Honda Research Institute EU
Honda Research Institute Europe GmbH
https://www.honda-ri.de/

Rejection strategies for Learning Vector Quantization - a comparison of probabilistic and deterministic approaches

Heiko Wersing, David Nebel, Barbara Hammer, Thomas Villmann, Lydia Fischer

2014

Preprint:

This is an accepted article published in Advances in Self-Organizing Maps and Learning Vector Quantization - Proceedings of the 10th International Workshop, WSOM. The final authenticated version is available online at: https://doi.org/10.1007/978-3-319-07695-9_10

Rejection Strategies for Learning Vector Quantization – a Comparison of Probabilistic and Deterministic Approaches

Lydia Fischer[1,2], David Nebel[3], Thomas Villmann[3],
Barbara Hammer[2], and Heiko Wersing[1]

1 – HONDA Research Institute Europe GmbH, Offenbach, Germany
2 – Bielefeld University, Germany
3 – University of Applied Sciences Mittweida, Germany

Abstract. In this contribution, we focus on reject options for prototype-based classifiers, and we present a comparison of reject options based on statistical models for prototype-based classification as compared to alternatives which are motivated by simple geometric principles. We compare the behavior of generative models such as Gaussian mixture models and discriminative ones to results from robust soft learning vector quantization. It turns out that (i) reject options based on simple geometric show a comparable quality as compared to reject options based on statistical approaches. This behavior of the simple options offers a nice alternative towards making a probabilistic modeling and allowing a more fine-grained control of the size of the remaining data in many settings. It is shown that (ii) discriminative models provide a better classification accuracy also when combined with reject strategies based on probabilistic models as compared to generative ones.

Keywords: prototype-based reject option, classification

1 Introduction

Learning vector quantization (LVQ) [15] constitutes a powerful and efficient classification strategy particularly suited for multi-class classification or online scenarios. It can be substantiated by strong mathematical guarantees for generalization behavior as well as learning dynamics for modern cost function based versions such as generalized LVQ (GLVQ) [21] or robust soft LVQ (RSLVQ) [23]. In application scenarios, however, perfect classification can rarely be achieved due to inherent noise in the data, overlap of classes, missing sensors, etc. Essentially, a reject option relaxes the constraint of a classifier to provide a class label for a given input with a low confidence value, rather an explicit 'don't know' is accepted as a return in such cases.

Note that most classifiers actually do provide a continuous value rather than a crisp output only such as the distance of a given data point to the decision boundary. Together with an appropriate threshold, these numbers could be taken

as a reject option. However, the real-valued outputs provided by the classifiers can usually not be interpreted as a confidence measure because their scaling is unclear and can vary locally. A variety of approaches is concerned with techniques how to turn these values into a statistical confidence [20, 27], or how to define appropriate, possibly local thresholds for a reject option which respects a different scaling of the values [9, 25]. Interestingly, while a number of efficient strategies have been realized for popular classification schemes like support vector machines or k-nearest neighbor classifiers [4, 20, 27, 7, 12, 9, 6], relatively little approaches address prototype-based learning strategies such as LVQ [25, 5, 13]. Another idea is the distance-based two stage approach from [16] which separately addresses outliers and ambiguous regions. An approach, which combines a reject option with empirical risk minimization for a binary classifier, is proposed in [11] which could be a direction of further research.

In this approach we investigate reject options for prototype-based learning schemes such as LVQ. In particular, we investigate approaches which are inspired by the geometric nature of LVQ classifiers and we compare these reject options to reject options based on confidence values. We consider the key question: Are these geometric approaches comparable to reject strategies based on confidence values of probabilistic models which can be optimal as shown in [4], and if so under which conditions? Therefore, we systematically compare the behavior of the measures to rejection strategies for probabilistic models. We vary (i) the rejection strategy, ranging from deterministic, geometric measures to reject options based on confidence values, (ii) the data set, ranging from artificial data to typical benchmarks, and (iii) the nature of the prototype-based model for which the reject option is taken, considering purely discriminative models in comparison to generative ones. Albeit both classifiers are derived as explicit probabilistic models. Purely discriminative ones are tailored to the classification task rather than the data, such that it is not clear whether reject strategies can be based on their confidence values. Similarly, it is not clear whether efficient deterministic strategies based on simple geometric quantities can reach the performance of rejection strategies on confidence values, the latter is supposed to require valid probabilistic models of the data. We will show that this is indeed the case for real life settings: heuristic reject strategies based on geometric considerations offer an alternative to measures based on a confidence value, thus offering a way towards reject strategies for purely deterministic LVQ schemes.

2 Probabilistic Prototype-Based Classification

Assume a data set $\boldsymbol{X}$ with elements of the real vector space $\mathbb{R}^n$. A prototype-based classifier is characterized by a set of prototypes $\boldsymbol{W} = \{\boldsymbol{w}_i \in \mathbb{R}^n\}_{i=1}^k$, which are equipped with labels $c(\boldsymbol{w}_i) \in \{1, \ldots, C\}$, if a classification into C classes is considered. Classification of a data point $\boldsymbol{x} \in \mathbb{R}^n$ takes place by a winner takes all (WTA) scheme: $\boldsymbol{x}$ is mapped to the label $c(\boldsymbol{x}) = c(\boldsymbol{w}_i)$ of the prototype $\boldsymbol{w}_i$ which is closest to $\boldsymbol{x}$ as measured in some distance measure. Often, the standard squared euclidean distance $\|\boldsymbol{x} - \boldsymbol{w}_i\|^2$ or a generalized quadratic form

$(\boldsymbol{x} - \boldsymbol{w}_i)^T \Lambda (\boldsymbol{x} - \boldsymbol{w}_i)$ with positive semi-definite matrix Λ is considered; generalizations to more general dissimilarity measures such as divergences, functional metrics, or general dissimilarities have also been proposed [26, 10].

Due to its simple classification scheme and the representation of the model in terms of few prototypes, prototype-based classification enjoys a wide popularity. Additional there are diverse learning techniques available to induce an appropriate model from a given data set. Popular learning techniques include the classical family of LVQ as proposed by Kohonen [15], generalizations of LVQ which establish the model by cost functions [21, 23], or unsupervised learning schemes equipped with posterior labeling like neural gas or extensions thereof [17, 2]. Here, we have a glimpse at two different strategies which play a role in the subsequent experiments. We only consider probabilistic LVQ models, because the results allow a direct use of a reject option on their confidence values.

RSLVQ: Robust soft learning vector quantization (RSLVQ) has been proposed as a probabilistic model which, in the limit of small bandwidth, yields update rules very similar to classical LVQ 2.1 [23]. The objective is given as

$$E = \sum_j \log p(y_j|\boldsymbol{x}_j, \boldsymbol{W}) = \sum_j \log \frac{p(\boldsymbol{x}_j, y_j|\boldsymbol{W})}{p(\boldsymbol{x}_j|\boldsymbol{W})} \tag{1}$$

where $p(\boldsymbol{x}_j|\boldsymbol{W}) = \sum_i p(\boldsymbol{w}_i) p(\boldsymbol{x}_j|\boldsymbol{w}_i)$ constitutes a mixture of Gaussians with prior probability $p(\boldsymbol{w}_i)$ usually taken uniformly over all prototypes. The probability $p(\boldsymbol{x}_j|\boldsymbol{w}_i)$ is usually taken as an isotropic Gaussian centered in $\boldsymbol{w}_i$ with fixed variance σ^2, or a generalization thereof with a more general covariance matrix. The probability $p(\boldsymbol{x}_j, y_j|\boldsymbol{W}) = \sum_i \delta_{c(\boldsymbol{x}_j)}^{c(\boldsymbol{w}_i)} p(\boldsymbol{w}_i) p(\boldsymbol{x}_j|\boldsymbol{w}_i)$ (δ_i^j being the Kronecker delta) restricts to the mixture components with the correct labeling. This likelihood ratio is optimized using a gradient technique. RSLVQ provides an explicit confidence value $p(y|\boldsymbol{x}, \boldsymbol{W})$ for every class y of a given data point $\boldsymbol{x}$.

GMM: Albeit RSLVQ is derived from a probabilistic model, its cost function is purely discriminative. This means model parameters do not necessarily yield to a good generative model for the observed data $\boldsymbol{x}$. As shown in [22], for example, this is not the case in general. In practice, generative data models are often trained in an unsupervised way, directly aiming at a representation of the data distribution $p(\boldsymbol{x})$, popular examples being Gaussian mixture models for density estimation. Here we consider a class-wise Gaussian mixture model (GMM) which aims at a representation of every class by optimizing the following data log-likelihood

$$E = \sum_j \log \left(\sum_i \delta_{c(\boldsymbol{x}_j)}^{c(\boldsymbol{w}_i)} p(\boldsymbol{w}_i) p(\boldsymbol{x}_j|\boldsymbol{w}_i) \right) \tag{2}$$

where $p(\boldsymbol{x}_j|\boldsymbol{w}_i)$ is a Gaussian distribution centered in $\boldsymbol{w}_i$, and $p(\boldsymbol{w}_i)$ is the class-wise prior of the prototype with $\sum_j \delta_{c(\boldsymbol{x}_j)}^{c(\boldsymbol{w}_j)} p(\boldsymbol{w}_j) = 1$. The model parameters can be optimized by means of a gradient technique or, alternatively, a classical EM scheme for every class, since the objective decomposes according to the class labels [3]. A GMM provides for each class y an explicit confidence measure

$p(y|\boldsymbol{x}, \boldsymbol{W}) = p(y)p(\boldsymbol{x}, c(\boldsymbol{x})|\boldsymbol{W}) / \sum_{z\in\{1,\ldots,C\}} p(z)p(\boldsymbol{x}, z|\boldsymbol{W})$ where, due to the training procedure, a generative data model representing the distribution on $\boldsymbol{x}$ is present. In this context $p(y)$ is the prior of the class with $\sum_{y\in\{1,\ldots,C\}} p(y) = 1$.

Since GMM and RSLVQ offer probabilistic models, the classification of a data point $\boldsymbol{x}$ can be based on the most likely class $\operatorname{argmax}_y p(y|\boldsymbol{x}, \boldsymbol{W})$. In practice, the resulting maximum y often corresponds to the class of the closest prototype such that a close resemblance to a classical WTA scheme is obtained.

3 Reject Options

What are possible rejection measures of prototype-based models which correlate to the confidence of a classification and, together with a rejection strategy such as a simple threshold, lead to a reject option? In general, a rejection measure constitutes a function $r : \mathbb{R}^n \to \mathbb{R}$, $r(\boldsymbol{x})$ indicating the certainty of the classification of a data point $\boldsymbol{x}$, together with an ordering direction, which specifies whether low or high values of $r(\boldsymbol{x})$ correspond to a high certainty of the classification. We assume that a rejection measure is always scaled in such a way that smaller values correspond to a lower certainty. We consider the following rejection measures:

Conf: Chow proved for a Bayes classifier with known class densities that a reject option on $r_{\text{Conf}}(\boldsymbol{x}) = \max_y p(y|\boldsymbol{x})$ reaches the optimum error-reject tradeoff: for a certain error rate (error probability) it minimizes the reject rate (reject probability)[4]. This means to reject a data point if $r_{\text{Conf}}(\boldsymbol{x}) < \theta$. This strategy relies on the assumption that a good probabilistic model of the data is given, otherwise guarantees as proved e. g. in [11] do not necessarily hold. Note that in regions with low class densities this measure can return high confidence values caused by normalization, thus it cannot exclude outliers. Our measure (Fig. 1) is inspired by the one of Bayes but the values are calculated by the mentioned models and not by a Bayes classifier.

Dist: This error measure is inspired by geometric considerations. It returns the distance of $\boldsymbol{x}$ to the closest decision boundary. Assume $\boldsymbol{w}^+$ and $\boldsymbol{w}^-$ correspond to prototypes with a different labeling and neighbored receptive fields with the belonging distances d^+ and d^- to $\boldsymbol{x}$. Then, the distance of a data point $\boldsymbol{x}$ to the decision boundary defined by these two prototypes is given as $r_{\text{Dist}}(\boldsymbol{x}) = \frac{|d^+ - d^-|}{2\|\boldsymbol{w}^+ - \boldsymbol{w}^-\|^2}$ (Fig. 1). If only one prototype per class is present, the prototypes $\boldsymbol{w}^+$ and $\boldsymbol{w}^-$ are given by the two closest prototypes of the data point $\boldsymbol{x}$. Provided a class is represented by more prototypes than one, the underlying topology has to be estimated using e.g. the Hebbian learning strategy as proposed in [18].

$\mathbf{d}^+$: This error measure is also geometrically inspired, treating points which are outliers with low confidence. This is measured by the squared distance to the closest prototype $r_{d^+}(\boldsymbol{x}) = -d^+(\boldsymbol{x})$ (Fig. 1).

Note that these reject options differ in the following items:

- *Motivation of r:* There are essentially two different reasons to reject a data point, which are referred to in the literature as a rejection because of an

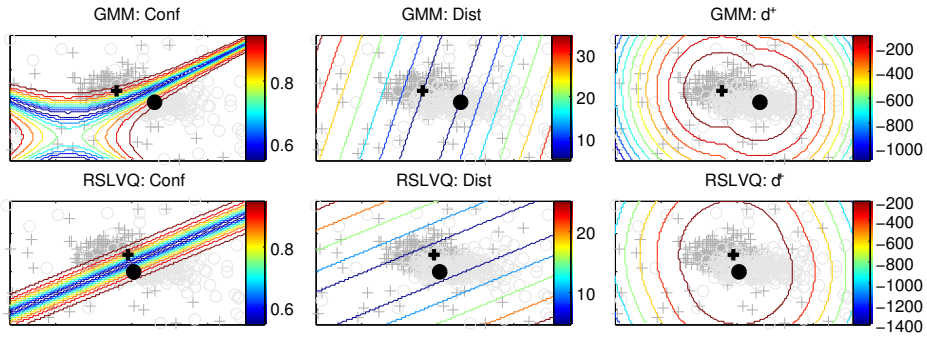


Fig. 1. Level curves of the considered reject options for a GMM and a RSLVQ model of an artificial 2D Gaussian data set. The black symbols are prototypes.

ambiguous classification, or a rejection because of the data point being an outlier [25]. The reject measures as given above follow different principles. **Conf** realizes a rejection because of ambiguity, since it requires that the maximum class probability reaches the threshold θ. Due to the normalization of probabilities, this results in a gap of the class probabilities. **Dist** explicitly realizes an ambiguous reject option by referring to the class boundary, while $\mathbf{d}^+$ realizes an outlier reject option.

- *Scaling of r:* For **Conf**, values are in the interval $[0, 1]$ allowing a direct interpretation as statistical confidence value. This fact offers a simple way to set an appropriate threshold due to external requirements regarding the confidence, for example. In contrast, the other measures take values in the real numbers, but their scaling is not clear. Since the scaling can even vary locally and it can depend nonlinearly on the confidence, a proper choice of a threshold is unclear. We will investigate global threshold strategies in experiments, yielding results comparable to reject options based on the confidence.
- *Requirements as regards the model:* The scaling of **Conf** as a confidence measures requires that a probabilistic model of the data is available. We investigate the effect of having a discriminative versus generative model in experiments, only the latter actually providing a valid representation of the input distribution in general.

These measures provide values indicating the confidence of a classification such that they give rise to a direct threshold-based rejection strategy: given $\theta \in \mathbb{R}$, points which fulfill $r(\boldsymbol{x}) < \theta$ are rejected. Since measures such as **Dist** and $\mathbf{d}^+$ aim at a rejection caused by different reasons. It can be worthwhile to combine several measures [25]. This leads to a more complex rejection strategy which depends on two thresholds. We refer to this measure as follows:

Comb: This measure combines the previous two reject options $r_{\text{Comb}}(\boldsymbol{x}) = (r_{\text{Dist}}(\boldsymbol{x}), r_{d^+}(\boldsymbol{x}))$ leading to a reject strategy based on a threshold vector $\boldsymbol{\theta} = (\theta_1, \theta_2)$: $\boldsymbol{x}$ is rejected if $r_{\text{Dist}}(\boldsymbol{x}) < \theta_1$ or $r_{d^+}(\boldsymbol{x}) < \theta_2$.

4 Experiments

We test the behavior of the different rejection measures in experiments, focusing on the following questions: What is the behavior of the measures regarding different characteristics of the model ranging from a discriminative to a generative one? What is the behavior of simple deterministic heuristics in comparison to rejection strategies based on confidence measures and do the latter require valid probabilistic models? Since probabilistic models are needed for an evaluation of **Conf**, we use the two probabilistic models RSLVQ and GMM. For all settings, RSLVQ and GMM are trained using one prototype per class. For RSLVQ, a global parameter σ^2 is optimized via cross-validation. For GMM, correlations are set to zero and local scalings of the dimensions are adapted by means of diagonal matrices attached to the prototypes which are optimized in an EM scheme. Training takes place until convergence using random initialization and without leave-one-out method. Convergence is assumed if the training error changes less than 10^{-5} during two sequenced training steps. We use the following data sets:

- *Gaussian clusters*: This data set consists of two artificially generated Gaussian clusters in two dimensions with overlap. These are overlaid with uniform noise in the plane. Data are randomly divided into training and test set.
- *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, ... [1]. Each data point consists of 19 real-valued image descriptors. The data set is decomposed into a training set of 210 data points and a test set of 2100 data points. Due to zero variance, dimensions 3 to 5 are deleted, and data are normalized by a z-transformation before training.
- *Tecator data*: The Tecator data set consists of 215 spectra with 100 spectral bands ranging from 850 nm to 1050 nm [24]. 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 means of binning the real values into two classes of equal size. Data are randomly split into a training set with 144 samples and test set with 71 samples.
- *Haberman*: The Haberman survival data set contains 306 instances from two classes indicating the survival for more than 5 years after breast cancer surgery [1]. Data are represented by three attributes related to the age, the year, and the number of positive axillary nodes detected. Data are randomly split into training and test set of equal size.

For all data sets, two models are trained: a probabilistic generative model by means of class-wise GMM, and a probabilistic discriminative model by means of RSLVQ. For the resulting models, the effect of a reject option is compared for different possible strategies as introduced above. 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). For **Comb**, a threshold vector is varied accordingly, and we report the result of the respective best combination. We denote the set of data points which are not rejected using θ as $\boldsymbol{X}_\theta$. The results

are depicted as graphs plotting the relative size $|\boldsymbol{X}_\theta|/|\boldsymbol{X}|$ versus the classification accuracy on $\boldsymbol{X}_\theta$ normalized by its size.

Figure 2 shows the results obtained for the different rejection strategies and data sets. The resulting graphs [19] display a smooth transition from the accuracy of the model without reject options to the limit value 1 (in the case of Gaussian clusters it goes to 0) which results if $|\boldsymbol{X}_\theta|$ approaches 0 (we leave out the value for the empty set at $|\boldsymbol{X}_\theta| = 0$). The classification accuracy on $\boldsymbol{X}_\theta$ does not change with θ if the classification accuracy is already 100 % (as is the case for the Tecator data set for RSLVQ), or if the errors are uniformly distributed over the range of the rejection measure r which is the case for the Haberman data set, for example. In the latter case, classes are imbalanced with the second class accounting for roughly one third of the data only, and LVQ models tend to represent only class one properly, such that class two accounts for errors equally distributed according to r. Note that the graphs are subject of noise if the size $|\boldsymbol{X}_\theta|$ approaches 0 which can be attributed to the small sample size $\boldsymbol{X}_\theta$. Accordingly, the graphs are not reliable for $|\boldsymbol{X}_\theta|/|\boldsymbol{X}| < 0.1$, and the corresponding parts of the graphs should be seen as an indicator only. We choose the values of θ equidistant between the extremal values of each single measure.

Interestingly, the control of the number of points which are not rejected, $|\boldsymbol{X}_\theta|$, depending on the threshold θ partially has gaps, as indicated in Fig. 2 by the straight parts of the curves and the ending of the curves at some size of $|\boldsymbol{X}_\theta| \gg 0$. Such gaps can occur provided the size of $\boldsymbol{X}_\theta$ changes abruptly with the threshold, which seems to be the case in some settings where a further increase of the thresholds leads to a rejection of all remaining data points. This is the fact for **Conf** for Gaussian clusters, Image Segmentation and Tecator for the GMM model, indicating that no points with confidence larger than a maximum threshold value θ exist. Interestingly these gaps can be observed for **Conf** for the generative models only, not the discriminative ones. Further, this behavior is observed for $\mathbf{d}^+$ for the data sets Gaussian clusters and Image Segmentation (both models) and Tecator (generative model). In contrast, the graphs of **Dist** and **Comb** do not have large gaps.

We can draw a few general conclusions from the graphs displayed in Fig. 2: In all cases, the discriminative model RSLVQ yields the same or better results as compared to generative GMM models, albeit the latter have a higher degree of freedom because of an adaptive diagonal matrix per prototype unlike RSLVQ, which relies on a global bandwidth only. This also holds for the full range of certainty values taken for the reject strategies, regardless of whether deterministic of probabilistic rejection measures are used. Thus, it seems advisable to focus on the discriminative task, where confidence based measure or deterministic measures can be used. As expected, reject strategies based on the confidence yields the best behavior in most cases, but it does not allow a smooth variation of the size of $\boldsymbol{X}_\theta$ for a large range in two of the settings. As mentioned in Section 3 **Conf** cannot exclude outliers. This is apparently not a problem for the used data sets, highlighting the applicability of the optimality criterion of Chow [4]. **Dist** seems to offer a reasonable strategy in all other settings, whereby the be-

havior is universally good for generative as well as discriminative models, and it relaxes the burden of computing an explicit confidence value. $\mathbf{d}^+$ gives better results than **Dist** in only one case (Gaussian clusters, GMM), and worse results than **Dist** in three cases (Gaussian cluster, RSLVQ; Image segmentation, both models; Tecator, GMM). Thus, in general, focusing on the discriminative nature seems advisable also as concerns the rejection strategy. As expected, **Comb** shows results comparable to the best of the two geometric reject options **Dist** and $\mathbf{d}^+$, but also requiring a more complex reject strategy by the combination of both values.

5 Conclusions

We have compared direct geometric reject options and their combination with Bayesian motivated reject options in a couple of benchmarks using models with different characteristics. The resulting observations are that geometric measures such as **Dist** behave equally good as probabilistic measures, while often allowing a more fine-grained control of the size of the rejected data set. In addition, they do not require explicit probabilistic models thus opening the way for an integration into powerful deterministic alternatives such as GLVQ [21]. The suitability of the approach to these settings is the topic of ongoing work [8].

While allowing for simple measures which are applicable for a wider range of models, the scaling of appropriate thresholds is not clear a priori and it depends on the data set at hand. In the literature, a few proposals how to automatically determine data-adapted values have been proposed [25], which can be transferred to our setting. They can even be extended to online scenarios, and LVQ classifiers offer intuitive life-long learning strategies [14].

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

References

1. K. Bache and M. Lichman. UCI machine learning repository, 2013.
2. O. Beyer and P. Cimiano. Online Semi-Supervised Growing Neural Gas. *Int. J. Neural Syst.*, 22(5), 2012.
3. C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.
4. C. K. Chow. On Optimum Recognition Error and Reject Tradeoff. *IEEE Transactions on Information Theory*, 16(1):41–46, 1970.
5. 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):84–94, 2000.
6. S. J. Delany, P. Cunningham, D. Doyle, and A. Zamolotskikh. Generating Estimates of Classification Confidence for a Case-Based Spam Filter. In *ICCBR*, pages 177–190, 2005.

7. P. R. Devarakota, B. Mirbach, and B. Ottersten. Confidence Estimation in Classification Decision: A Method for Detecting Unseen Patterns. In *Int. Conf. on Advances in Pattern Recognition (ICAPR 2007)*, 2006.
8. L. Fischer, B. Hammer, and H. Wersing. Rejection Strategies for Learning Vector Quantization, submitted.
9. G. Fumera, F. Roli, and G. Giacinto. Reject option with multiple thresholds. *Pattern Recognition*, 33(12):2099–2101, Dec. 2000.
10. B. Hammer, D. Hofmann, F.-M. Schleif, and X. Zhu. Learning vector quantization for (dis-)similarities. *Neurocomputing*, accepted.
11. R. Herbei and M. H. Wegkamp. Classification with reject option. *Can J Statistics*, 34(4):709–721, 2006.
12. R. Hu, S. J. Delany, and B. M. Namee. Sampling with Confidence: Using k-NN Confidence Measures in Active Learning. In *Proceedings of the UKDS Workshop at 8th Int. Conf. on Case-based Reasoning*, ICCBR'09, pages 181–192, 2009.
13. E. Ishidera, D. Nishiwaki, and A. Sato. A confidence value estimation method for handwritten Kanji character recognition and its application to candidate reduction. *Int. J. on Document Analysis and Recognition*, 6(4):263–270, Apr. 2004.
14. 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.
15. T. Kohonen. *Self-Organization and Associative Memory*. Springer Series in Information Sciences, Springer-Verlag, third edition, 1989.
16. T. Landgrebe, D. M. J. Tax, P. Paclík, and R. P. W. Duin. The interaction between classification and reject performance for distance-based reject-option classifiers. *Pattern Recognition Letters*, 27(8):908–917, 2006.
17. T. Martinetz, S. Berkovich, and K. Schulten. Neural-gas Network for Vector Quantization and its Application to Time-Series Prediction. *IEEE-Transactions on Neural Networks*, 4(4):558–569, 1993.
18. T. Martinetz and K. Schulten. Topology representing networks. *Neural Networks*, 7(3):507–522, 1994.
19. M. S. A. Nadeem, J.-D. Zucker, and B. Hanczar. Accuracy-Rejection Curves (ARCs) for Comparing Classification Methods with a Reject Option. In *MLSB*, pages 65–81, 2010.
20. J. C. Platt. Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods. In *Advances in Large Margin Classifiers*. MIT Press, May 23 1999.
21. A. Sato and K. Yamada. Generalized Learning Vector Quantization. In *Advances in Neural Information Processing Systems*, volume 7, pages 423–429, 1995.
22. P. Schneider, M. Biehl, and B. Hammer. Distance learning in discriminative vector quantization. *Neural Computation*, 21(10):2942–2969, 2009.
23. S. Seo and K. Obermayer. Soft Learning Lector Quantization. *Neural Computation*, 15(7):1589–1604, Jul 2003.
24. H. H. Thodberg. Tecator data set, contained in StatLib Datasets Archive, 1995.
25. A. Vailaya and A. K. Jain. Reject Option for VQ-Based Bayesian Classification. In *Int. Conf. on Pattern Recognition (ICPR)*, pages 2048–2051, 2000.
26. T. Villmann and S. Haase. Divergence-Based Vector Quantization. *Neural Computation*, 23(5):1343–1392, 2011.
27. T.-F. Wu, C.-J. Lin, and R. C. Weng. Probability Estimates for Multi-class Classification by Pairwise Coupling. *J. of Machine Learning Research*, 5:975–1005, 2004.

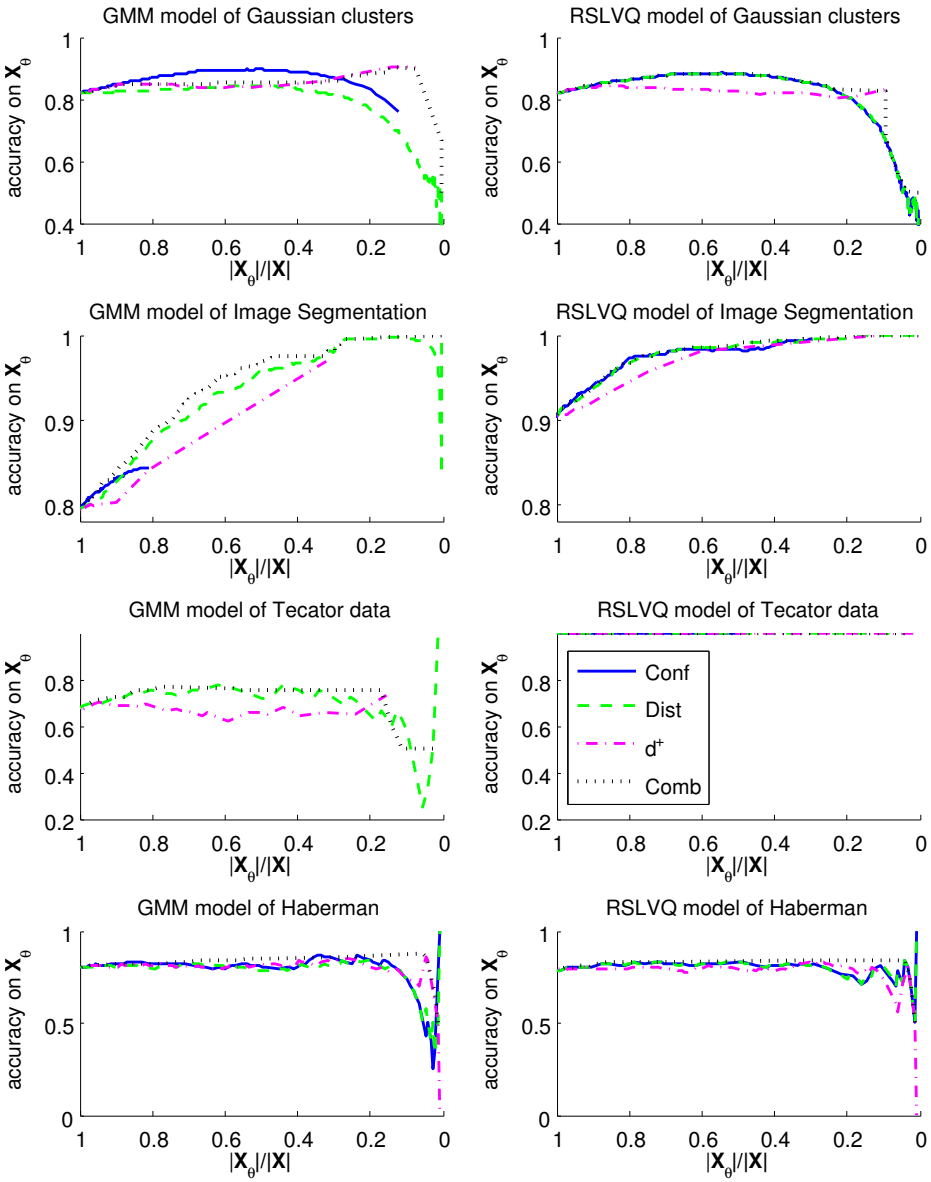


Fig. 2. Results of different rejection options when applied to generative or discriminative models trained for different data sets. We report the relative size of $\boldsymbol{X}_\theta$ as compared to the accuracy of the classifier on this set [19].