ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

Rejection Strategies for Learning Vector Quantization

Lydia Fischer[1,2], Barbara Hammer[2] and Heiko Wersing[1] *

1 – HONDA Research Institute Europe GmbH,
Carl-Legien-Str. 30, 63065 Offenbach - Germany

2 – Bielefeld University, Universitätsstr. 25, 33615 Bielefeld - Germany

Abstract. We present prototype-based classification schemes, e. g. learning vector quantization, with cost-function-based and geometrically motivated reject options. We evaluate the reject schemes in experiments on artificial and benchmark data sets. We demonstrate that reject options improve the accuracy of the models in most cases, and that the performance of the proposed schemes is comparable to the optimal reject option of the Bayes classifier in cases where the latter is available.

1 Motivation

Powerful machine learning methods such as recent learning vector quantization (LVQ) models based on cost functions or support vector machines and linear time approximations thereof provide state of the art classification algorithms for automated data analysis [1, 2, 3, 4]. Their linear time complexity, high accuracy, and excellent generalization ability make them suitable also for large data sets. However, generalization bounds and training algorithms rely on the assumption of data being i.i.d. This limits the suitability for big data analysis, streaming data which displays a trend, in presence of outliers, or regions of strong overlap in the data. These cases require enhancing the classifier by measures of certainty that a model has taken a classification decision for a certain point or a data region. Such reject options constitute a first step towards incremental adaptation of the model complexity tailored to data regions with a high degree of uncertainty.

While there exist popular extensions of SVM to provide a confidence value of the classification [5, 6, 7] and first models have been proposed for distance-based k-nearest neighbor approaches [8], only few approaches address prototype-based classifiers [9, 10] thereby lacking a comparison to theoretically motivated alternatives such as explicit stochastic models. In this contribution, we are interested in efficient, online-computable reject options for LVQ classifiers and their behavior in comparison to mathematically well founded statistical models. For this purpose, we address the cost-function based models generalized LVQ (GLVQ) [11] and generalized matrix LVQ (GMLVQ) [1] as well as the probabilistic counterpart robust soft LVQ (RSLVQ) [2]. We propose simple geometric reject options for these models which can be computed efficiently in online scenarios, and we compare these reject options to more costly alternatives based on a probabilistic modeling and an optimum reject option for the Bayes classifier [12].

*BH gratefully acknowledges funding by the CITEC center of excellence. LF acknowledges funding by the CoR-Lab Research Institute for Cognition and Robotics and gratefully acknowledges the financial support from Honda Research Institute Europe.

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

2 Learning Vector Quantization

Suppose a data set $(\mathbf{x}_j, y_j) \in \mathbb{R}^n \times \{1, \ldots, C\}$ with data points $\mathbf{x}$ and labels y is given. A LVQ classifier is characterized by a set of prototypes $\mathbf{W} = \{\mathbf{w}_i \in \mathbb{R}^n\}_{i=1}^k$ equipped with class labels $c(\mathbf{w}_i) \in \{1, \ldots, C\}$. A given point $\mathbf{x}_j$ is classified according to the label of the closest prototype, the winner, as measured in the squared Euclidean distance $\|\mathbf{x} - \mathbf{w}\|^2$ or any other distance.

Given training data, GLVQ [11] optimizes the location of prototypes by means of a stochastic gradient descent on the cost function

$$E = \sum_j \Phi((d^+(\mathbf{x}_j) - d^-(\mathbf{x}_j))/(d^+(\mathbf{x}_j) + d^-(\mathbf{x}_j)))$$

where Φ is a monotonic increasing function, e. g. the logistic function. $d^\pm$ is the distance to the closest prototypes $\mathbf{w}^\pm$ of the correct/incorrect class. A generalization of GLVQ towards a general quadratic form $(\mathbf{x} - \mathbf{w}_i)^T \Lambda (\mathbf{x} - \mathbf{w}_i)$ with positive semi-definite matrix Λ has been proposed under the acronym GMLVQ [1]. This cost function strongly correlates to the classification error since a data point is classified correctly iff the nominator of the cost function is smaller than zero. Further, the nominator can be linked to the hypothesis margin of the classifier which influences its generalization ability [1]. Note that the value of the fraction ranges in the interval $(-1, 1)$ with -1 indicating a certain classification because d^+ is much smaller than d^-. Due to its excellent performance in practice [13], we will consider a reject option related to these costs.

RSLVQ [2] optimizes the data log likelihood of a probabilistic model:

$$E = \sum_j \log p(y_j|\mathbf{x}_j, \mathbf{W}) = \sum_j \log \left(p(\mathbf{x}_j, y_j|\mathbf{W})/p(\mathbf{x}_j|\mathbf{W})\right)$$

$p(\mathbf{x}_j|\mathbf{W}) = \sum_i p(\mathbf{w}_i) p(\mathbf{x}_j|\mathbf{w}_i)$ is a mixture of Gaussians with uniform prior probability $p(\mathbf{w}_i)$ and Gaussian probability $p(\mathbf{x}_j|\mathbf{w}_i)$ centered in $\mathbf{w}_i$ which is iosotropic with fixed variance or, more generally, uses a general covariance matrix. The probability $p(\mathbf{x}_j, y_j|\mathbf{W}) = \sum_i \delta_{c(\mathbf{x}_j)}^{c(\mathbf{w}_i)} p(\mathbf{w}_i) p(\mathbf{x}_j|\mathbf{w}_i)$ (δ_i^j is the Kronecker delta) restricts to mixture components with correct labeling. Relying on a probability model, RSLVQ provides an explicit confidence value $p(y|\mathbf{x}, \mathbf{W})$ for every pair $\mathbf{x}$ and y, paying the price of a higher computational complexity for learning.

3 Reject Options

A reject option relaxes the constraint on a classifier to provide a class label for every input. We will consider reject options which are based on certainty measures. Given a certainty measure $r : \mathbb{R}^n \to \mathbb{R}$ for the classification of a point $\mathbf{x}$ and a threshold $\theta \in \mathbb{R}$, a simple reject option is to reject $\mathbf{x}$ iff $r(\mathbf{x}) < \theta$. As mentioned in [14], uncertainty can have two different reasons: points being outliers, or points being located in ambiguous regions. As we will discuss, certainty measures take these two causes into account to different degrees. Further, certainty measures differ according to their scaling, allowing a uniform threshold θ iff $r(\mathbf{x})$ is normalized, and they differ according to their computational complexity and online computability, i. e. efficiency. We investigate the following reject options:

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

RelSim: The relative similarity is inspired by the cost function of GLVQ. Its suitability for a rejection measure has been mentioned in [11] already. It is $r_{\mathrm{RelSim}}(\mathbf{x}) = (d^- - d^+)/(d^- + d^+)$. r_{RelSim} is efficient, normalized to $(0, 1)$, and takes both, ambiguity and outlier rejection into account.

Dist: As certainty measure, we consider the disambiguity of the classification as measured by the distance of a point to the closest decision boundary of the classifier. The distance of a point $\mathbf{x}$ to the hyperplane separating the receptive fields of $\mathbf{w}^+$ and $\mathbf{w}^-$ is given by $r_{\mathrm{Dist}}(\mathbf{x}) = (|d^+ - d^-|) \,/\, \left(2\|\mathbf{w}^+ - \mathbf{w}^-\|^2\right)$. This formula is directly applicable if every class is modeled by only one prototype. Otherwise, the underlying topology has to be estimated using e. g. Hebbian learning [15]. This certainty measure is efficient but not normalized.

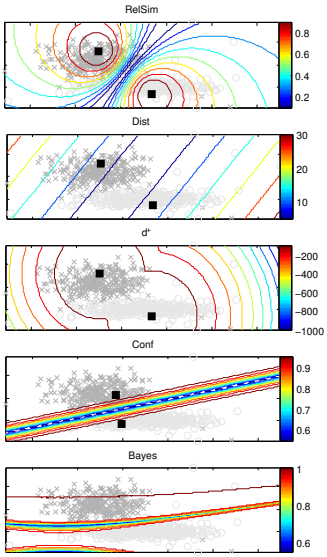

Fig. 1: Isobars of the measures for an artificial two class problem with Gaussian clusters. Black squares are GLVQ/RSLVQ prototypes.

$\mathbf{d}^+$: Outliers can be identified by their distance to the closest prototype d^+. We use this information for an outlier-based certainty measure as basis for a reject option: $r_{d^+}(\mathbf{x}) = -d^+(\mathbf{x})$. This measure is efficient but not normalized.

Comb: This measure combines the previous two reject options $r_{\mathrm{Comb}}(\mathbf{x}) = (r_{\mathrm{Dist}}(\mathbf{x}), r_{d^+}(\mathbf{x}))$ leading to a reject strategy based on a threshold vector $\theta = (\theta_1, \theta_2)$: $\mathbf{x}$ is rejected if $r_{\mathrm{Dist}}(\mathbf{x}) < \theta_1$ or $r_{d^+}(\mathbf{x}) < \theta_2$. The measure takes into account ambiguity and outliers, but it requires two threshold parameters. For evaluation, we refer to the best combination of both thresholds determined via exhaustive search, which is no longer efficient but can serve as a baseline for comparison.

Conf: Classifiers based on probabilistic models such as RSLVQ provide a direct confidence value of the classification: $r_{\mathrm{Conf}}(\mathbf{x}) = \max_y \hat{p}(y|\mathbf{x})$ with the estimated probability $\hat{p}(\cdot)$. This measure is normalized and, depending on the probability model, it takes into account ambiguous regions. The drawback is that it can only be used for probabilistic models such as RSLVQ which has higher complexity as compared to GLVQ or GMLVQ. Conf serves as baseline for an evaluation whether simple geometric measures can reach the quality of probabilistic models.

Bayes: The Bayes classifier provides class probabilities for each class provided the data distribution is known. The corresponding reject option $r_{\mathrm{Bayes}}(\mathbf{x}) = \max_y p(y|\mathbf{x})$ is optimal in the sense of an error-reject trade-off [12]. We will use it as ground truth for an artificial data set with known underlying distribution.

4 Experiments

We evaluate the results of the reject strategies in a 10-fold repeated cross-validation with ten repeats for RSLVQ, GLVQ, and GMLVQ with one prototype

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

per class. The following data sets are used:

- *Gaussian clusters*: This data set contains two artificially generated overlapping 2D Gaussian clusters. These are overlaid with uniform noise.
- *Image Segmentation*: The image segmentation data set consists of 2310 data points representing small patches from outdoor images with 7 different classes with equal distribution such as brickface, sky, ... [16]. Each data point consists of 19 real-valued image descriptors.
- *COIL-20*: The Columbia Object Image Database Library (COIL-20) consists of gray scaled images of twenty objects [17]. The objects are rotated in 5° steps, so that there are 72 images per object. The data set contains 1440 data points which are 16384 dimensional. We use PCA [18] to reduce the dimensionality to 30. The task is to classify each single object.
- *Tecator data*: The tecator data set consists of 215 spectra with 100 spectral bands ranging from 850 nm to 1050 nm [19]. The task is to predict the fat content of the probes, which is turned into a two class classification problem to predict a high/low fat content by binning into two classes of equal size.
- *Haberman*: The Haberman survival data set contains 306 instances from two classes indicating the survival for more than 5 years after breast cancer surgery [16]. Data are represented by three attributes related to the age, the year, and the number of positive axillary nodes detected.

We report the effect of the different reject strategies for the different models RSLVQ, GLVQ, and GMLVQ where applicable. Thereby, we vary the reject threshold θ in small steps from no reject (which corresponds to the original model) to full reject (i. e. no data point is classified). $\mathbf{X}_\theta$ denotes the points which are not rejected using θ. The results are reported as graphs of the relative size $\mathbf{X}_\theta/\mathbf{X}$ versus the classification accuracy on $\mathbf{X}_\theta$ normalized by its size.

Figure 2 displays all results. For RSLVQ for Gaussian clusters, Conf, Dist and Comb provide nearly the same shape as the optimal reject option of the Bayes classifier. The same behavior occurs for RelSim, Dist and Comb with the G(M)LVQ model. Only d^+ shows poor results not only for Gaussian clusters but also for the other data sets. This can be an indicator that there are less outliers than ambiguous data points in the data. In nearly all models and for all data sets, Comb constitutes the best reject measure, but being based on two thresholds it is not efficient. RelSim and Dist provide comparable curves with less effort except for Tecator (GLVQ) and Haberman (RSLVQ, GLVQ). As a conclusion, we can see that RelSim in combination with GMLVQ offers an efficient certainty measure with a quality comparable to optimum reject strategies in almost all settings, but releasing the burden of an explicit probabilistic modeling for Conf or optimization of different objectives for Comb.

5 Conclusion

We have compared several efficient geometric reject measures for prototype-based approaches with statistical reject strategies on benchmark data sets. We

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

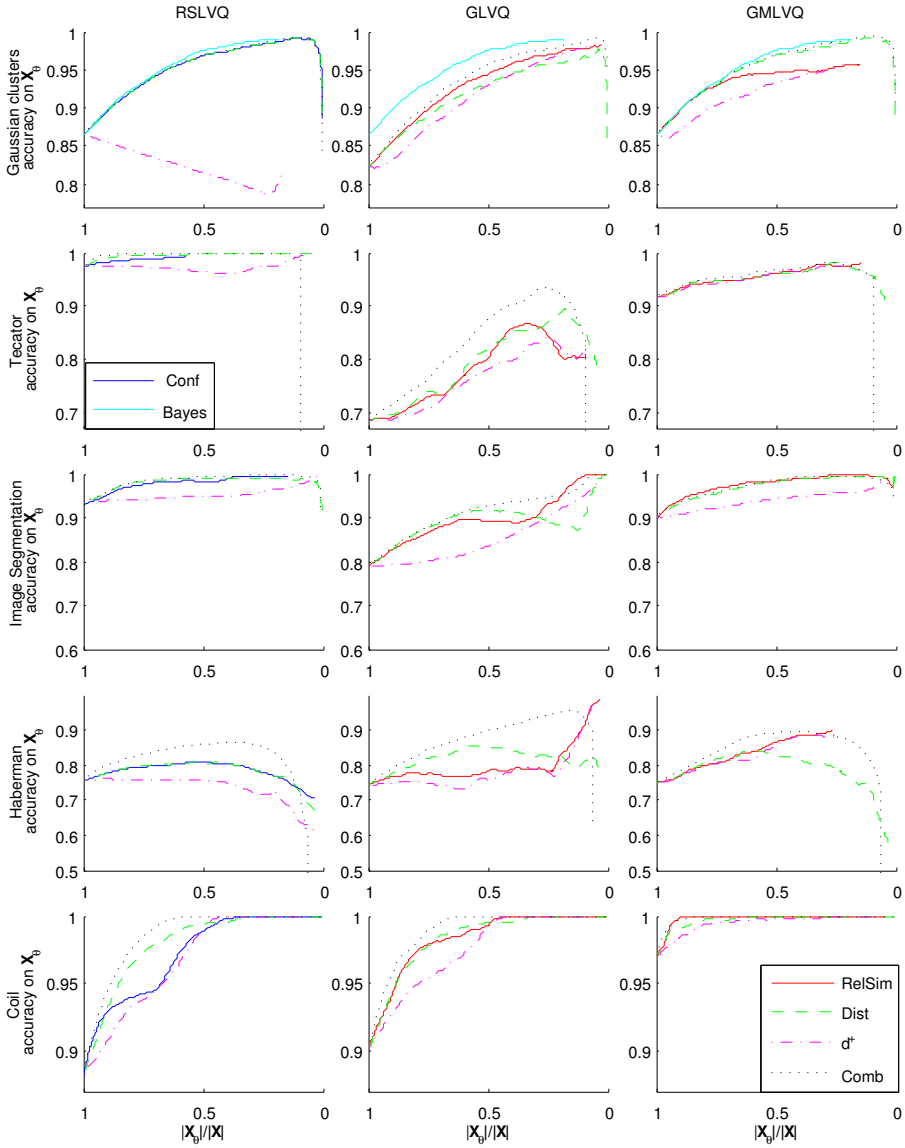

Fig. 2: Average results of mentioned rejection options when applying RSLVQ, GLVQ and GMLVQ models trained for different data sets. We display the relative size of $\mathbf{X}_\theta$ vs. the accuracy of the classifier on this set. The averaged curve is plotted, where at least 80 % of the single runs deliver a value.

applied the reject options to different models: GLVQ as popular LVQ scheme based on a cost function, GMLVQ which uses a metric adaption, and RSLVQ

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

which provides a statistically motivated discriminative model. We showed that geometrically motivated measures (RelSim, Dist, Comb) can be used to improve the accuracy of a model and they lead to results comparable to optimum Bayes reject strategies (e. g. using GMLVQ and RelSim) but releasing the burden of explicit statistical modeling. This opens the way towards the design of efficient life-long model adaptation for popular prototype-based classifiers such as GMLVQ: the model complexity can easily be tailored online towards regions with a low certainty of the classification, e. g. introducing novel prototypes which are capable of representing novel aspects of the data.

References

[1] P. Schneider, M. Biehl, and B. Hammer. Adaptive Relevance Matrices in Learning Vector Quantization. *Neural Computation*, 21(12):3532–3561, 2009.

[2] S. Seo and K. Obermayer. Soft Learning Lector Quantization. *Neural Computation*, 15(7):1589–1604, Jul 2003.

[3] C. Campbell and Y. Ying. *Learning with Support Vector Machines*. Morgan and Claypool, 2011.

[4] I. W. Tsang, J. T. Kwok, and P.-M. Cheung. Core vector machines: Fast SVM training on very large data sets. *Journal of Machine Learning Research*, 6:363–392, 2005.

[5] J. C. Platt. Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods. In *Adv. in Large Margin Classifiers*. MIT Press, 1999.

[6] T.-F. Wu, C.-J. Lin, and R. C. Weng. Probability Estimates for Multi-class Classification by Pairwise Coupling. *Journal of Machine Learning Research*, 5:975–1005, August 2004.

[7] G. Fumera and F. Roli. Support Vector Machines with Embedded Reject Option. In *Proceedings of the Int. Workshop on Pattern Recognition with Support Vector Machines (SVM2002), Niagara Falls*, pages 68–82. Springer, 2002.

[8] R. Hu, S. J. Delany, and B. Mac Namee. Sampling with Confidence: Using k-NN Confidence Measures in Active Learning. In *Proceedings of the UKDS Workshop at 8th International Conference on Case-based Reasoning*, ICCBR'09, pages 181–192, 2009.

[9] E. Ishidera, D. Nishiwaki, and A. Sato. A confidence value estimation method for handwritten Kanji character recognition and its application to candidate reduction. *International Journal on Document Analysis and Recognition*, 6(4):263–270, April 2004.

[10] S. Kirstein, H. Wersing, H.-M. Gross, and E. Körner. A Life-Long Learning Vector Quantization Approach for Interactive Learning of Multiple Categories. *Neural Networks*, 28:90–105, 2012.

[11] A. Sato and K. Yamada. Generalized Learning Vector Quantization. In *Advances in Neural Information Processing Systems*, volume 7, pages 423–429, 1995.

[12] C. K. Chow. On Optimum Recognition Error and Reject Tradeoff. In *IEEE Transactions in Information Theory*, volume 16(1), pages 41–16, 1970.

[13] M. Biehl, K. Bunte, and P. Schneider. Analysis of flow cytometry data by matrix relevance learning vector quantization. *PLoS ONE*, 8(3):e59401, 2013.

[14] A. Vailaya and A. K. Jain. Reject Option for VQ-Based Bayesian Classification. In *International Conference on Pattern Recognition (ICPR)*, pages 2048–2051, 2000.

[15] T. Martinetz and K. Schulten. Topology representing networks. *Neural Networks*, 7:507–522, 1994.

[16] K. Bache and M. Lichman. UCI machine learning repository, 2013.

[17] S. A. Nene, S. K. Nayar, and H. Murase. Columbia Object Image Library (COIL-20). *Technical Report CUCS-005-96*, February 1996.

[18] L. J. P. van der Maaten. Matlab Toolbox for Dimensionality Reduction, March 2013. http://homepage.tudelft.nl/19j49/Matlab_Toolbox_for_Dimensionality_Reduction.html.

[19] H. H. Thodberg. Tecator data set, contained in StatLib Datasets Archive, 1995.