Supervised data analysis and reliability estimation with exemplary application for spectral data

Frank-Michael Schleif[a,*,1], Thomas Villmann[b], Matthias Ongyert[c]

[a] *University of Leipzig, Department of Medicine, Computational Intelligence Group, Semmelweisstrasse 10, 04103 Leipzig, Germany*

[b] *University of Applied Science Mittweida, Department of Mathematics, Physics and Computer Science, Technikumplatz 17, 09648 Mittweida, Germany*

[c] *Steria Mummert Consulting AG, Neumarkt 10, 04109 Leipzig, Germany*

Summary

The analysis and classification of data, is a common task in multiple fields of experimental research such as bioinformatics, medicine, satellite remote sensing or chemometrics leading to new challenges for an appropriate analysis. For this purpose different machine learning methods have been proposed. These methods usually do not provide information about the reliability of the classification. This however is a common requirement in e.g. medicine and biology. In this line the present contribution offers an approach to enhance classifiers with reliability estimates in the context of prototype vector quantization. This extension can also be used to optimize precision or recall of the classifier system and to determine items which are not classifiable. This can lead to significantly improved classification results. The method is exemplarily presented on satellite remote spectral data but is applicable to a wider range of data sets.

Key words: spectral analysis, reliability estimation, classifier optimization, conformal prediction, rejection region, conformal thresholding

*Corresponding author at: Computational Intelligence Group, Department of Medicine, University of Leipzig, Semmelweisstrasse 10, 04103 Leipzig, Germany. Tel.: +49(0)3419718955; Fax.: +49(0)3419718849.

Email addresses: schleif@informatik.uni-leipzig.de (Frank-Michael Schleif)

URL: http://www.uni-leipzig.de/~compint (Frank-Michael Schleif)

[1] ACKNOWLEDGMENTS: The authors are grateful to Alex Gammerman and Luo Zhiyuan for helpful discussions on Hedging predictions (both Computer Learning Research Center (CLRC), Royal Holloway, University of London, UK). Further we would like to thank Erzsebet Merenyi for stimulating discussions on the analysis of satellite remote sensing data (Department of Electrical and Computer Engineering, Rice University, Houston, TX, USA) and Michel Verleysen for fruitful discussions on functional data analysis (Department of Electricity, Universite catholique de Louvain). This work was supported in parts by the Federal Ministry of Education and Research under PNR: 934000-545, and a research contract with the Bruker Daltonik GmbH, Leipzig, Germany.

1. Introduction

The generation of classification models, is a common task in multiple fields of experimental research such as bioinformatics, medicine, satellite remote sensing or chemometrics [23, 25]. Reliability estimation of the obtained classification models is frequently required. In traditional statistics this information is usually provided by significance levels whereas for machine learning models such estimators are rare. Recently a learning theoretical approach for this problem was proposed by [33], called *conformal prediction*. We adapt this model for utilization of prototype-based classifiers like Learning Vector Quantization (LVQ) namely Supervised Relevance Neural Gas (SRNG) [32]. This model classifies each sample prototype-based and additionally offers a level of its classification reliability.

We demonstrate the capabilities of this method for classification of satellite remote sensing spectral data. For this type of data true color images allow a visual control of classification accuracy [8]. In this specific application another aspect is given by the functional character of the data which requires an adequate handling [19, 23, 29]. In particular we favor the usage of functional distances for similarity determination instead of standard euclidean metric.

The paper is organized as follows. First we briefly introduce the main ingredients for our model. We start with a short review of the Supervised Relevance Neural GAS (SRNG) for prototype based classification [32] and demonstrate how this approach can deal with different types of metrics including a functional metric. Thereafter the method of conformal prediction [33] is discussed in the light of prototype based classifiers. It is shown how a thresholding approach can be employed in the analysis of functional spectral data combining the two measures of confidence and credibility as derived from conformal predictions. The experimental settings of our approach are defined. In the experimental section we apply our framework on data obtained from remote satellite imaging. The data are analyzed in detail and some new findings are made which have not been reported so far. The paper is closed by a summary and a discussion of open points and research directions.

2. Material and Methods

2.1. Supervised Neural Gas for functional Data

Supervised Neural Gas (SNG) [10] is considered as a representative for prototype based classification approaches as introduced by Kohonen [15]. Different prototype classifiers have been proposed so far [10, 15, 21] as improvements of the original approach. The SNG combines the idea of neighborhood cooperativeness during learning from the unsupervised Neural Gas algorithm (NG)introduced in [18] with the supervised Generalized learning vector quantizer (GLVQ) as given in [21]. Subsequently we give the basic notations and some remarks to the integration of alternative metrics into Supervised Neural Gas (SNG). Details on SNG including convergence proofs can be found in [10].

Let us first clarify some notations: Let $c_{\mathbf{v}} \in \mathcal{L}$ be the label of input $\mathbf{v}$, $\mathcal{L}$ a set of labels (classes) with $\#\mathcal{L} = N_{\mathcal{L}}$. Let $V \subseteq \mathbb{R}^{D_V}$ be a finite set of inputs $\mathbf{v}$. LVQ uses a fixed number of prototypes (weight vectors, codebook vectors) for each class. Let $\mathbf{W} = \{\mathbf{w}_{\mathbf{r}}\}$ be the set of all codebook vectors and $c_{\mathbf{r}}$ be the class label of $\mathbf{w}_{\mathbf{r}}$. Furthermore, let $\mathbf{W}_c = \{\mathbf{w}_{\mathbf{r}} | c_{\mathbf{r}} = c\}$ be the subset of prototypes assigned to class $c \in \mathcal{L}$ and $\mathbf{W}_c$ is the cardinality of $\mathbf{W}_c$.

In vector quantization a stimulus vector $\mathbf{v} \in V$ is mapped onto that neuron $\mathbf{s} \in A$ the pointer $\mathbf{w}_s$ of which is closest to the presented stimulus vector $\mathbf{v}$,

$$\Psi^{\lambda}_{V \to \mathcal{A}} : \mathbf{v} \mapsto \mathbf{s}(\mathbf{v}) = \mathrm{argmin}_{\mathbf{r} \in A} d^{\lambda}(\mathbf{v}, \mathbf{w}_{\mathbf{r}}) \tag{1}$$

$d^{\lambda}(\mathbf{v}, \mathbf{w})$ is an arbitrary differentiable similarity[2] measure, which may depend on a parameter vector λ. For the moment we take λ as fixed. The neuron $\mathbf{s}(\mathbf{v})$ is called winner or best matching unit. The subset of the input space

$$\Omega^{\lambda}_{\mathbf{r}} = \{\mathbf{v} \in V : \mathbf{r} = \Psi_{V \to A}(\mathbf{v})\} \tag{2}$$

which is mapped to a particular neuron $\mathbf{r}$ according to (1), forms the (masked) receptive field of that neuron forming a Voronoi tessellation. If the class information of the weight vector is used, the boundaries $\partial\Omega^{\lambda}_{\mathbf{r}}$ generate the decision boundaries for classes. A training algorithm should adapt the prototypes such that for each class $c \in \mathcal{L}$, the corresponding codebook vectors $\mathbf{W}_c$ represent the class as accurately as possible. This means that the set of points in any given class $V_c = \{\mathbf{v} \in V | c_{\mathbf{v}} = c\}$, and the union $\mathcal{U}_c = \bigcup_{\mathbf{r}|_{\mathbf{w}_{\mathbf{r}} \in \mathbf{W}_c}} \Omega_{\mathbf{r}}$ of receptive fields of the corresponding prototypes should differ as little as possible.

We suppose to have m data vectors $\mathbf{v}_i$. As pointed out in [10], the neighborhood learning for a given input $\mathbf{v}_i$ with label c is applied to the subset $\mathbf{W}_c$. The respective cost function is

$$Cost_{SNG}(\gamma) = \sum_{i=1}^{m} \sum_{\mathbf{r}|\mathbf{w}_{\mathbf{r}} \in \mathbf{W}_{c_i}} \frac{h_{\gamma}(\mathbf{r}, \mathbf{v}_i, \mathbf{W}_{c_i}) \cdot f(\mu_{\lambda}(\mathbf{r}, \mathbf{v}))}{C(\gamma, K_{c_i})} \tag{3}$$

with $f(x) = (1 + \exp(-x))^{-1}$, $h_{\gamma}(\mathbf{r}, \mathbf{v}, \mathbf{W}) = \exp\left(-\frac{k_{\mathbf{r}}(\mathbf{v}, \mathbf{W})}{\gamma}\right)$ and $\mu_{\lambda}(\mathbf{r}, \mathbf{v}) = \frac{d^{\lambda}_{\mathbf{r}} - d^{\lambda}_{\mathbf{r}_-}}{d^{\lambda}_{\mathbf{r}} + d^{\lambda}_{\mathbf{r}_-}}$ whereby $d^{\lambda}_{\mathbf{r}_-}$ is defined as the squared distance to the best matching prototype but labeled with $c_{\mathbf{r}_-} \neq c_{\mathbf{v}}$, say $\mathbf{w}_{\mathbf{r}_-}$ and $d^{\lambda}_{\mathbf{r}} = d^{\lambda}(\mathbf{v}, \mathbf{w}_{\mathbf{r}})$. For a detailed formal analysis of SNG we refer to [10].

2.1.1. Incorporation of a functional metric to SNG

As pointed out before, the similarity measure $d^{\lambda}(\mathbf{v}, \mathbf{w})$ is only required to be differentiable with respect to λ and $\mathbf{w}$. The triangle inequality has not to be fulfilled necessarily. This leads to a great freedom in the choice of suitable measures and allows the usage of non-standard metrics in a natural way. We now review a functional metric as given in [16]. This type of metric is especially suited in case of functional data because it takes consecutive points into account which is a natural property in case of functional data. In [16] a successful application of this

[2] A similarity measure is a non-negative real-valued function, which, in contrast to a distance measure does not necessarily fulfill the triangle inequality and the symmetry property.

type of metric was shown using the well known *tecator* data provided in [2].

The corresponding derivations can be plugged into the above equations leading to SNG with a functional metric, whereby the data are functions represented by vectors and, hence, the vector dimensions are spatially correlated. A similar situation can be observed for satellite spectra as demonstrated in [26].

Common vector processing does not take the spatial order of the coordinates into account. As a consequence, the functional aspect of spectral data is lost. For proteom spectra the order of signal features (peaks) is due to the nature of the underlying biological samples and the measurement procedure. The masses of measured chemical compounds are given ascending and peaks encoding chemical structures with a higher mass follows chemical structures with lower masses. In addition multiple peaks with different masses may encode parts of the same chemical structure and hence are correlated.

Lee proposed a distance measure taking the functional structure into account by involving the previous and next values of x_i in the i-Th term of the sum, instead of x_i alone. Assuming a constant sampling period τ, the proposed norm is:

$$\mathcal{L}_p^{FCC}(\mathbf{v}) = \left(\sum_{k=1}^{D} \left(A_k(\mathbf{v}) + B_k(\mathbf{v}) \right)^p \right)^{\frac{1}{p}} \tag{4}$$

with

$$A_k(\mathbf{v}) = \begin{cases} \frac{\tau}{2}|v_k| & \text{if } 0 \leq v_k v_{k-1} \\ \frac{\tau}{2} \frac{v_k^2}{|v_k|+|v_{k-1}|} & \text{if } 0 > v_k v_{k-1} \end{cases} \tag{5}$$

$$B_k(\mathbf{v}) = \begin{cases} \frac{\tau}{2}|v_k| & \text{if } 0 \leq v_k v_{k+1} \\ \frac{\tau}{2} \frac{v_k^2}{|v_k|+|v_{k+1}|} & \text{if } 0 > v_k v_{k+1} \end{cases} \tag{6}$$

are respectively of the triangles on the left and right sides of x_i. Just as for L_p, the value of p is assumed to be a positive integer. At the left and right ends of the sequence, x_0 and x_D are assumed to be equal to zero. The derivatives for the functional metric taking $p = 2$ are given in [16]. Now we consider the scaled functional norm where each dimension v_i is scaled by a parameter $\lambda_i > 0$ $\lambda_i \in (0, 1]$ and $\sum_i \lambda_i = 1$. Then the scaled functional norm is:

$$\mathcal{L}_p^{FCC}(\lambda\mathbf{v}) = \left(\sum_{k=1}^{D} \left(A_k(\lambda\mathbf{v}) + B_k(\lambda\mathbf{v}) \right)^p \right)^{\frac{1}{p}} \tag{7}$$

with

$$A_k(\lambda\mathbf{v}) = \begin{cases} \frac{\tau}{2}\lambda_k|v_k| & \text{if } 0 \leq v_k v_{k-1} \\ \frac{\tau}{2} \frac{\lambda_k^2 v_k^2}{\lambda_k|v_k|+\lambda_{k-1}|v_{k-1}|} & \text{else} \end{cases} \tag{8}$$

$$B_k(\lambda\mathbf{v}) = \begin{cases} \frac{\tau}{2}\lambda_k|v_k| & \text{if } 0 \leq v_k v_{k+1} \\ \frac{\tau}{2} \frac{\lambda_k^2 v_k^2}{\lambda_k|v_k|+\lambda_{k+1}|v_{k+1}|} & \text{else} \end{cases} \tag{9}$$

The corresponding derivations can be found in [26]. Using this parametrization one can emphasize/neglect different parts of the function for classification. This distance measure can be put into SNG as shown above and has been applied subsequently in the analysis of the spectra. SNG with a parametrized metric is subsequently referred as SRNG. The functional metric will be just referred as FUNC and will be always used with metric adaptation if not stated otherwise.

2.2. Conformal prediction - Reliability estimation

In the analysis of spectral data the determination of a classifier is a difficult task. The data are functional and in general high dimensional and only few assumptions about the specific nature (e.g. distributions) of the data can be made. Due to this reasons an analysis using classical statistics such as statistical tests for group comparisons, Linear discriminant analysis or partial least squares methods (see e.g. [12] for an overview) can not be applied, in general. Alternatively so called soft methods, with only minor assumptions about the specific properties of the data, are used. Typical representants of these type are prototype based classifiers such as the formerly mentioned Supervised Relevance Neural Gas [11] and variants or the famous Support Vector Machines (SVM) [28]. These methods have already proven to be appropriate for the analysis of spectral data [22, 24] also in case of very high dimensional complex problems. A drawback of these methods, in contrast to classical statistics, is the lack of reliability measures, which similar to the well known $p-$ or q-values can be used to judge the significance or reliability of a taken decisions. Only few attempts were made to give reliability estimates for these methods (see e.g. [5, 7]). Thereby the reliability estimate can be helpful to judge on the reliability of a decision but also in a more generic framework to improve the overall performance of the classifier. Reliability sometimes also referred as confidence, has been subject of a quite new theory called conformal prediction as introduced in [33]. These theory directly aims on the determination of confidence and as a second measure credibility of classifier decisions. The stability of the algorithm presented here follows immediately from the stability analysis of conformal prediction as provided in [33] because our approach is directly derived from it. According to this analysis the algorithm is stable in stochastic sense. Thereby the type of the classifier is not much limited but it is assumed that a so called *non-conformity* measure is available, revealing relevant knowledge of the classification decision. Subsequently we introduce the relevant parts of the conformal prediction approach and detail how it can be used in the analyzed experiments.

2.2.1. Settings

For the introduction to conformal prediction we switch to a more practical notation. Let the training data $z_i = (x_i, y_i) \in \mathbb{Z} = \mathbb{X} \times \mathbb{Y}$, $i = 1 \ldots L$ be given by the data points $x_i \in \mathbb{X} = \mathbb{R}^{D_V}$ and their labels $y_i \in \mathbb{Y} = \{1, 2, \ldots, N_{\mathcal{L}}\}$ belonging to one of the $N_{\mathcal{L}}$ classes. Furthermore let x_{L+1} be a new observed data point with unknown label y_{L+1} and classification / prediction $\hat{y}_{L+1}$.

For conformal prediction we need the following terms: *conformal prediction algorithm*, *prediction region*, *nonconformity measure*,*r-value*, *error rate* ϵ, *confidence & credibility*, *exchangeability* and *validity* which are explained in more detail subsequently.

In our setting the *conformal prediction algorithm* computes for the given training data $(z_i)_{i=1,\ldots,L}$, the observed data

point x_{L+1} and a chosen error rate ϵ the *prediction region* $\Gamma^\epsilon(z_1, \ldots, z_l, x_{L+1}) \subset \mathbb{Y}$ consisting of 0 to n possible labels. The applied method ensures us that if the z_i are exchangeable[3] then

$$P(y_{L+1} \notin \Gamma^\epsilon(z_1, \ldots, z_l, x_{L+1})) \leq \epsilon \tag{10}$$

holds asymptotically for $L \to \infty$ for each distribution of $\mathbb{Z}$. One says that the predictor is *asymptotically valid*. It is important to mention, that the probability is an unconditional one what means, that if we repeat the process of drawing samples x_{L+1} and generating Γ^ϵ n times we will find with respect to statistical fluctuations that in less than $\epsilon \times n$ cases the real label y_{L+1} is not under the predicted labels of Γ^ϵ. It does not mean, that for a certain x_{L+1} y_{L+1} is in Γ^ϵ with probability $> 1 - \epsilon$. As counter example one considers an empty prediction region for which this conditional probability becomes exactly zero. Such cases may happen if the observed sample x_{L+1} is extremely rare in $\mathbb{X}$ ($\mathbb{Z}$) in such a way, that it is not typical with respect to the given training data. So it does not effect the error rate (10).

The conformal prediction algorithm is illustrated in Figure 1 and (11)-(15). The non conformity measure $A(D_i, z_i)$ is used to calculate a non conformity value α_i that estimates how badly z_i fits to the representative data D_i={$z_1, \ldots, z_{L+1}$ z_i}. For a certain prediction $\hat{y}$ one calculates its *r-value* by adding $z_{L+1} = (x_{L+1}, \hat{y})$ (11) to the training data (12), calculating the α_i by checking each z_i against the rest (13) and retrieving $r_{\hat{y}}$ as the relative amount of samples that are as bad or worse conformal to all remaining examples (14). For a reasonable non conformity measure A α_{L+1} should be small if x_{L+1} is typical and the prediction $\hat{y}$ is right and typical for the data point x_{L+1}. This involves a high $r_{\hat{y}}$ and a membership of $\hat{y}$ in Γ^ϵ for most ϵ. If x_{L+1} is untypically or $\hat{y}$ is wrong A should detect this mismatch and generate a big α_{L+1}. In this case only a few examples of the training data have a greater nc-value such that $r_{\hat{y}}$ will be quiet small. As a consequence $\hat{y}$ will only be contained in Γ^ϵ for smaller ϵ.

$$\forall \hat{y} \in \mathbb{Y}$$

$$z_{L+1} \stackrel{\text{def}}{=} (x_{L+1}, \hat{y}) \tag{11}$$

$$\forall i \in \{1, \ldots L+1\}$$

$$D_i = \{z_1, \ldots, z_{L+1}\} \setminus \{z_i\} \tag{12}$$

$$\alpha_i = A(D_i, z_i) \tag{13}$$

$$r_{\hat{y}} = \frac{|\{\alpha_i : \alpha_i \geq \alpha_{L+1}|}{L+1} \tag{14}$$

$$\Gamma^\epsilon = \{y : r_{\hat{y}} > \epsilon\} \tag{15}$$

2.2.2. Non Conformity Measure

As explained in the previous section the non conformity measure should evaluate the fit of a test example z_i to representative data D_i. It is those part of the method that can incorporate detailed knowledge about the data distribution. In our setting its

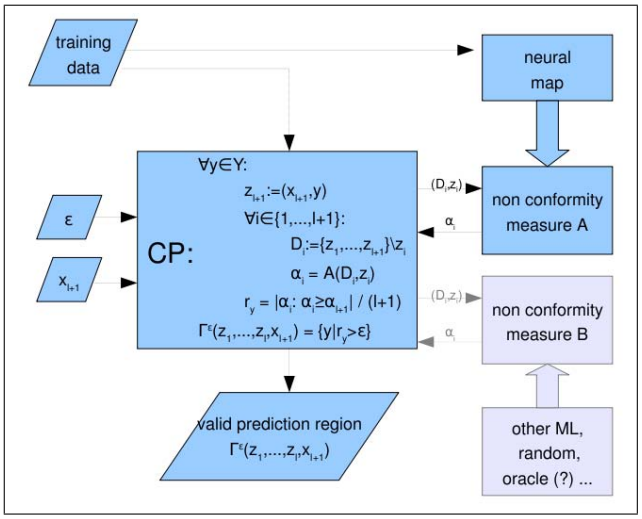

Figure 1: Schema of cooperating parts in conformal prediction. As it can be seen the conformal prediction algorithm uses the training data and an arbitrary non conformity measure to generate valid prediction regions. Normally one wants to incorporate a neural map to calculate meaningful nc values.

the place to use the learned neural map (as A in Figure 1). Nevertheless one can use any arbitrary real valued function [4] (10) but maybe with negative impact on prediction efficiency (see 2.2.3). To apply this on a prototype based situation one has to think about how the match between arbitrary z_i and D_i could be managed. A obvious solution, to learn a neural map with each individual D_i and match z_i against it, would entail high computational costs, because this has to be done for all the L *one left out* multi-sets D_i for each of the $N_{\mathcal{L}}$ test objects $(x_{L+1}, \hat{y}_{j=1,\ldots,N_{\mathcal{L}}})$ in the conformal prediction algorithm. Our solution lies in the arbitrariness of A [5]. We can ignore matching z_i exactly against D_i but instead use the whole training data without z_{L+1}, therefore learning must be performed only once. The lost amount of information will be small if the number of training data is high, so that adding z_i but leaving out z_{L+1} will not change learning results dramatically.

Obvious measures for prototype based methods are k nearest neighbor methods for example:

$$\alpha_i = \frac{\sum_{j=1}^{k} d_{ij}^{+}}{\sum_{j=1}^{k} d_{ij}^{-}} \tag{16}$$

with d_{ij}^{+} being the given distance between x_i and the j-th nearest prototype with identical label y_i and d_{ij}^{-} being the given distance between x_i and the j-th nearest prototype with a label different to y_i. Other measures are conceivable.

2.2.3. Prediction Region

The prediction region $\Gamma^\epsilon(z_1, \ldots, z_l, x_{l+1})$ stands in the center of conformal prediction. It contains for a given error rate ϵ the

[3] *exchangeability* is a weak condition: e.g. independently and identically distributed random variables are exchangeable [33]

[4] *Any* measureable function on $\mathbb{Z}^{(*)} \times \mathbb{Z}$ taking values in extended real line is called a non conformity measure

[5] This could be a constant function or a relatively to z_i fixed random value leaving out D_i at all

possible labels of $\mathbb{Y}$ that ensures (10). But how can we use it and how will it change for different values ϵ and different A?

Suppose we are using a meaningful non conformity measure A. If we would set ϵ nearly to 0 then conformal prediction has to produce Γ's that makes nearly no error at all, which can only be satisfied if Γ contains all possible labels. Of course such a prediction bears no information. But if we slowly raise ϵ we allow some rare errors to occur and as a benefit the conformal prediction algorithm excludes some unlikely labels from our prediction region and increasing its information content. In detail those $\hat{y}$ are discarded whose r-value is less equal ϵ, that means only a few z_i are as *non conformal* as $z_{L+1} = (x_{L+1}, \hat{y})$. This is a strong indicator that z_{L+1} does not belongs to the distribution $\mathbb{Z}$ and so $\hat{y}$ seams not to be the right label. If one further raises ϵ only those $\hat{y}$ will remain in Γ that can produce a high r-value meaning that the corresponding z_{L+1} is rated as very typical by A.

So one can trade the error rate against information content. The most useful prediction is those containing exactly one label. Therefore two error rates are of particular interest, ϵ_1 being the smallest ϵ and ϵ_2 being the greatest ϵ so that $|\Gamma^\epsilon| = 1$. ϵ_2 is the r-value of the best and ϵ_1 is the r-value of the second best label y. So the prediction can be summarized as

$$\zeta(\text{confidence}) = 1 - \epsilon_1 = 1 - r_{y_{2\text{nd}}} \tag{17}$$

$$\kappa(\text{credibility}) = \epsilon_2 = r_{y_{1\text{st}}} \tag{18}$$

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 (un)typical and not an outlier.

As mentioned in 2.2.2 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 great error rates, meaning that $\epsilon_2 - \epsilon_1$ being large for typical data. That means, that a good measure can give useful information already for an ensured (10) small error rate ϵ_1 and on the other hand one would have to face up a high average error rate ϵ_2 to exclude the right label from the prediction region.

For practical applications here we are only interested on prediction regions with $|\Gamma^\epsilon| = 1$. For these regions natural measures of confidence and credibility become available by application of the conformal prediction methodology. These two values combined with the conformal prediction (predicted label) can be employed subsequently not only to estimate the pointwise reliability of the classification but also to improve the classifier system, by means of a thresholding approach.

2.2.4. Recall/Precision/Thresholding with Conformal Prediction

Recall-Precision graphs are very common in the field of information retrieval (IR) to estimate the performance of the considered IR-system [17]. Here we use this graphs in combination with a thresholding to improve the overall classifier performance. Thereby the recall $\mathcal{R}$ and the precision $\mathcal{P}$ are defined as:

ID	frequency range	label	resolution	bits
1	0.45–0.52	blue	30×30	8
2	0.52–0.60	green	30×30	8
3	0.63–0.69	red	30×30	8
4	0.76–0.90	near IR	30×30	8
5	1.55–1.75	mid IR	30×30	8
7	2.08–2.35	mid IR	30×30	8

Table 1: Characteristics of the Landsat imaging device

$$\mathcal{R} = \frac{C}{L} \quad \mathcal{P} = \frac{C^+}{C} \tag{19}$$

with C as the number of classified (not rejected) data points and C^+ as the number of correct classified data points. Further we introduce a so called rejection set $\mathcal{S}_r$ and an acceptance set $\mathcal{S}_a$

$$\mathcal{S}_r = \{x_i : \zeta_i < \zeta_t \vee \kappa_i < \kappa_t\} \quad \mathcal{S}_a = \{x_i : \zeta_i \geq \zeta_t \wedge \kappa_i \geq \kappa_t\} \tag{20}$$

with ζ_t/κ_t as the user defined confidence/credibility thresholds. For a chosen threshold pair ζ_t/κ_t the definitions for recall $\mathcal{R}$ and the precision $\mathcal{P}$ are adapted in the natural way using the acceptance region such as the thresholded recall $\mathcal{R}_{(\zeta_t,\kappa_t)}$ and the thresholded precision $\mathcal{P}_{(\zeta_t,\kappa_t)}$ become:

$$\mathcal{R}_{(\zeta_t,\kappa_t)} = \frac{|\mathcal{S}_a|}{L} \quad \mathcal{P}_{(\zeta_t,\kappa_t)} = \frac{|\mathcal{S}_a|^+}{|\mathcal{S}_a|} \tag{21}$$

with $|\mathcal{S}_a|$ as the number of classified (not rejected) data points in the acceptance set and $|\mathcal{S}_a|^+$ as the number of correct classified data points in the acceptance set. An example of such a recall/precision graph for different thresholds ζ_t/κ_t is given in Figure 3.

3. Data description

We applied the algorithm to a large real world data set: a multi-spectral LANDSAT TM satellite image of the Colorado area. Airborne and satellite-borne remote sensing spectral images consist of an array of multi-dimensional vectors (spectra) assigned to particular spatial regions (pixel locations) reflecting the response of a spectral sensor at various wavelengths. A spectrum is a characteristic pattern that provides a clue to the surface material within the respective surface element. The utilization of these spectra includes areas such as mineral exploration, land use, forestry, ecosystem management, assessment of natural hazards, water resources, environmental contamination, biomass and productivity; and many other activities of economic significance [20].

Spectral images can formally be described as a matrix $\mathbf{S} = \mathbf{v}^{(x,y)}$, where $\mathbf{v}^{(x,y)} \in \mathbb{R}^{D_\mathcal{V}}$ is the vector (spectrum) at pixel location (x, y) with $D_\mathcal{V} = 6$. The description of the spectral bands is given in Table 1. The elements $v_i^{(x,y)}$, $i = 1 \ldots D_\mathcal{V}$ of spectrum $\mathbf{v}^{(x,y)}$ reflect the responses of a spectral sensor at a suite of wavelengths [4]. The spectrum is a characteristic fingerprint pattern that identifies the averaged content of the surface material within the area defined by pixel (x, y). The individual

2-dimensional image $\mathbf{S}_i = v_i^{(x,y)}$ at wavelength i is called the ith image band. The data density $\mathcal{P}(\mathcal{V})$ may vary strongly within the data. Sections of the data space can be very densely populated while other parts may be extremely sparse, depending on the materials in the scene and on the spectral bandpasses of the sensor.

In addition to dimensionality and volume, other factors, specific to remote sensing, can make the analyses of hyperspectral images even harder. For example, given the richness of data, the goal is to separate many cover classes, however, surface materials that are significantly different for an application may be distinguished by very subtle differences in their spectral patterns. The pixels can be mixed, which means that several different materials may contribute to the spectral signature associated with one pixel. This may lead to an unsafe prediction. Training data may be scarce for some classes, and classes may be represented very unevenly (see Table 2). All the above difficulties motivate research into advanced and novel approaches. However it should be mentioned, that the presented approach is not limited to this type of application, but can be applied to a wider range of (spectral) or feature driven imaging analysis such as MALDI-Imaging [8, 30], raman spectroscopy of tissue slices or the analysis of microscopic images [3, 31] to name just a few.

The image was taken very close to colorado springs using satellites of LANDSAT-TM type[6]. The satellite produced pictures of the earth in 7 different spectral bands. The ground resolution in meter is 30×30 for the bands $1-5$ and band 7. Band 6 (thermal band) has a resolution of 60×60 only and, therefore, it is often dropped. The LANDSAT TM bands were strategically determined for optimal detection and discrimination of vegetation, water, rock formations and cultural features within the limits of broad band multi-spectral imaging. The spectral information, associated with each pixel of a LANDSAT scene is represented by a vector $\mathbf{v} \in \mathcal{V} \subseteq \mathbb{R}^{D_\mathcal{V}}$ with $D_\mathcal{V} = 6$. The aim of any classification algorithm is to subdivide this data space into subsets of data points, with each subset corresponding to specific surface covers such as forest, industrial region, etc. The feature categories are specified by prototype data vectors (training spectra). Additionally, the Colorado image is completely labeled by experts [7]. There are 14 labels describing different vegetation types and geological formations. The detailed labeling of the classes is given in Table 2, here we also specify the used coloring for the subsequently generated images as obtained from the classification models[8]. The colors where chosen such that similar materials get similar colors in the RGB space. In addition we show plots of the data using the HSV color space whereby the H channel encodes the class ($1-14$ scaled to the full range), S the results of the confidence measure ζ and V the results for the credibility measure κ. Using this setting a perfect recognition/prediction results in colors with high saturation and colorimetry (v - channel), whereas less perfect detected data points reduce the saturation and/or the colorimetry such that they appear darker and more dirty.

Thereby, the label probability varies in a wide range. The size of the image is 1907×1784 pixels [9].

Label	class	R	G	B	ground cover	#pixels
a	1	0	128	0	Scotch pine	581424
b	2	128	0	128	Douglas fir	355145
c	3	128	0	0	Pine / fir	181036
d	4	192	0	192	Mixed pines	272282
e	5	0	255	0	Mixed pines	144334
f	6	255	0	0	Aspen/Pines	208152
g	7	255	255	255	No veg.	170196
h	8	128	60	0	Aspen	277778
i	9	0	0	255	Water	16667
j	10	0	255	255	Moist meadow	97502
k	11	255	255	0	Bush land	127464
l	12	255	128	0	Pastureland	267495
m	13	0	128	128	Dry meadow	675048
n	14	128	128	128	Alpine veg.	27556
o	15	0	0	0	misclassif.	-

Table 2: Short description of the different classes of the satellite image, the used similarity based coloring (in RGB space) and the number of pixel present for each class.

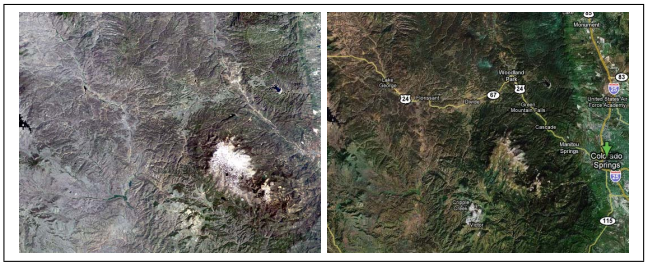

Figure 2: True coloring of the satellite data. Left the coloring in accordance to the RGB channels of the original data (data approx 1990), right a up to data image as obtained from [9]

4. Experiments and Results

To get a valid setting of the experiments the data have been split into multiple sets, such that three data splits are obtained. These sets are named as *tuning set* (TRS) with 1500 data points per class, the *crossvalidation set* (CRS) with 3.381.079 data points has been used in a 5×5 cross validation, thereby we call each test set as the *rest set* (RS) of this crossvalidation. For the set TRS and CRS the points have been selected randomly from the original data set such that each class is equally represented.

[6] Thanks to M. Augusteijn (University of Colorado) for providing this image.

[7] Its known that an exact ground truth labeling is complicated to obtain in this field and also effects such as the granularity may significantly effect the data and hence the label precision (e.g snow may appear as water). Under this light imprecision of the labeling is a general problem for multiple data sets.

[8] For better visualization in b/w the misclassifications are sometimes also given with white coloring. Due to some specifics of the given labeling with respect to the information encoded in the data, as pointed out in the text, the class 7 (also white) is often subject of misclassifications, anyway. Colored versions of the image can be obtained from the corresponding author.

[9] Thereby 9 pixel have a unclear label and have been removed.

method	$\mathbf{W}_c^5$	$\mathbf{W}_c^{10}$	$\mathbf{W}_c^{20}$	$\mathbf{W}_c^{50}$	$\mathbf{W}_c^{100}$
EUC-rec	90.50	92.1	93.7	95.0	96.5
EUC-pre	89.8	91.4	92.2	92.2	92.3
SEUC-rec	91.1	92.5	93.9	95.2	96.5
SEUC-pre	90.5	91.7	92.7	92.4	92.9

Table 3: Tuning results evaluated by recognition and prediction for metric Euc and SEUC varying the map size parameter of SRNG.

The TRS has been used for parameter tuning studies, thereby the data points have been split into a training and a test set such that 1000 points where used for training and 500 points to determine the optimal parameters. In additional experiments it has been verified that alternative set sizes of the cross-validation do not change the results significantly as long as the data statistics is sufficiently preserved. For details on this topic we refer to [1].

The parameter tuning part has been done for SRNG with standard and scaled Euclidean metric (SNG/SRNG). The identified optimal settings for the basic parameters of S(R)NG with conformal prediction were transferred to the other models. The SRNG with appropriate parametrization has been subsequently applied to the prior not used data in the CRS data set and evaluated in a 5×5-fold crossvalidation scheme. From the crossvalidation runs, showing very small variances between the different models, we choose the first model to label the whole satellite image. In the following we detail the three stages of our experiments, followed by an additional analysis employing conformal prediction in a thresholding experiment.

4.1. Parameter tuning

As already mentioned the SRNG parameters have been optimized on a very small subset of the original data set using the TRS split. Thereby the following parameters have been subject of optimization: map size, as the number $\mathbf{W}_c$ of prototypes per class in a range of $\{5, 10, 20, 50, 100\}$ and the parameter k of the k-NN based non-conformity measure. The remaining parameters of SRNG have been chosen in accordance to [32] with 200 training cycles for each experiment. First we analyzed the effect of different map sizes, as shown in Figure 3 using precision/recall graphs we also took the prediction accuracy of the model (on the test set of TRS) to judge the appropriate size. We observed that for a fixed $k = 1$ of the non-conformity measure a map size of 100 would give best results. However we found also that already 10 prototypes per class constitute a similar performance, therefore a map size of 10 balancing performance and model complexity was chosen as the final setting. Fixing the model size of 10 prototypes per class we varied the parameter k of the non-conformity measure in a range of $\{1, 3, 5\}$. Again we employed the recall/precision graphs and observed that for a $k = 1$ the dispersion of the overall precision was optimal. It should be mentioned that for the other metrics these parameters have been found to be stable as well as depicted in Table 3.

The use of Recall/Precision graphs motivates the use of a threshold to balance between recall and precision (see Figure 3). This of course is very problem dependent and should probably not be automated[10]. Thereby the conformal prediction approach reports the reliability parameters (ζ, κ) for each data point as ideal candidates for thresholding. Here we determined the thresholding parameters for three points (95% recall, break-even, end) point using the first model of the crossvalidation part (a model with optimized parameters) given in Table 4. The 95% recall can be considered as a natural criterion which allows to omit 5% of the points, occurring quite often for the analysis of real data. The second point in our analysis is the break-even point, which can be considered at that point of the recall/precision graph at which a break in the recall/precision graph can be found (e.g. a ascent of ≈ 1 for a tangent fitted against the graph). The third point is an extreme of the graph at which a further removement of points does not significantly improve the precision of the classifier[11].

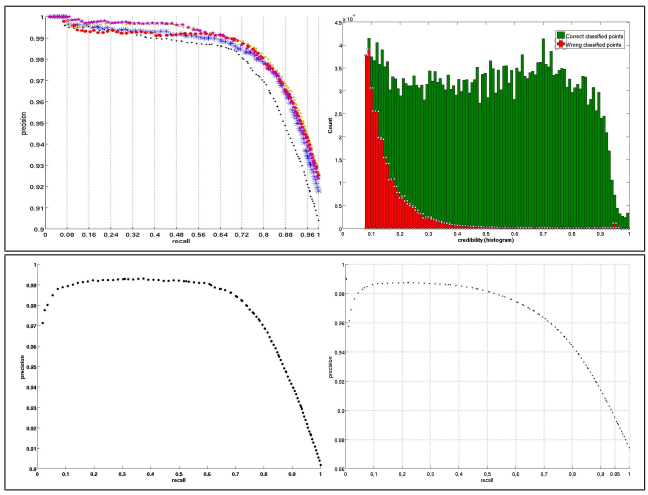

Figure 3: Recall/Precision plots for the different map sizes using SRNG (without thresholding). The curves are given as: map size 5 (dots/black), 10 (stars/blue), 20 (circle/red), 50 (filled star/magenta), 100 (arrow/yellow). The second plot shows a histogram of the credibilities determined for all data points. The third plot shows pairs of (ζ, κ) using the optimal map size 10 with $k = 1$ in the non-conformity measure for scaled Euclidean metric and the last plot a similar curve for the FUNC metric. This plot and the histogram may be employed to determined an appropriate threshold used later on.

The identified thresholding parameters have been used later on to get optimal precision / recall values of the classifier on the remaining (never prior used rest data RS).

4.2. Cross validation results

The SRNG with a map size of $\mathbf{W}_c = 10$ and $k = 1$ for the non-conformity measure was applied on the given satellite remote sensing data using the data subset CRS. Thereby the SRNG was trained using the three considered metrics namely, standard Euclidean metric (EUC), scaled Euclidean

[10] In principle it is possible to get an automatic threshold determination e.g. by line fitting on the recall/precision graph - but this is not the focus of this paper.

[11] It should be mentioned that the generated recall/precision graph may not give a graph as a function but a cloud of distributed point. In this case we determine the convex hull of the cloud. It may also happen that the mentioned three points do not exist but only the 95% point. For our experiments it was always possible to determine the three mentioned points.

point	(κ/ζ)
$EUC_{95\%}$	0.09/0.95
EUC_{break}	0.20/0.98
EUC_{end}	0.44/0.99
$SEUC_{95\%}$	0.12/0.92
$SEUC_{break}$	0.20/0.97
$SEUC_{end}$	0.40/0.99
$FUNC_{95\%}$	0.11/0.92
$FUNC_{break}$	0.11/0.95
$FUNC_{end}$	0.47/0.96

Table 4: Optimal thresholding parameters for (ζ, κ) as obtained by manual inspection of the recall/precision graph of one SRNG model with the different metrics.

metric	Rec.	Pred. mean	Pred. std.
EUC	n.a.	92.6	0.2
SEUC	n.a.	92.3	0.23
FUNC	n.a.	(87.4)	–

Table 5: Crossvalidation results for SRNG with metric Euc, SEUC, FUNC using the optimized parameters, without thresholding. For the FUNC metric only one model has been calculated.

metric (SEUC) and the functional metric (FUNC). The results for recognition and prediction in a 5-fold crossvalidation, without thresholding, are depicted in Table 5.

One observes that the recognition and prediction accuracies are very high with close or above 90%. An analysis of the different confusion matrices supports these finding and shows also that all classes are sufficiently modeled. These findings support the results published in [32]. Interestingly the differences between the different metrics are very small. Nevertheless the metric SEUC allows the identification of discriminating features. A typical ranking of the features for SEUC is obtained as in Table 6 and visualizations of the results using the whole image are shown in 4 and 5.

4.3. Thresholding

While the results found so far are already very promising we were looking for further improvements as well as a more detailed reliability estimation than plain cross validation accuracies or confusion matrices. Therefore we employed the conformal prediction methodology on the remaining test sets RS. The results for recall and precision using the different threshold are given in Table 7 for comparison the thresholding was also applied on the data used in for the first cross validation.

metric	D_1	D_2	D_3
SEUC	$0.08(1E^{-2})$	$0.14(2E^{-2})$	$0.24(2E^{-2})$
FUNC	0.12	0.19	0.20
metric	D_4	D_5	D_6
SEUC	$0.3(1E^{-2})$	$0.24(1E^{-2})$	0.0(0)
FUNC	0.26	0.23	0.0

Table 6: Relevance profile for the metric SEUC and FUNC. For the SEUC mean and standard deviation are shown.

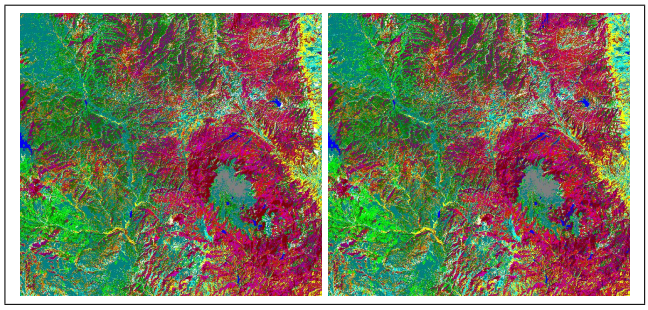

Figure 4: RGB plot for the colorado image. The left plot shows the image with the given labeling and the right plot the same image but with a predicted labeling using conformal prediction and SRNG (EUC). The color table is given as is in Table 2.

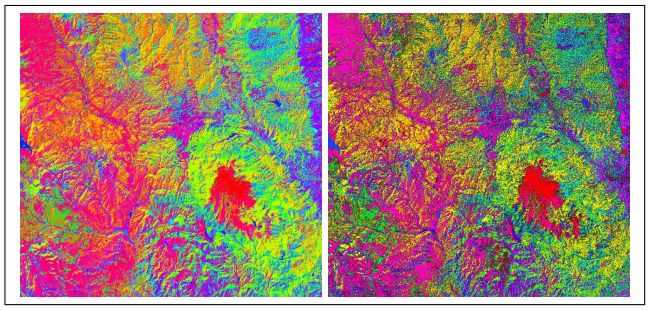

Figure 5: HSV plot for the colorado image. The left plot shows the image with the given labeling $H = labeling/14, S = 1, V = 1$ and the right plot the same image but with a predicted labeling using conformal prediction, and $S = \zeta/median(\zeta)$, $V = \kappa/median(\kappa)$ using one of the determined models. The HSV coloring is easier to interpret using conformal prediction but the coloring is not semantically related to the ground material.

point	Recall/Precision (CRS-1)	Recall/Precision (RS)
$EUC_{95\%}$	95.05/94.59	93.89/93.18
EUC_{break}	79.52/97.99	76.49/97.64
EUC_{end}	56.93/98.90	52.49/99.07
$SEUC_{95\%}$	95.18/92.20	91.54/93.08
$SEUC_{break}$	79.49/97.05	77.80/97.01
$SEUC_{end}$	60.72/99.04	60.45/98.93
$FUNC_{95\%}$	95.11/91.02	94.15/89.83
$FUNC_{break}$	77.11/95.07	79.07/94.64
$FUNC_{end}$	31.60/98.68	48.38/98.24

Table 7: Recall and precision values by application of the thresholding on SRNG-EUC, SRNG-SEUC and SRNG-FUNC using different thresholds for (ζ, κ). It can be seen that there is strong difference between the metrics but the SRNG-FUNC metric performs slightly worse than the others. As expected a more restrictive threshold (reducing the recall) improves the precision up to 99% in this case. Thereby also in case of a larger number of assignments to the *unclassified* state (EUC_{break}, $SEUC_{break}$, $FUNC_{break}$) the structural information of the satellite image is still kept as shown in Figure 6.

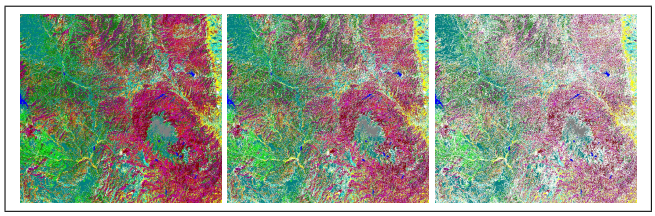

Figure 6: Visualization of the thresholding using SRNG-EUC with different threshold. Its clearly visible that the borders of the classes are subject of uncertainty but also (as pointed out later on) different interesting findings can be made with respect to the safety of a classification considering different thresholds. The first plot shows the classified (recall) pixel at a threshold of $EUC_{95\%}$, the second plot EUC_{break} and the third for EUC_{end} respectively. It can be seen that the number of rejected points (assigned to class 15 - colored white) is increasing. This helps to identify regions which are safe or unsafe with respect to the classification even if the predicted labeling is still correct.

Multiple results can be found in the thresholding approach by considering Table 7 as well as the HSV plots on the differently thresholded RS data (see Figure 6). As a first point we see, that the thresholding improves the precision, not only on the CRS-1 data, which is expected, but also on the prior not used RS data. This observation is valid for all three thresholding points. Considering the Figure 6 we find that removing 50% of the data points still keeps the structural information encoded in the image. The removed points are in general located at the class boundaries which are a natural source of uncertainty with respect to the classification decision. The points removed at the $EUC_{95\%}$ level, again mainly account for class border points but there is also a significant amount of points which appear to be inside of classes. In fact confidence and credibility of the points are in general quite high. This however implies that the classifier was quite sure in its decision, nevertheless these points have been found to be classified wrong. A closer inspection of these points reveals that the most of it belong to the class 7 which is *no vegetation* but are classified to class 14 *alpine vegetation* (not vice versa), this is surprising but considering Figure 2 (left) the effect becomes clear. Miss classification to class 14 do always occur where the true-color image shows snow-coverage, this is due to the fact that the region labeled as *alpine vegetation* (class 14) is completely covered by snow at the time point of taking the satellite image. Hence class 14 should - with respect to the measured data - better be labeled as *snow* than *alpine vegetation*. The effect is depicted in more detail in Figure 7. There it also becomes visible, that this error in the labeling accounts for a larger number of misclassifications. Considering this case high values of confidence and credibility combined with misclassifications maybe in fact an indicator for a wrong labeling or contradictory data (see also Figure 8).

A further region of interesting points is depicted in Figure 9, nearby the Lake George (see Figure 2 (right)).

Thereby multiple misclassifications of class 3 (pine/fir) and class 2 (Douglas fir) to class 9 (water) have been found. Considering both images in Figure 2 we found that the effected pine/fir points are near to water regions. Figure 2 (right) suggests that water level may have changed and hence this miss classification are explainable also.

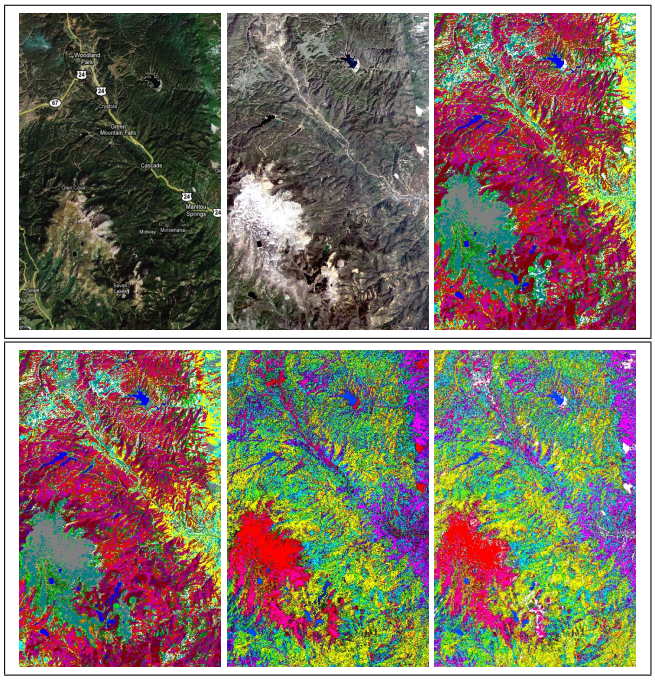

Figure 7: Region with stronger misclassifications related to the alpine-vegetation class (14). Top row shows a zoom into the region close to the alpine region. Left up to date image of the region, next true color view of this regions dating back to approximately 1990, third plot with the RGB coloring of the original labeling of the map. The second row shows the results as obtained by SRNG-EUC with conformal prediction. The plot on the left shows a coloring in RGB with the conformal predictions, the plot in the middle the HSV image using confidence and credibility. Only few dark regions (low credibility/confidence) can be found in the lower part of second plot, second row. Interestingly these items (class 12 pastureland) are not misclassified but only unsafe. But there are also regions of high confidence/credibility which are labeled as class 14 or class 7 (vegetation free).

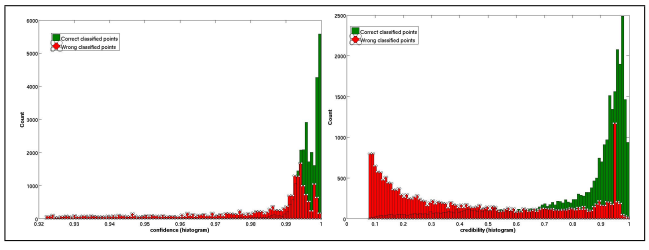

Figure 8: Confidence and credibility histogram plots. The plot helps to identify regions of high confidence with respect to the classification decision. It is also visible that there exist a larger amount with high confidence but wrong labeling - which fits to the findings presented in Figure 7.

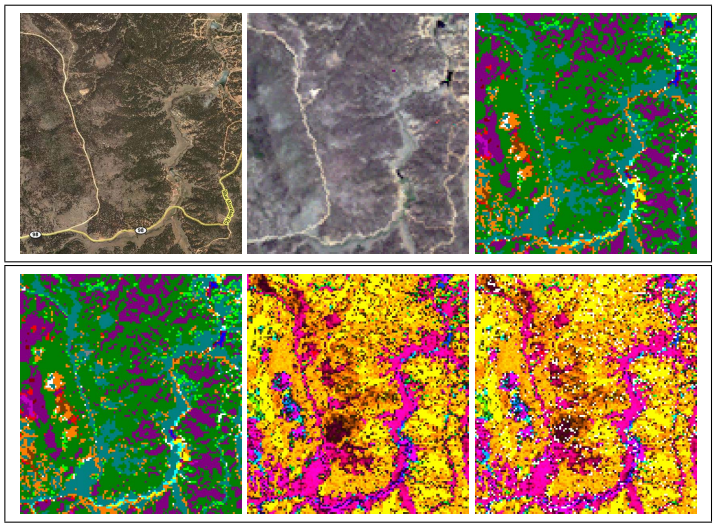

Figure 9: Region with stronger misclassifications. Top row shows a zoom into the region close to lake George. Left up to date image of the region, next true color view of this regions dating back to approximately 1990, third plot with the RGB coloring of the original labeling of the map. The second row shows the results as obtained by SRNG-EUC with conformal prediction. The plot on the left shows a coloring in RGB with the conformal predictions, the plot in the middle the HSV image using confidence and credibility. Here already dark/dirty regions can be detected indicating pixels with unsafe labeling. This is supported by an analysis of the miss classifications (right plot) where misclassified item are colored white. A closer inspection of the region using the true color maps (first row) supports these findings. The discrimination problems occur between class 13 (dry meadow - gray blue), class 12 (grass pastureland - yellow) and class 1 (Scottish fir - dark green).

5. Conclusions

A method for the reliability estimation and optimization of prototype based classifiers has been presented. Thereby the approach incorporates conformal prediction to determine a threshold based on recall/precision analysis and to get reliability estimates for the classification of single items. By use of these measures the performance of the classifier can be tailored with respect to optimal recall and / or precision. This in general improves the interpretability of the generated classifications as shown here exemplarily for satellite remote sensing images. Further a classification can be analyzed with respect to its reliability and also the state of *not classifiable* can be supported. Especially the new class of unclassifiable entries is relevant in multiple classification tasks such as cases involving a classifier based automatic labeling of samples from medicine [26, 27], psychology [13, 14] or bio security domains [6], to name just a few. In these fields the confidence of the classification plays an essential role and the proposed approach offers a better interpretability of the results. In a next step the method will be applied to larger cohorts of spectral data obtained from MALDI-Imaging experiments [30]. Beside of the different promising aspects of the methods there are also some points which could be improved. Currently the choice and parametrization of the non-conformity measure must be optimized by crossvalidation a procedure which is only possible if a sufficient amount of samples is available. In future work, different non-conformity measures should be analyzed with respect to their properties under different conditions to get more generic knowledge about the behavior of a chosen measure. This knowledge could be used to simplify the formerly mentioned parametrization and choice.

References

[1] M. Aupetit. Homogeneous bipartition based on multidimensional ranking. In *Proc. Of European Symposium on Artificial Neural Networks (ESANN'2008)*, pages 259–264, Evere, Belgium, 2008. d-side publications.

[2] C. Blake and C. Merz. UCI repository of machine learning databases., 1998. available at: http://www.ics.uci.edu/ mlearn/MLRepository.html.

[3] C. Brüß, F. Bollenbeck, F.-M. Schleif, W. Weschke, T. Villmann, and U. Seiffert. Fuzzy image segmentation with fuzzy labelled neural gas. In *Proc. of ESANN 2006*, pages 563–569, 2006.

[4] N. W. Campbell, B. T. Thomas, and T. Troscianko. Neural networks for the segmentation of outdoor images. In *Solving Engineering Problems with Neural Networks. Proceedings of the International Conference on Engineering Applications of Neural Networks (EANN'96). Syst. Eng. Assoc, Turku, Finland*, volume 1, pages 343–6, 1996.

[5] L. P. Cordella, P. Foggia, C. Sansone, F. Tortorella, and M. Vento. Reliability parameters to improve combination strategies in multi-expert systems. *Pattern Analysis and Applications*, 2(3):205–214, 1999.

[6] National Reseach Council. *Chemical and Biological Terrorism: Research and Development to Improve Civilian Medical Response*. National Academy Press, 1999.

[7] C. de Stefano, C. Sansone, and M. Vento. 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(1):84–93, 2000.

[8] S.-O. Deininger, M. Gerhard, and F.-M. Schleif. Statistical classification and visualization of maldi-imaging data. In *Proc. of CBMS 2007*, pages 403–405, 2007.

[9] Google. Free available images of colorado springs, 2008. http://maps.google.de/maps [key word: colorado springs] (last visit 17032008).

[10] B. Hammer, M. Strickert, and Th. Villmann. Supervised neural gas with general similarity measure. *Neural Processing Letters*, 21(1):21–44, 2005.

[11] B. Hammer and Th. Villmann. Generalized relevance learning vector quantization. *Neural Networks*, 15(8-9):1059–1068, 2002.

[12] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, New York, 2001.

[13] W. Hermann, B. Eggers, H. Barthel, D. Clark, Th. Villmann, S. Hesse, F. Grahmann, H.-J. Kühn, O. Sabri, and A. Wagner. Correlation between automated writing movements and striatal dopaminergic innervation in patients with Wilson's disease. *Journal of Neurology*, 249(8):1082–1087, 2002.

[14] W. Hermann, A. Wagner, H.-J. Khn, F. Grahmann, and T. Villmann. Classification of fine-motoric disturbances in Wilson's disease using artificial neural networks. *Acta Neurologica Scandinavia*, 111(6):400–406, 2005.

[15] Teuvo Kohonen. *Self-Organizing Maps*, volume 30 of *Springer Series in Information Sciences*. Springer, Berlin, Heidelberg, 1995. (Second Extended Edition 1997).

[16] J. Lee and M. Verleysen. Generalization of the l_p norm for time series and its application to self-organizing maps. In M. Cottrell, editor, *Proc. of Workshop on Self-Organizing Maps (WSOM) 2005*, pages 733–740, Paris, Sorbonne, 2005.

[17] C. Manning and H. Schuetze. *Foundations of Statistical Natural Language Processign*. MIT Press, London, 1999.

[18] Thomas M. Martinetz, Stanislav G. Berkovich, and Klaus J. Schulten. 'Neural-gas' network for vector quantization and its application to time-series prediction. *IEEE Trans. on Neural Networks*, 4(4):558–569, 1993.

[19] J.O. Ramsay and B. W. Silverman. *Functional Data Analysis*. Springer, New York, 2006.

[20] J. A. Richards and X. Jia. *Remote Sensing Digital Image Analysis*. Springer, New York, 1999. 3rd Ed.

[21] A. Sato and K. Yamada. Generalized learning vector quantization. In D. S. Touretzky, M. C. Mozer, and M. E. Hasselmo, editors, *Advances in Neural Information Processing Systems 8. Proceedings of the 1995 Conference*, pages 423–9. MIT Press, Cambridge, MA, USA, 1996.

[22] F.-M. Schleif, U. Clauss, Th. Villmann, and B. Hammer. Supervised relevance neural gas and unified maximum separability analysis for classification of mass spectrometric data. In *Proceedings of ICMLA 2004*, pages 374–379. IEEE Press, December 2004.

[23] F.-M. Schleif, B. Hammer, and Th. Villmann. Supervised neural gas for functional data and its application to the analysis of clinical proteom spectra. In *Proc. of IWANN 2007*, pages 1036–1044, 2007.

[24] F.-M. Schleif, M. Lindemann, P. Maass, M. Diaz, J. Decker, T. Elssner, M. Kuhn, and H. Thiele. Support vector classification of proteomic profile spectra based on feature extraction with the bi-orthogonal discrete wavelet transform. *Computing and Visualization in Science*, pages DOI: 10.1007/s00791–008–0087–z, 2008.

[25] F.-M. Schleif, T. Villmann, T. Elssner, J. Decker, and M. Kostrzewa. Machine learning and soft-computing in bioinformatics - a short journey. In *Proc. of FLINS 2006*, pages 541–548, 2006.

[26] F.-M. Schleif, T. Villmann, B. Hammer, M. v. d. Werff, A. Deelder, and R. Tollenaar. Analysis of Spectral Data in Clinical Proteomics by use of Learning Vector Quantizers. In T. G. Smolinski, M. G. Milanova, and A.-E. Hassanien, editors, *Computational Intelligence in Biomedicine and Bioinformatics*, volume 1, pages 141–167. Springer, New York, NY, USA, 2008.

[27] F.-M. Schleif, T. Villmann, M. Kostrzewa, B. Hammer, and A. Gammerman. Cancer informatics by prototype networks in mass spectrometry. *Artificial Intelligence in Medicine*, page PMID:18778925, 2008.

[28] V Vapnik. *Statistical Learning Theory*. Wiley, New York, 1998.

[29] T. Villmann, E. Merenyi, and U. Seiffert. Machine learning approaches and pattern recognition for spectral data. In *Proc. Of European Symposium on Artificial Neural Networks (ESANN'2008)*, pages 433–444, Evere, Belgium, 2008. d-side publications.

[30] T. Villmann, F.-M. Schleif, B. Hammer, and M. Kostrzewa. Exploration of mass-spectrometric data in clinical proteomics using learning vector quantization methods. *Briefings in Bioinformatics*, 9(2):129–143, 2008.

[31] T. Villmann, M. Strickert, C. Brüß, F.-M. Schleif, and U. Seiffert. Visualization of fuzzy information in fuzzy-classification for image segmentation using MDS. In M. Verleysen, editor, *Proc. Of European Symposium on Artificial Neural Networks (ESANN'2007)*, pages 103–108, Brussels, Belgium, 2007. d-side publications.

[32] Th. Villmann, E. Merényi, and B. Hammer. Neural maps in remote sensing image analysis. *Neural Networks*, 16(3-4):389–403, 2003.

[33] V. Vovk, A. Gammerman, and G. Shafer. *Algorithmic Learning in a Random World*. Springer, New York, 2005.