the model by considering unlabeled information in this case is very limited.

In all cases, the incorporation of information about unlabeled data into the classifier leads to an increased, at least equal, classification accuracy of the resulting model, since the additionally available information can better be taken into account to optimize the class boundaries. Thus, S3-C-RGLVQ constitutes a very promising method to infer a high quality semi-supervised prototype-based classifier for general dissimilarity data sets which offers point-wise measures for confidence and credibility about the classification.

Table 2: Classification accuracy (% ± std) for real life data.

Data	S3-C-RGLVQ	RGLVQ
swissprot	**81.06 ± 5.53**	79.37 ± 4.78
chromosome	78.88 ± 3.28	78.78 ± 3.70
sonatas	**77.98 ± 3.94**	71.99 ± 2.92
zongker	**87.93 ± 0.84**	86.48 ± 1.50
vibrio	**98.76 ± 0.47**	97.40 ± 0.84

6. Conclusions

In this contribution, we have developed an efficient semi-supervised classification technique for general dissimilarity data, which represents the decisions in the form of prototypes, based on the conformal prediction concept and relational prototype-based classifier. It naturally inherits the merits from both techniques. Due to a prototypical representation, unlike many alternative black-box techniques, it offers the possibility of a direct inspection of the classifier by humans. Further, unlike kernel-based alternatives such as kernel GLVQ (Qin and Suganthan, 2004) or relevance vector machine (Tipping, 2001), this technique does not require that data are embeddable into Euclidean space, rather, a general symmetric dissimilarity matrix is sufficient. For those alternative techniques to deal with dissimilarity data, extra preprocessing steps have to be added as described by (Pekalska and Duin, 2005). Due to the properties of conformal prediction, instead of providing only a predicted label, it also permits to identify the safety of the prediction by means of point-wise measures for confidence and credibility. Thereby the 'secure' unlabeled data can be exploited and used to optimize the trained model, at the same time the 'insecure' training data can be identified and accordingly the complexity of the model is adapted.

We demonstrated the quality of the technique on different SSL data sets. As a results, a powerful semi-supervised learning algorithm can be derived, which in most cases achieves comparable results to semi-supervised SVM and with direct interpretability of the classification in term of the prototypes. It works especially well for non i.i.d labeled data. Duo to the multi-class capability of prototype-based method, it can directly deal with multi-class data sets. Furthermore, it does not degenerate the learning performance by incorporating additional information of unlabeled data which is still a crucial issue in the semi-supervised learning (Singh et al., 2008; Li and Zhou, 2011; Zhu and Goldberg, 2009).

One central problem of this technique as introduced above has not yet been considered in this letter: we used a global value *prc* to identify the secure region of the training data in every iteration. It may cause some uncertainty issues at the earlier stages of retraining as we have seen in the checkerboard data, if the number of prototypes is not sufficiently high and the prototypes are not sufficiently distributed in the data space. In spite of the fact that this potential issue can be partially solved by the nature of prototype-based method, i.e. its stability against outliers, it should be more seriously studied, e.g. using a local value *prc* for each iteration to more precisely identify the high confidence items. Future work will also address the model sparsity for large scale problem and linear approximation techniques as introduced in (Zhu et al., 2012).

Acknowledgments

This research and development project is funded by the German Federal Ministry of Education and Research (BMBF) within the Leading-Edge Cluster Competition and managed by the Project Management Agency Karlsruhe (PTKA). The authors are responsible for the contents of this publication. Funding in the frame of the centre of excellence 'Cognitive Interaction Technologies' (CITEC) is grate- fully acknowledged. The second author was kindly supported by a Marie Curie Intra-European Fellowship (IEF) FP7-PEOPLE-2012-IEF (FP7-327791-ProMoS).

References

Amis, G.P., Carpenter, G.A., 2010. Self-supervised artmap. Neural Networks 23, 265–282.

Bacciu, D., Starita, A., 2009. Expansive competitive learning for kernel vector quantization. Pattern Recognition Letters 30, 641–651.

Biehl, M., Ghosh, A., Hammer, B., Bengio, Y., 2006. Dynamics and generalization ability of lvq algorithms, in: Journal of Machine Learning Research.

Boeckmann B, e., 2003. The swiss-prot protein knowledgebase and its supplement trembl. Nucleic Acids Research 31, 365–370.

Chapelle, O., Schölkopf, B., Zien, A. (Eds.), 2006. Semi-Supervised Learning. MIT Press, Cambridge, MA.

Cilibrasi, R., Vitányi, P.M.B., 2005. Clustering by compression. IEEE Transactions on Information Theory 51, 1523–1545.

Cordella, L.P., Foggia, P., Sansone, C., Tortorella, F., Vento, M., 1999. Reliability parameters to improve combination strategies in multi-expert systems. Pattern Analysis and Applications 2, 205–214.

Cruz-Barbosa, R., Vellido, A., 2010. Semi-supervised geodesic generative topographic mapping. Pattern Recognition Letters 31, 202–209.

Duin, R.P., 2012. PRTools. URL: http://www.prtools.org.

Ghosh, A.K., 2012. A probabilistic approach for semi-supervised nearest neighbor classification. Pattern Recognition Letters 33, 1127–1133.

Gieseke, F., Airola, A., Pahikkala, T., Kramer, O., 2012. Sparse quasi-newton optimization for semi-supervised support vector machines, pp. 45–54.

Grbovic, M., Vucetic, S., 2013. Decentralized estimation using distortion sensitive learning vector quantization. Pattern Recognition Letters 34, 963–969.

Gusfield, D., 1997. Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology. Cambridge University Press.

Haasdonk, B., Bahlmann, C., 2004. Learning with distance substitution kernels. Pattern Recognition - Proc. of the 26th DAGM Symposium .

Hammer, B., Hofmann, D., Schleif, F.M., Zhu, X., 2013. Learning vector quantization for (dis-)similarities. Neurocomputing doi:http://dx.doi.org/10.1016/j.neucom.2013.05.054.

Hammer, B., Strickert, M., Villmann, T., 2005. On the generalization ability of GRLVQ networks. Neural Processing Letters 21, 109–120.

Jain, A.K., Zongker, D., 1997. Representation and recognition of handwritten digits using deformable templates. IEEE Trans. Pattern Anal. Mach. Intell. 19, 1386–1391. doi:10.1109/34.643899.

Lee, H.j., Cho, S., 2006. Application of lvq to novelty detection using outlier training data. Pattern Recognition Letters 27, 1572–1579.

Li, Y.F., Zhou, Z.H., 2011. Towards making unlabeled data never hurt, in: Getoor, L., Scheffer, T. (Eds.), ICML, Omnipress. pp. 1081–1088.

Maier, T., Klebel, S., Renner, U., Kostrzewa, M., 2006. Fast and reliable maldi-tof ms–based microorganism identification. Nature Methods .

Mantrach, A., van Zeebroeck, N., Francq, P., Shimbo, M., Bersini, H., Saerens, M., 2011. Semi-supervised classification and betweenness computation on large, sparse, directed graphs. Pattern Recognition 44, 1212–1224.

Mokbel, B., Hasenfuss, A., Hammer, B., 2009. Graph-based representation of symbolic musical data, in: GbRPR, pp. 42–51.

Neuhaus, M., Bunke, H., 2006. Edit distance based kernel functions for structural pattern classification. Pattern Recognition 39, 1852–1863.

Ortiz, A., Górriz, J., Ramírez, J., Martínez-Murcia, F., 2013. Lvq-svm based cad tool applied to structural mri for the diagnosis of the alzheimer's disease. Pattern Recognition Letters 34, 1725–1733.

Ortiz-Bayliss, J., Terashima-Marín, H., Conant-Pablos, S., 2012. Learning vector quantization for variable ordering in constraint satisfaction problems. Pattern Recognition Letters In Press.

Papadopoulos, H., Proedrou, K., Vovk, V., Gammerman, A., 2002. Inductive confidence machines for regression, in: Elomaa, T., Mannila, H., Toivonen, H. (Eds.), ECML, pp. 345–356.

Pekalska, E., Duin, R., 2005. The dissimilarity representation for pattern recognition. World Scientific.

Pekalska, E., Duin, R.P.W., Günter, S., Bunke, H., 2004. On not making dissimilarities euclidean, in: Fred, A.L.N., Caelli, T., Duin, R.P.W., Campilho, A.C., de Ridder, D. (Eds.), SSPR/SPR, Springer. pp. 1145–1154.

Proedrou, K., Nouretdinov, I., Vovk, V., Gammerman, A., 2002. Transductive confidence machines for pattern recognition, in: Elomaa, T., Mannila, H., Toivonen, H. (Eds.), ECML, Springer. pp. 381–390.

Qin, A.K., Suganthan, P., 2004. A novel kernel prototype-based learning algorithm, in: Pattern Recognition, Proceedings of 17th ICPR, pp. 621–624.

Rajadell, O., Garcia-Sevilla, P., Dinh, V., Duin, R., 2011. Semi-supervised hyperspectral pixel classification using interactive labeling, in: WHISPERS, 2011 3rd Workshop on, pp. 1–4. doi:10.1109/WHISPERS.2011.6080905.

Sato, A., Yamada, K., 1995. Generalized learning vector quantization, in: Touretzky, D.S., Mozer, M., Hasselmo, M.E. (Eds.), NIPS, MIT Press. pp. 423–429.

Schleif, F.M., Zhu, X., Hammer, B., 2014. Sparse conformal prediction for dissimilarity data. Annals of Mathematics and Artificial Intelligence (AMAI) doi:10.1007/s10472-014-9402-1.

Seo, S., Obermayer, K., 2003. Soft learning vector quantization. Neural Computation 15, 1589–1604.

Shafer, G., Vovk, V., 2008. A tutorial on conformal prediction. Journal of Machine Learning Research 9, 371–421.

Singh, A., Nowak, R.D., Zhu, X., 2008. Unlabeled data: Now it helps, now it doesn't, in: Koller, D., Schuurmans, D., Bengio, Y., Bottou, L. (Eds.), NIPS, Curran Associates, Inc.. pp. 1513–1520.

Song, E., Huang, D., Ma, G., Hung, C.C., 2011. Semi-supervised multi-class adaboost by exploiting unlabeled data. Expert Systems with Applications 38, 6720–6726.

de Stefano, C., Sansone, C., Vento, M., 2000. To reject or not to reject: that is the question: an answer in case of neural classifiers. IEEE Transactions on Systems, Man and Cybernetics Part C 30, 84–93.

Subramanya, A., Bilmes, J., 2011. Semi-supervised learning with measure propagation. Journal of Machine Learning Research 12, 3311–3370.

Suzuki, J., Fujino, A., Isozaki, H., 2007. Semi-supervised structured output learning based on a hybrid generative and discriminative approach, in: EMNLP-CoNLL, ACL. pp. 791–800.

Tanha, J., van Someren, M., Afsarmanesh, H., 2014. Boosting for multiclass semi-supervised learning. Pattern Recognition Letters 37, 63 – 77.

Tian, X., Gasso, G., Canu, S., 2012. A multiple kernel framework for inductive semi-supervised svm learning. Neurocomputing 90, 46–58.

Tipping, M.E., 2001. Sparse bayesian learning and the relevance vector machine. J. Mach. Learn. Res. 1, 211–244.

Trosset, M.W., Priebe, C.E., Park, Y., Miller, M.I., 2008. Semisupervised learning from dissimilarity data. Computational Statistics and Data Analysis 52, 4643 – 4657. doi:10.1016/j.csda.2008.02.030.

Vovk, V., 2012a. Conditional validity of inductive conformal predictors. Journal of Machine Learning Research - Proceedings Track 25, 475–490.

Vovk, V., 2012b. Cross-conformal predictors. CoRR abs/1208.0806.

Vovk, V., Gammerman, A., Shafer, G., 2005. Algorithmic Learning in a Random World. Springer, New York.

Xu, L., Schuurmans, D., 2005. Unsupervised and semi-supervised multi-class support vector machines, in: Veloso, M.M., Kambhampati, S. (Eds.), AAAI, AAAI Press / The MIT Press. pp. 904–910.

Zhu, X., Gisbrecht, A., Schleif, F.M., Hammer, B., 2012. Approximation techniques for clustering dissimilarity data. Neurocomputing 90, 72–84.

Zhu, X., Goldberg, A.B., 2009. Introduction to semi-supervised learning. Synthesis Lectures on Artif. Intell. and Machine Learning 3, 1–130.

Zhu, X., Schleif, F.M., Hammer, B., 2013. Semi-supervised vector quantization for proximity data, in: ESANN, pp. 89–94.