View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by DSpace@MIT

Journal of Machine Learning Research 11 (2010) 491-516 Submitted 3/09; Revised 8/09; Published 2/10

Classification Using Geometric Level Sets

Kush R. Varshney KRV@MIT.EDU
Alan S. Willsky WILLSKY@MIT.EDU
Laboratory for Information and Decision Systems
Massachusetts Institute of Technology
Cambridge, MA 02139-4309, USA

Editor: Ulrike von Luxburg

Abstract

A variational level set method is developed for the supervised classification problem. Nonlinear classifier decision boundaries are obtained by minimizing an energy functional that is composed of an empirical risk term with a margin-based loss and a geometric regularization term new to machine learning: the surface area of the decision boundary. This geometric level set classifier is analyzed in terms of consistency and complexity through the calculation of its ε-entropy. For multicategory classification, an efficient scheme is developed using a logarithmic number of decision functions in the number of classes rather than the typical linear number of decision functions. Geometric level set classification yields performance results on benchmark data sets that are competitive with well-established methods.

Keywords: level set methods, nonlinear classification, geometric regularization, consistency, complexity

1. Introduction

Variational level set methods, pioneered by Osher and Sethian (1988), have found application in fluid mechanics, computational geometry, image processing and computer vision, computer graphics, materials science, and numerous other fields, but have heretofore found little application in machine learning. The goal of this paper is to introduce a level set approach to the archetypal machine learning problem of supervised classification. We propose an implicit level set representation for classifier decision boundaries, a margin-based objective regularized by a surface area penalty, and an Euler-Lagrange descent optimization algorithm for training.

Several well-developed techniques for supervised discriminative learning exist in the literature, including the perceptron algorithm (Rosenblatt, 1958), logistic regression (Efron, 1975), and support vector machines (SVMs) (Vapnik, 1995). All of these approaches, in their basic form, produce linear decision boundaries. Nonlinear boundaries in the input space can be obtained by mapping the input space to a feature space of higher (possibly infinite) dimension by taking nonlinear functions of the input variables. Learning algorithms are then applied to the new higher-dimensional feature space by treating each dimension linearly. They retain the efficiency of the input lower-dimensional space for particular sets of nonlinear functions that admit the kernel trick (Schölkopf and Smola, 2002).

As an alternative to kernel methods for generalizing linear methods, we propose finding nonlinear decision boundaries directly in the input space. We propose an energy functional for clas-

©2010 Kush R. Varshney and Alan S. Willsky.

sification that is composed of an empirical risk term that uses a margin-based loss function and a complexity term that is the length of the decision boundary for a two-dimensional input space and the surface area of the decision boundary more generally. The empirical risk term is standard in many classification methods. What is new in this work is the measurement of decision boundary complexity by surface area, an inherently geometric quantity, and the idea of using variational level set methods for optimization in discriminative learning.

We use the term *contour* to refer to a one-dimensional curve in a two-dimensional space, a two-dimensional surface in a three-dimensional space, and generally a $D-1$ dimensional hypersurface in a D-dimensional space. Classifier decision boundaries partition the input space into regions corresponding to the different class labels. If the region corresponding to one class label is composed of several unconnected pieces, then the corresponding decision boundary is composed of several unconnected pieces; we refer to this entire collection of decision boundaries as the contour. The level set representation is a flexible, implicit representation for contours that does not require knowing the number of disjoint pieces in advance. The contour is represented by a smooth, Lipschitz continuous, scalar-valued function, known as the *level set function*, whose domain is the input space. The contour is implicitly specified as the zero level set of the level set function.

Level set methods entail not only the representation, but also the minimization of an energy functional whose argument is the contour.[1] For example, in foreground-background image segmentation, a popular energy functional is mean squared error of image intensity with two different ‘true’ image intensities inside the contour and outside the contour. Minimizing this energy produces a good segmentation when the two regions differ in image intensity. In order to perform the minimization, a gradient descent approach is used. The first variation of the functional is found using the calculus of variations; starting from an initial contour, a gradient flow is followed iteratively to converge to a minimum. This procedure is known as *curve evolution* or *contour evolution*.

The connection between level set methods (particularly for image segmentation) and classification has been noticed before, but to the best of our knowledge, there has been little prior work in this area. Boczko et al. (2006) only hint at the idea of using level set methods for classification. Tomczyk and Szczepaniak (2005) do not consider fully general input spaces. Specifically, examples in the training and test sets must be pixels in an image with the data vector containing the spatial index of the pixel along with other variables. Cai and Sowmya (2007) do consider general feature spaces, but have a very different energy functional than our margin-based loss functional. Theirs is based on counts of training examples in grid cells and is similar to the mean squared error functional for image segmentation described earlier. Their learning is also based on one-class classification rather than standard discriminative classification, which is the framework we follow. Yip et al. (2006) use variational level set methods for density-based clustering in general feature spaces, rather than for learning classifiers.

Cremers et al. (2007) dichotomize image segmentation approaches into those that use spatially continuous representations and those that use spatially discrete representations, with level set methods being the main spatially continuous approaches. There have been methods using discrete representations that bear some ties to our methods. An example of a spatially discrete approach uses normalized graph cuts (Shi and Malik, 2000), a technique that has been used extensively in unsupervised learning for general features unrelated to images as well. Normalized decision boundary surface area is implicitly penalized in this discrete setting. Geometric notions of complexity in su-

[1] In the image segmentation literature, variational energy minimization approaches often go by the name *active contours*, whether implemented using level set methods or not.

pervised classification tied to decision boundary surface area have been suggested by Ho and Basu (2002), but also defined in a discrete way related to graph cuts. In contrast, the continuous formulation we employ using level sets involves very different mathematical foundations, including explicit minimization of a criterion involving surface area. Moreover, the continuous framework—and in particular the natural way in which level set functions enter into the criterion—lead to new gradient descent algorithms to determine optimal decision boundaries. By embedding our criterion in a continuous setting, the surface area complexity term is defined intrinsically rather than being defined in terms of the graph of available training examples.

There are some other methods in the literature for finding nonlinear decision boundaries directly in the input space related to image segmentation, but these methods use neither contour evolution for optimization, nor the surface area of the decision boundary as a complexity term, as in the level set classification method proposed in this paper. A connection is drawn between classification and level set image segmentation in Scott and Nowak (2006) and Willett and Nowak (2007), but the formulation is through decision trees, not contour evolution. Tomczyk (2005), Tomczyk and Szczepaniak (2006), and Tomczyk et al. (2007) present a simulated annealing formulation given the name adaptive potential active hypercontours for finding nonlinear decision boundaries in both the classification and clustering problems; their work considers the use of radial basis functions in representing the decision boundary. Pölzlbauer et al. (2008) construct nonlinear decision boundaries in the input space from connected linear segments. In some ways, their approach is similar to active contours methods in image segmentation such as snakes that do not use the level set representation: changes in topology of the decision boundary in the optimization are difficult to handle. (The implicit level set representation takes care of topology changes naturally.)

The theory of classification with Lipschitz functions was discussed by von Luxburg and Bousquet (2004). As mentioned previously, level set functions are Lipschitz functions and the specific level set function that we use, the *signed distance function*, has a unit Lipschitz constant. Von Luxburg and Bousquet minimize the Lipschitz constant, whereas in our formulation, the Lipschitz constant is fixed. The von Luxburg and Bousquet formulation requires the specification of a subspace of Lipschitz functions over which to optimize in order to prevent overfitting, but does not resolve the question of how to select this subspace. The surface area penalty that we propose provides a natural specification for subspaces of signed distance functions.

The maximum allowable surface area parameterizes nested subspaces. We calculate the ε-entropy (Kolmogorov and Tihomirov, 1961) of these signed distance function subspaces and use the result to characterize geometric level set classification theoretically. In particular, we look at the consistency and convergence of level set classifiers as the size of the training set grows. We also look at the Rademacher complexity (Koltchinskii, 2001; Bartlett and Mendelson, 2002) of level set classifiers.

For the multicategory classification problem with $M > 2$ classes, typically binary classification methods are extended using the *one-against-all* construction (Hsu and Lin, 2002). The one-against-all scheme represents the classifier with M decision functions. We propose a more parsimonious representation of the multicategory level set classifier that uses $\log_2 M$ decision functions.[2] A collection of $\log_2 M$ level set functions can implicitly specify M regions using a binary encoding like a Venn diagram (Vese and Chan, 2002). This proposed logarithmic multicategory classification is

[2] It is certainly possible to use one-against-all with the proposed level set classification methodology. In fact, there are M-category level set methods that use M level set functions (Samson et al., 2000; Paragios and Deriche, 2002), but they are less parsimonious than the approach we follow.

new, as there is no logarithmic formulation for M-category classification in the machine learning literature. The energy functional that is minimized has a multicategory empirical risk term and surface area penalties on $\log_2 M$ contours.

The level set representation of classifier decision boundaries, the surface area regularization term, the logarithmic multicategory classification scheme, and other contributions of this paper are not only interesting academically, but also practically. We compare the classification performance of geometric level set classification on several binary and multicategory data sets from the UCI Repository and find the results to be competitive with many classifiers used in practice.

Level set methods are usually implemented on a discretized grid, that is the values of the level set function are maintained and updated on a grid. In physics and image processing applications, it nearly always suffices to work in two- or three-dimensional spaces. In classification problems, however, the input data space can be high-dimensional. Implementation of level set methods for large input space dimension becomes cumbersome due to the need to store and update a grid of that large dimension. One way to address this practical limitation is to represent the level set function by a superposition of radial basis functions (RBFs) instead of on a grid (Cecil et al., 2004; Slabaugh et al., 2007; Gelas et al., 2007). We follow this implementation strategy in obtaining classification results.

In Section 2, we detail geometric level set classification in the binary case, giving the objective to be minimized and the contour evolution to perform the minimization. In Section 3, we provide theoretical analysis of the binary level set classifier given in Section 2. The main result is the calculation of the ε-entropy of the space of level set classifiers as a function of the maximum allowable decision boundary surface area; this result is then applied to characterize consistency and complexity. Section 4 goes over multicategory level set classification. In Section 5, we describe the RBF level set implementation and use that implementation to compare the classification test performance of geometric level set classification to the performance of several other classifiers. Section 6 concludes and provides a summary of the work.

2. Binary Geometric Level Set Classification

In the standard binary classification problem, we are given training data $\{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)\}$ with data vectors $\mathbf{x}_i \in \Omega \subset \mathbb{R}^D$ and class labels $y_i \in \{+1, -1\}$ drawn according to some unknown probability density function $p_{\mathbf{X},Y}(\mathbf{x}, y)$ and we would like to learn a classifier $\hat{y} : \Omega \to \{+1, -1\}$ that classifies previously unseen data vectors well. A popular approach specifies the classifier as $\hat{y}(\mathbf{x}) = \operatorname{sign}(\varphi(\mathbf{x}))$, where φ is a scalar-valued function. This function is obtained by minimizing the empirical risk over a model class $\mathcal{F}$:

$$\min_{\varphi \in \mathcal{F}} \sum_{i=1}^{n} \mathrm{L}(y_i \varphi(\mathbf{x}_i)). \tag{1}$$

A wide variety of margin-based loss functions L are employed in different classification methods, including the logistic loss in logistic regression, the hinge loss in the SVM, and the exponential loss in boosting: $\mathrm{L}_{\text{logistic}}(z) = \log(1 + e^{-z})$, $\mathrm{L}_{\text{hinge}}(z) = \max\{0, 1 - z\}$, and $\mathrm{L}_{\text{exponential}}(z) = e^{-z}$ (Bartlett et al., 2006; Lin, 2004). The loss function on which classifiers are typically evaluated is the misclassification zero-one loss: $\mathrm{L}_{\text{zero-one}}(z) = \operatorname{step}(-z)$, where step is the Heaviside unit step function.

Limiting the model class is a way to control the complexity of the learned classifier in order to increase its generalization ability. This is the central idea of the structural risk minimization principle (Vapnik, 1995). A model subclass can be specified directly through a constraint in the optimization by taking $\mathcal{F}$ to be the subclass in (1). Alternatively, it may be specified through a regularization term J with weight λ:

$$\min_{\varphi\in\mathcal{F}}\sum_{i=1}^{n}\mathrm{L}(y_i\varphi(\mathbf{x}_i))+\lambda\mathrm{J}(\varphi), \tag{2}$$

where $\mathcal{F}$ indicates a broader class within which the subclass is delineated by $\mathrm{J}(\varphi)$.

We propose a geometric regularization term novel to machine learning, the surface area of the decision boundary:

$$\mathrm{J}(\varphi)=\oint_{\varphi=0}d\mathbf{s}, \tag{3}$$

where $d\mathbf{s}$ is an infinitesimal surface area element on the decision boundary. Decision boundaries that shatter more points are more tortuous than decision boundaries that shatter fewer points. The regularization functional (3) promotes smooth, less tortuous decision boundaries. It is experimentally shown in Varshney and Willsky (2008) that with this regularization term, there is an inverse relationship between the regularization parameter λ and the Vapnik-Chervonenkis (VC) dimension of the classifier. An analytical discussion of complexity is provided in Section 3.3. The empirical risk term and regularization term must be properly balanced as the regularization term by itself drives the decision boundary to be a set of infinitesimal hyperspheres.

We now describe how to find a classifier that minimizes (2) with the new regularization term (3) using the level set methodology. As mentioned in Section 1, the level set approach implicitly represents a $(D-1)$-dimensional contour C in a D-dimensional space Ω by a scalar-valued, Lipschitz continuous function φ known as the level set function (Osher and Fedkiw, 2003). The contour is the zero level set of φ. Contour C partitions Ω into the regions $\mathcal{R}$ and $\mathcal{R}^c$, which can be simply connected, multiply connected, or composed of several components. The level set function $\varphi(\mathbf{x})$ satisfies the properties: $\varphi(\mathbf{x})<0$ for $\mathbf{x}\in\mathcal{R}$, $\varphi(\mathbf{x})>0$ for $\mathbf{x}\in\mathcal{R}^c$, and $\varphi(\mathbf{x})=0$ for $\mathbf{x}$ on the contour C.

The level set function is often specialized to be the signed distance function, including in our work. The magnitude of the signed distance function at a point equals its distance to C, and its sign indicates whether it is in $\mathcal{R}$ or $\mathcal{R}^c$. The signed distance function satisfies the additional constraint that $\|\nabla\varphi(\mathbf{x})\|=1$ and has Lipschitz constant equal to one. Illustrating the representation, a contour in a $D=2$-dimensional space and its corresponding signed distance function are shown in Figure 1. For classification, we take $\mathcal{F}$ to be the set of all signed distance functions on the domain Ω in (2).

The general form of objective functionals in variational level set methods is

$$\mathrm{E}(C)=\int_{\mathcal{R}}g_{\mathrm{r}}(\mathbf{x})d\mathbf{x}+\oint_{C}g_{\mathrm{b}}(C(\mathbf{s}))d\mathbf{s}, \tag{4}$$

where g_{r} is a region-based function and g_{b} is a boundary-based function. Note that the integral in the region-based functional is over $\mathcal{R}$, which is determined by C. Region-based functionals may also be integrals over $\mathcal{R}^c$, in place of or in addition to integrals over $\mathcal{R}$. In contour evolution, starting from some initial contour, the minimum of (4) is approached iteratively via a flow in the negative gradient direction. If we parameterize the iterations of the flow with a time parameter t, then it may be shown using the calculus of variations that the flow of C that implements the gradient descent is

$$\frac{\partial C}{\partial t}=-g_{\mathrm{r}}\mathbf{n}-\left(g_{\mathrm{b}}\kappa-\langle\nabla g_{\mathrm{b}},\mathbf{n}\rangle\right)\mathbf{n}, \tag{5}$$

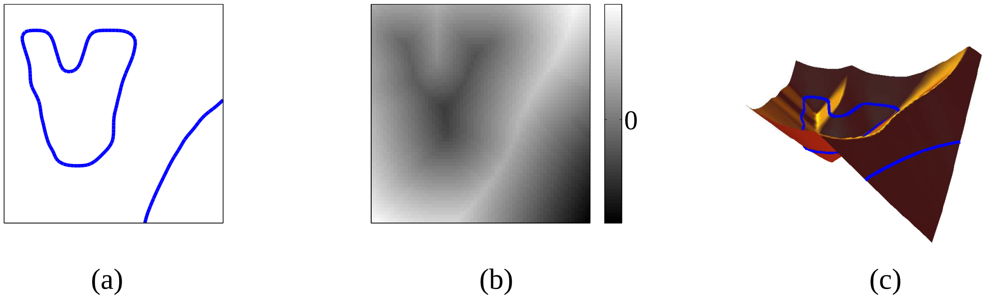

Figure 1: An illustration of the signed distance function representation of a contour with $D = 2$. The contour is shown in (a), its signed distance function is shown by shading in (b), and as a surface plot marked with the zero level set in (c).

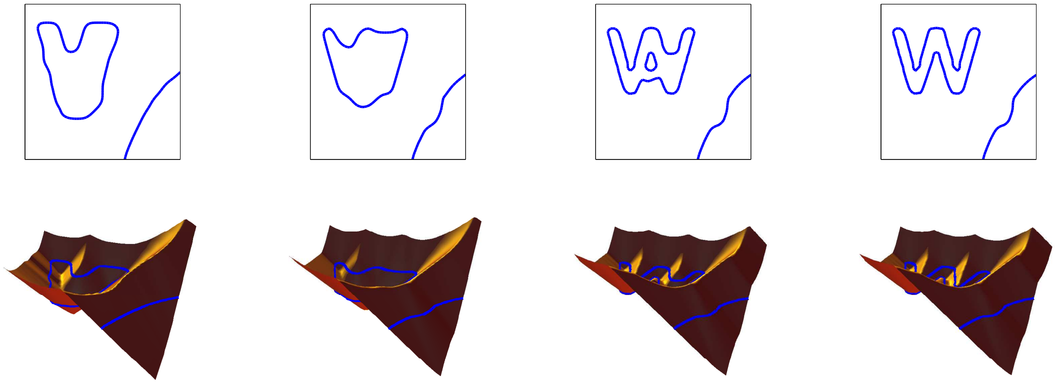

Figure 2: Iterations of an illustrative curve evolution proceeding from left to right. The top row shows the curve and the bottom row shows the corresponding signed distance function.

where $\mathbf{n}$ is the outward unit normal to C, and κ is its mean curvature (Caselles et al., 1997; Osher and Fedkiw, 2003). The mean curvature of a surface is an extrinsic measure of curvature from differential geometry that is the average of the principal curvatures. If the region-based function is integrated over $\mathcal{R}^c$, then the sign of the first term in (5) is reversed.

The flow of the contour corresponds to a flow of the signed distance function. The unit normal to the contour is $\nabla\varphi$ in terms of the signed distance function and the mean curvature is $\nabla^2\varphi$. The level set flow corresponding to (5) is

$$\frac{\partial\varphi(\mathbf{x})}{\partial t} = -g_{\mathrm{r}}(\mathbf{x})\nabla\varphi(\mathbf{x}) - \left(g_{\mathrm{b}}(\mathbf{x})\nabla^2\varphi(\mathbf{x}) - \langle\nabla g_{\mathrm{b}}(\mathbf{x}), \nabla\varphi(\mathbf{x})\rangle\right)\nabla\varphi(\mathbf{x}). \tag{6}$$

Figure 2 illustrates iterations of contour evolution.

For the classification problem, we have the following energy functional to be minimized:

$$\mathrm{E}(\varphi) = \sum_{i=1}^{n} \mathrm{L}(y_i\varphi(\mathbf{x}_i)) + \lambda \oint_C d\mathbf{s}. \tag{7}$$

The surface area regularization is a boundary-based functional with $g_\mathrm{b} = 1$ and the margin-based loss can be expressed as a region-based functional with g_r incorporating $\mathrm{L}(y_i\varphi(\mathbf{x}_i))$. Applying (4)–(6) to this energy functional yields the gradient descent flow

$$\left.\frac{\partial\varphi(\mathbf{x})}{\partial t}\right|_{\mathbf{x}=\mathbf{x}_i} = \begin{cases} \mathrm{L}(y_i\varphi(\mathbf{x}_i))\nabla\varphi(\mathbf{x}_i) - \lambda\nabla^2\varphi(\mathbf{x}_i)\nabla\varphi(\mathbf{x}_i), & \varphi(\mathbf{x}_i) < 0 \\ -\mathrm{L}(y_i\varphi(\mathbf{x}_i))\nabla\varphi(\mathbf{x}_i) - \lambda\nabla^2\varphi(\mathbf{x}_i)\nabla\varphi(\mathbf{x}_i), & \varphi(\mathbf{x}_i) > 0 \end{cases}. \tag{8}$$

In doing the contour evolution, note that we never compute the surface area of the decision boundary, which is oftentimes intractable, but just its gradient descent flow.

The derivative (8) does not take the constraint $\|\nabla\varphi\| = 1$ into account: the result of updating a signed distance function using (8) is not a signed distance function. Frequent reinitialization of the level set function as a signed distance function is important because (7) depends on the magnitude of φ, not just its sign. This reinitialization is done iteratively using (Sussman et al., 1994):

$$\frac{\partial\varphi(\mathbf{x})}{\partial t} = \operatorname{sign}(\varphi(\mathbf{x}))(1 - \|\nabla\varphi(\mathbf{x})\|).$$

With linear margin-based classifiers, including the original primal formulation of the SVM, the concept of margin is equivalent to Euclidean distance from the decision boundary in the input space. With kernel methods, however, this equivalence is lost; the quantity referred to as the margin, $y\varphi(\mathbf{x})$, is not the same as distance from $\mathbf{x}$ to the decision boundary in the input space. As discussed by Akaho (2004), oftentimes it is of interest to maximize the minimum distance to the decision boundary in the input space among all of the training examples. With the signed distance function representation, the margin $y\varphi(\mathbf{x})$ is equivalent to Euclidean distance from the decision boundary and hence is a satisfying nonlinear generalization to linear margin-based methods.

We now present two synthetic examples to illustrate this approach and its behavior. In both examples, there are $n = 1000$ points in the training set with $D = 2$. The first example has 502 points with label $y_i = -1$ and 498 points with label $y_i = +1$ and is separable by an elliptical decision boundary. The second example has 400 points with label $y_i = -1$ and 600 points with label $y_i = +1$ and is not separable by a simple shape, but has the -1 labeled points in a strip.

In these two examples, in the other examples in the rest of the paper, and in the performance results of Section 5.2, we use the logistic loss function for L in the objective (7). In these two examples, the surface area penalty has weight $\lambda = 0.5$; the value $\lambda = 0.5$ is a default parameter value that gives good performance with a variety of data sets regardless of their dimensionality D and can be used if one does not wish to optimize λ using cross-validation. Contour evolution minimization requires an initial decision boundary. In the portion of Ω where there are no training examples, we set the initial decision boundary to be a uniform grid of small components; this small seed initialization is common in level set methods. In the part of Ω where there are training examples, we use the locations and labels of the training examples to set the initial decision boundary. We assign a positive value to the initial signed distance function in locations of positively labeled examples and a negative value in locations of negatively labeled examples. The initial decision boundaries for the two examples are shown in the top left panels of Figure 3 and Figure 4.

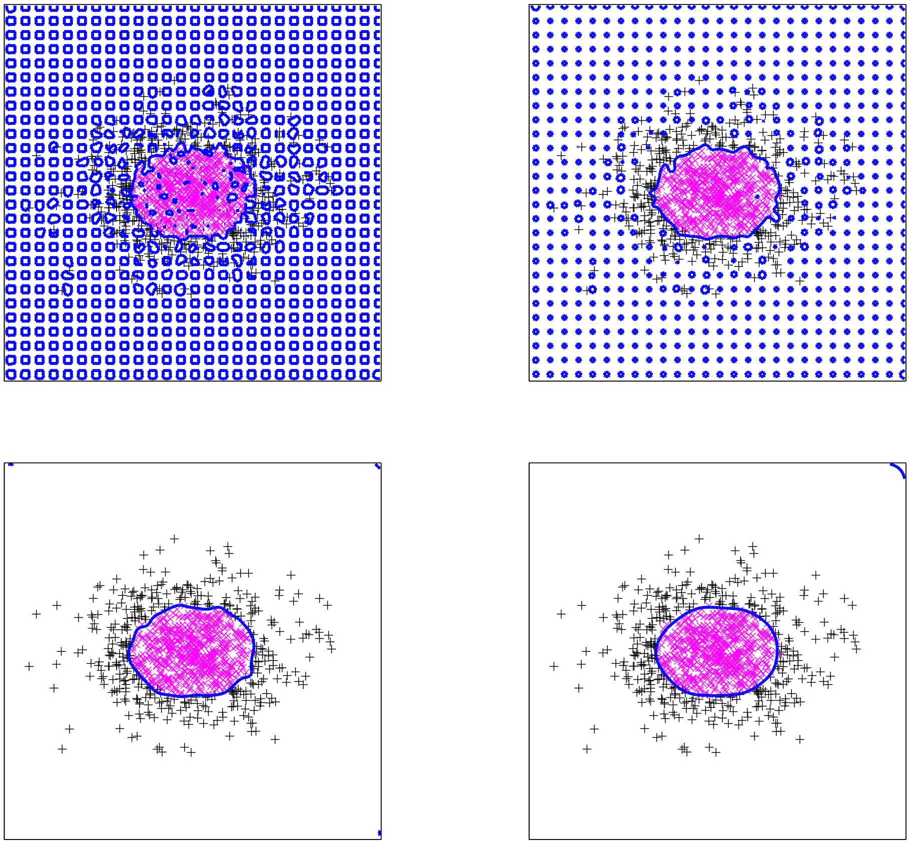

Figure 3: Curve evolution iterations with $\lambda = 0.5$ for an example training set proceeding in raster scan order starting from the top left. The magenta $\times$ markers indicate class label -1 and the black $+$ markers indicate class label $+1$. The blue line is the decision boundary.

Two intermediate iterations and the final decision boundary are also shown in Figure 3 and Figure 4. Solutions are as expected: an elliptical decision boundary and a strip-like decision boundary have been recovered. In the final decision boundaries of both examples, there is a small curved piece of the decision boundary in the top right corner of Ω where there are no training examples. This piece is an artifact of the initialization and the regularization term, and does not affect classifier performance. (The corner piece of the decision boundary is a minimal surface, a surface of zero mean curvature, which is a critical point of the surface area regularization functional (3), but not the global minimum. It is not important, assuming we have a representative training set.)

For a visual comparison of the effect of the surface area penalty weight, we show the solution decision boundaries of the geometric level set classifier for two other values of λ, 0.005 and 0.05, with the data set used in the example of Figure 4. As can be seen in comparing this figure with the bottom right panel of Figure 4, the smaller the value of λ, the longer and more tortuous the decision boundary. Small values of λ lead to overfitting.

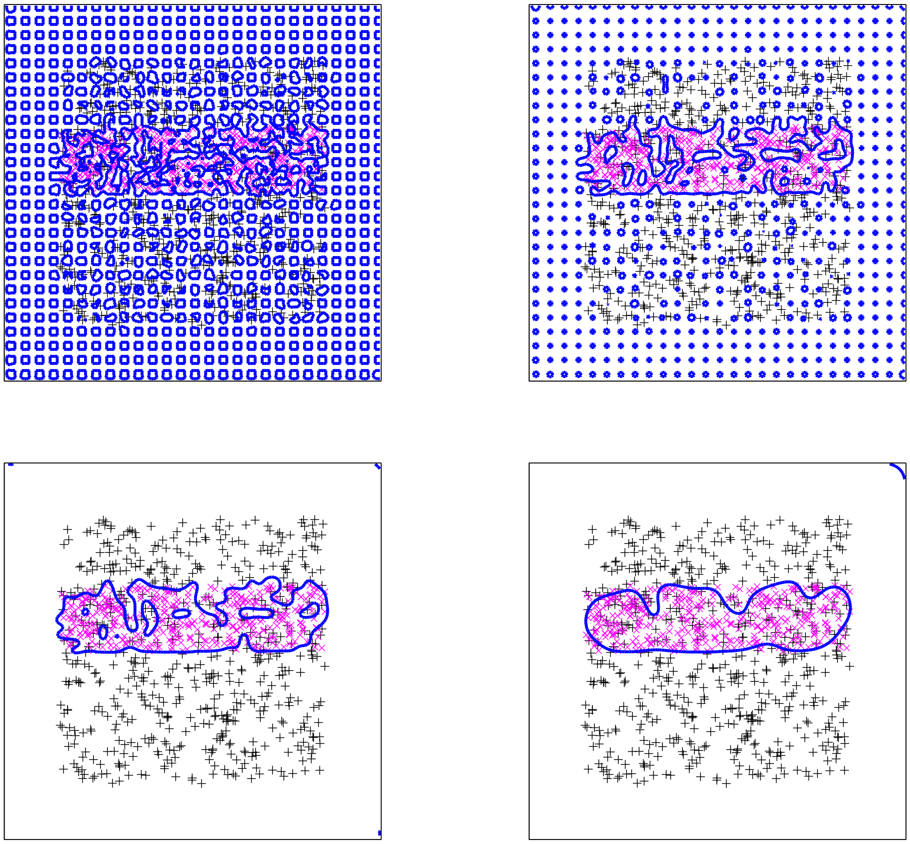

Figure 4: Curve evolution iterations with $\lambda = 0.5$ for an example training set proceeding in raster scan order starting from the top left. The magenta $\times$ markers indicate class label -1 and the black $+$ markers indicate class label $+1$. The blue line is the decision boundary.

In this section, we have described the basic method for nonlinear margin-based binary classification based on level set methods and illustrated its operation on two synthetic data sets. The next two sections build upon this core binary level set classification in two directions: theoretical analysis, and multicategory classification.

3. Consistency and Complexity Analysis

In this section, we provide analytical characterizations of the consistency and complexity of the level set classifier with surface area regularization described in Section 2. The main tool used in these characterizations is ε-entropy. Once we have an expression for the ε-entropy of the set of geometric level set classifiers, we can then apply consistency and complexity results from learning theory that are based on it. The beginning of this section is devoted to finding the ε-entropy of the space of signed distance functions with a surface area constraint with respect to the uniform or L_∞ metric on

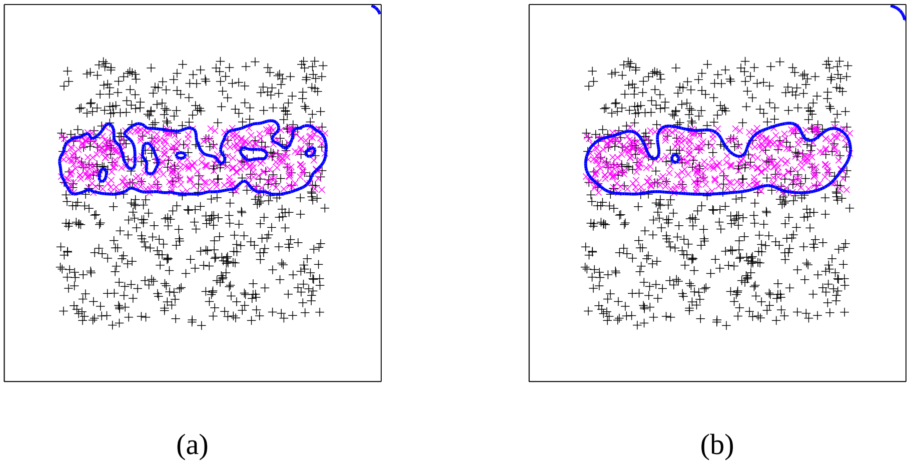

Figure 5: Solution decision boundaries with (a) $\lambda = 0.005$ and (b) $\lambda = 0.05$ for an example training set. The magenta $\times$ markers indicate class label -1 and the black $+$ markers indicate class label $+1$. The blue line is the decision boundary.

functions. The end of the section gives results on classifier consistency and complexity. The main findings are that level set classifiers are consistent, and that complexity is monotonically related to the surface area constraint, and thus the regularization term can be used to prevent underfitting and overfitting.

3.1 ε-Entropy

The ε-covering number of a metric space is the minimal number of sets with radius not exceeding ε required to cover that space. The ε-entropy is the base-two logarithm of the ε-covering number. These quantities are useful values in characterizing learning (Kulkarni, 1989; Williamson et al., 2001; Lin, 2004; von Luxburg and Bousquet, 2004; Steinwart, 2005; Bartlett et al., 2006). Kolmogorov and Tihomirov (1961), Dudley (1974, 1979), and others provide ε-entropy calculations for various classes of functions and various classes of sets, but the particular class we are considering, signed distance functions with a constraint on the surface area of the zero level set, does not appear in the literature. The second and third examples in Section 2 of Kolmogorov and Tihomirov (1961) are related, and the general approach we take for obtaining the ε-entropy of level set classifiers is similar to those two examples.

In classification, it is always possible to scale and shift the data and this is often done in practice. Without losing much generality and dispensing with some bothersome bookkeeping, we consider signed distance functions defined on the unit hypercube, that is $\Omega = [0,1]^D$, and we employ the uniform or L_∞ metric, $\rho_\infty(\varphi_1, \varphi_2) = \sup_{\mathbf{x}\in\Omega} |\varphi_1(\mathbf{x}) - \varphi_2(\mathbf{x})|$. We denote the set of all signed distance functions whose zero level set has surface area less than s by $\mathcal{F}_s$, its ε-covering number with respect to the uniform metric as $N_{\rho_\infty,\varepsilon}(\mathcal{F}_s)$, and its ε-entropy as $H_{\rho_\infty,\varepsilon}(\mathcal{F}_s)$. We begin with the $D = 1$ case and then come to general D.

Figure 6a shows a signed distance function over the unit interval. Due to the $\|\nabla\varphi\| = 1$ con-

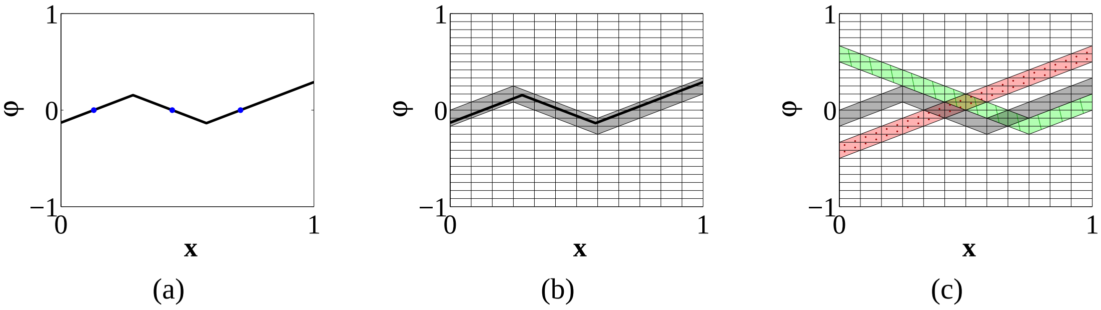

Figure 6: A $D = 1$-dimensional signed distance function in $\Omega = [0,1]$ is shown in (a), marked with its zero level set. The ε-corridor with $\varepsilon = \frac{1}{12}$ that contains the signed distance function is shown in (b), shaded in gray. The ε-corridor of (b), whose center line has three zero crossings is shown in (c), again shaded in gray, along with an ε-corridor whose center line has two zero crossings, shaded in green with stripes, and an ε-corridor whose center line has one zero crossing, shaded in red with dots.

straint, its slope is either $+1$ or -1 almost everywhere. The slope changes sign exactly once between two consecutive points in the zero level set. The signed distance function takes values in the range between positive and negative one.[3] In the $D = 1$ context, by surface area we mean the number of points in the zero level set, for example three in Figure 6a.

In finding $H_{\rho_\infty,\varepsilon}(\mathcal{F}_s)$, we will use sets known as *ε-corridors*, which are particular balls of radius ε measured using ρ_∞ in the space of signed distance functions. We use the terminology of Kolmogorov and Tihomirov (or translator Hewitt), but our definition is slightly different than theirs. An ε-corridor is a strip of height 2ε for all **x**. Let us define $\nu = \lceil \varepsilon^{-1} \rceil$. At $\mathbf{x} = 0$, the bottom and top of a corridor are at $2j\varepsilon$ and $2(j+1)\varepsilon$ respectively for some integer j, where $-\nu \leq 2j < \nu$. The slope of the corridor is either $+1$ or -1 for all **x** and the slope can only change at values of **x** that are multiples of ε. Additionally, the center line of the ε-corridor is a signed distance function, changing slope halfway between consecutive points in its zero level set and only there. The ε-corridor in which the signed distance function of Figure 6a falls is indicated in Figure 6b. Other ε-corridors are shown in Figure 6c.

By construction, each signed distance function is a member of exactly one ε-corridor. This is because since at $\mathbf{x} = 0$ the bottom and top of ε-corridors are at consecutive integer multiples of 2ε and since the center line of the corridor is a signed distance function, each signed distance function starts in one ε-corridor at $\mathbf{x} = 0$ and does not escape from it in the interval $(0,1]$. Also, an ε-corridor whose center line has s points in its zero level set contains only signed distance functions with at least s points in their zero level sets.

[3] There are several ways to define the signed distance function in the degenerate cases ($\mathcal{R} = \Omega$, $\mathcal{R}^c = \emptyset$) and ($\mathcal{R} = \emptyset$, $\mathcal{R}^c = \Omega$), including the assignments $-\infty$ and $+\infty$, or -1 and $+1$ (Delfour and Zolésio, 2001). For our purposes, it suffices to say that we have chosen a unique function for the $\mathcal{R} = \Omega$ case and a unique function for the $\mathcal{R}^c = \Omega$ case.

Theorem 1 *The ε-entropy of the set of signed distance functions defined over $\Omega = [0,1]$ with zero level set having less than s points is:*

$$H_{\rho_\infty,\varepsilon}(\mathcal{F}_s) = \log_2\left(\sum_{k=1}^{s}\binom{\lceil\varepsilon^{-1}\rceil - 1}{k-1}\right) + 1.$$

Proof Since ε-corridors only change slope at multiples of ε, we can divide the abscissa into ν pieces. (Each piece has width ε except the last one if ε^{-1} is not an integer.) In each of the ν subintervals, the center line of a corridor is either wholly positive or wholly negative. Enumerating the full set of ε-corridors is equivalent to enumerating binary strings of length ν. Thus, without a constraint s, there are 2^ν ε-corridors. Since, by construction, ε-corridors tile the space of signed distance functions, $N_{\rho_\infty,\varepsilon}(\mathcal{F}) = 2^\nu$.

With the s constraint on ε-corridors, the enumeration is equivalent to twice the number of compositions of the positive integer ν by a sum of s or less positive integers. Twice because for every composition, there is one version in which the first subinterval of the corridor center is positive and one version in which it is negative. As an example, the red corridor in Figure 6c can be composed with two positive integers $(5+7)$, the green corridor by three $(7+4+1)$, and the gray corridor by four $(1+4+4+3)$. The number of compositions of ν by k positive integers is $\binom{\nu-1}{k-1}$. Note that the zero-crossings are unordered for this enumeration and that the set $\mathcal{F}_s$ includes all of the signed distance functions with surface area smaller than s as well. Therefore:

$$N_{\rho_\infty,\varepsilon}(\mathcal{F}_s) = 2\sum_{k=1}^{s}\binom{\nu-1}{k-1}.$$

The result then follows because $H_{\rho_\infty,\varepsilon}(\mathcal{F}_s) = \log_2 N_{\rho_\infty,\varepsilon}(\mathcal{F}_s)$. ■

The combinatorial formula in Theorem 1 is difficult to work with, so we give a highly accurate approximation.

Theorem 2 *The ε-entropy of the set of signed distance functions defined over $\Omega = [0,1]$ with zero level set having less than s points is:*

$$H_{\rho_\infty,\varepsilon}(\mathcal{F}_s) \approx \lceil\varepsilon^{-1}\rceil + \log_2\Phi\left(\frac{2s - \lceil\varepsilon^{-1}\rceil}{\sqrt{\lceil\varepsilon^{-1}\rceil - 1}}\right),$$

where Φ is the standard Gaussian cumulative distribution function (cdf).

Proof Note that for a binomial random variable Z with $(\nu - 1)$ Bernoulli trials having success probability $\frac{1}{2}$:

$$\Pr[Z < z] = 2^{-\nu}\cdot 2\sum_{k=1}^{z}\binom{\nu-1}{k-1},$$

and that $N_{\rho_\infty,\varepsilon}(\mathcal{F}_s) = 2^\nu \Pr[Z < s]$. The result follows from the de Moivre-Laplace theorem and continuity correction, which are used to approximate the binomial distribution with the Gaussian distribution. ■

The central limit theorem tells us that the approximation works well when the Bernoulli success probability is one half, which it is in our case, and when the number of trials is large, which corresponds to small ε. The continuous approximation is better in the middle of the domain, when $s \approx \nu/2$, than in the tails. However, in the tails, the calculation of the exact expression in Theorem 1 is tractable. Since Φ is a cdf taking values in the range zero to one, $\log_2 \Phi$ is nonpositive. The surface area constraint only serves to reduce the ε-entropy.

The ε-entropy calculation in Theorem 1 and Theorem 2 is for the $D = 1$ case. We now discuss the case with general D. Recall that $\Omega = [0,1]^D$. Once again, we construct ε-corridors that tile the space of signed distance functions. In the one-dimensional case, the ultimate object of interest for enumeration is a string of length ν with binary labels. In the two-dimensional case, the corresponding object is a ν-by-ν grid of ε-by-ε squares with binary labels, and in general a D-dimensional Cartesian grid of hypercubes of volume ε^D, ν on each side. The surface area of the zero level set is the number of interior faces in the Cartesian grid whose adjoining ε^D hypercubes have different binary labels.

Theorem 3 *The ε-entropy of the set of signed distance functions defined over $\Omega = [0,1]^D$ with zero level set having surface area less than s is:*

$$H_{\rho_\infty,\varepsilon}(\mathcal{F}_s) \approx \left\lceil \varepsilon^{-1} \right\rceil^D + \log_2 \Phi\left(\frac{2s - D\left(\left\lceil \varepsilon^{-1} \right\rceil - 1\right)\left\lceil \varepsilon^{-1} \right\rceil^{D-1} - 1}{\sqrt{D\left(\left\lceil \varepsilon^{-1} \right\rceil - 1\right)\left\lceil \varepsilon^{-1} \right\rceil^{D-1}}} \right),$$

where Φ is the standard Gaussian cdf.

Proof In the one-dimensional case, it is easy to see that the number of segments is ν and the number of interior faces is $\nu - 1$. For a general D-dimensional Cartesian grid with ν hypercubes on each side, the number of hypercubes is ν^D and the number of interior faces is $D(\nu-1)\nu^{D-1}$. The result follows by substituting ν^D for ν and $D(\nu-1)\nu^{D-1}$ for $\nu - 1$ in Theorem 2. ■

Theorem 2 is a special case of Theorem 3 with $D = 1$. It is common to find the dimension of the space D in the exponent of ε^{-1} in ε-entropy calculations as we do here.

The ε-entropy calculation for level set classifiers given here enables us to analytically characterize their consistency properties as the size of the training set goes to infinity in Section 3.2 through ε-entropy-based classifier consistency results. The calculation also enables us to characterize the Rademacher complexity of level set classifiers in Section 3.3 through ε-entropy-based complexity results.

3.2 Consistency

In the binary classification problem, with training set of size n drawn from $p_{\mathbf{X},Y}(\mathbf{x},y)$, a consistent classifier is one whose probability of error converges in the limit as n goes to infinity to the probability of error of the Bayes optimal decision rule. The optimal decision rule to minimize the probability of error is $\hat{y}^*(\mathbf{x}) = \operatorname{sign}(p_{Y|\mathbf{X}}(Y = 1|\mathbf{X} = \mathbf{x}) - \frac{1}{2})$. Introducing notation, let the probability of error achieved by this decision rule be R^*. Also denote the probability of error of a level set classifier $\operatorname{sign}(\varphi^{(n)})$ learned from a training set of size n as $R(\operatorname{sign}(\varphi^{(n)}))$. For consistency, it is required that $R(\operatorname{sign}(\varphi^{(n)})) - R^*$ converge in probability to zero.

The learned classifier $\operatorname{sign}(\varphi^{(n)})$ minimizes the energy functional (7), and consequently the properties of $R(\operatorname{sign}(\varphi^{(n)}))$ are affected by both the margin-based loss function L and by the regularization term. Lin (2004), Steinwart (2005), and Bartlett et al. (2006) have given conditions on the loss function necessary for a margin-based classifier to be consistent. Common margin-based loss functions including the logistic loss and exponential loss meet the conditions. Lin calls a loss function that meets the necessary conditions *Fisher-consistent*. Fisher consistency of the loss function is not enough, however, to imply consistency of the classifier overall. The regularization term must also be analyzed; since the regularization term based on surface area we introduce is new, so is the following analysis.

Concentrating on the surface area regularization, we adapt Theorem 4.1 of Lin, which is based on ε-entropy. The analysis is based on the method of sieves, where sieves $\mathcal{F}_n$ are an increasing sequence of subspaces of a function space $\mathcal{F}$. In our case, $\mathcal{F}$ is the set of signed distance functions on Ω and the sieves, $\mathcal{F}_{s(n)}$, are subsets of signed distance functions whose zero level sets have surface area less than $s(n)$, that is $\oint_{\varphi=0} d\mathbf{s} < s(n)$. Such a constraint is related to the regularization expression $\mathrm{E}(\varphi)$ given in (7) through the method of Lagrange multipliers, with λ inversely related to $s(n)$. In the following, the function $s(n)$ is increasing in n and thus the conclusions of the theorem provide asymptotic results on consistency as the strength of the regularization term decreases as more training samples are made available. The sieve estimate is:

$$\varphi^{(n)} = \arg \min_{\varphi \in \mathcal{F}_{s(n)}} \sum_{i=1}^{n} \mathrm{L}(y_i \varphi(\mathbf{x}_i)). \tag{9}$$

Having found $H_{\rho_\infty,\varepsilon}(\mathcal{F}_s)$ in Section 3.1, we can apply Theorem 4.1 of Lin (2004), yielding the following theorem.

Theorem 4 Let L be a Fisher-consistent loss function in (9); let $\tilde{\varphi} = \arg\min_{\varphi\in\mathcal{F}} E[\mathrm{L}(Y\varphi(\mathbf{X}))]$, where $\mathcal{F}$ is the space of signed distance functions on $[0,1]^D$; and let $\mathcal{F}_{s(n)}$ be a sequence of sieves. Then for sieve estimate $\varphi^{(n)}$, we have[4]

$$R(\operatorname{sign}(\varphi^{(n)})) - R^* = O_P\left(\max\left\{n^{-\tau}, \inf_{\varphi\in\mathcal{F}_{s(n)}} \int (\varphi(\mathbf{x}) - \tilde{\varphi}(\mathbf{x}))^2 p_{\mathbf{X}}(\mathbf{x}) d\mathbf{x}\right\}\right),$$

where

$$\tau = \begin{cases} \frac{1}{3}, & D = 1 \\ \frac{1}{4} - \frac{\log\log n}{2\log n}, & D = 2 \\ \frac{1}{2D}, & D \geq 3 \end{cases}.$$

Proof The result is a direct application of Theorem 4.1 of Lin (2004), which is in turn an application of Theorem 1 of Shen and Wong (1994). In order to apply this theorem, we need to note two things. First, that signed distance functions on $[0,1]^D$ are bounded (by a value of 1) in the L_∞ norm. Second, that there exists an A such that $H_{\rho_\infty,\varepsilon}(\mathcal{F}_s) \leq A\varepsilon^{-D}$. Based on Theorem 3, we see that $H_{\rho_\infty,\varepsilon}(\mathcal{F}_s) \leq \nu^D$ because the logarithm of the cdf is nonpositive. Since $\nu = \lceil \varepsilon^{-1} \rceil$, if ε^{-1} is an integer, then $H_{\rho_\infty,\varepsilon}(\mathcal{F}_s) \leq \varepsilon^{-D}$ and otherwise there exists an A such that $H_{\rho_\infty,\varepsilon}(\mathcal{F}_s) \leq A\varepsilon^{-D}$. ■

[4] The notation $Z_n = O_P(\zeta_n)$ means that the random variable Z_n is bounded in probability at the rate ζ_n (van der Vaart, 1998).

Clearly $n^{-\tau}$ goes to zero as n goes to infinity. Also, $\inf_{\varphi \in \mathcal{F}_{s(n)}} \int (\varphi(\mathbf{x}) - \tilde{\varphi}(\mathbf{x}))^2 p_{\mathbf{X}}(\mathbf{x}) d\mathbf{x}$ goes to zero when $s(n)$ is large enough so that the surface area constraint is no longer applicable.[5] Thus, level set classifiers are consistent.

3.3 Rademacher Complexity

The principal idea of the structural risk minimization principle is that the generalization error is the sum of an empirical risk term and a capacity term (Vapnik, 1995). The two terms should be sensibly balanced in order to achieve low generalization error. Here, we use the ε-entropy of signed distance functions constrained in decision boundary surface area to characterize the capacity term. In particular, we look at the Rademacher complexity of $\mathcal{F}_s$ as a function of s (Koltchinskii, 2001; Bartlett and Mendelson, 2002).

The Rademacher average of a class $\mathcal{F}$, denoted $\hat{R}_n(\mathcal{F})$, satisfies (von Luxburg and Bousquet, 2004):

$$\hat{R}_n(\mathcal{F}) \leq 2\varepsilon + \frac{4\sqrt{2}}{\sqrt{n}} \int_{\frac{\varepsilon}{4}}^{\infty} \sqrt{H_{\rho_{2,n},\varepsilon'}(\mathcal{F})} d\varepsilon',$$

where $\rho_{2,n}(\varphi_1, \varphi_2) = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (\varphi_1(\mathbf{x}_i) - \varphi_2(\mathbf{x}_i))^2}$ is the empirical ℓ_2 metric. We found $H_{\rho_\infty,\varepsilon}(\mathcal{F}_s)$ for signed distance functions with surface area less than s in Section 3.1, and $H_{\rho_{2,n},\varepsilon}(\mathcal{F}) \leq H_{\rho_\infty,\varepsilon}(\mathcal{F})$. Thus, we can characterize the complexity of level set classifiers via the Rademacher capacity term (von Luxburg and Bousquet, 2004):

$$C_{\text{Rad}}(\mathcal{F}_s, n) = 2\varepsilon + \frac{4\sqrt{2}}{\sqrt{n}} \int_{\frac{\varepsilon}{4}}^{\infty} \sqrt{H_{\rho_\infty,\varepsilon'}(\mathcal{F})} d\varepsilon'. \tag{10}$$

With $\Omega = [0,1]^D$, the upper limit of the integral in (10) is one rather than infinity because ε cannot be greater than one.

In Figure 7, we plot C_{Rad} as a function of s for three values of D, and fixed ε and n. Having a fixed ε models the discretized grid implementation of level set methods. As the value of s increases, decision boundaries with more area are available. Decision boundaries with large surface area are more complex than smoother decision boundaries with small surface area. Hence the complexity term increases as a function of s. We have also empirically found the same relationship between the VC dimension and the surface area penalty (Varshney and Willsky, 2008). Consequently, the surface area penalty can be used to control the complexity of the classifier, and prevent underfitting and overfitting. The Rademacher capacity term may be used in setting the regularization parameter λ.

4. Multicategory Geometric Level Set Classification

Thus far, we have considered binary classification. In this section, we extend level set classification to the multicategory case with $M > 2$ classes labeled $y \in \{1, \dots, M\}$. We represent the decision boundaries using $m = \lceil \log_2 M \rceil$ signed distance functions $\{\varphi_1(\mathbf{x}), \dots, \varphi_m(\mathbf{x})\}$. Using such a set of level set functions we can represent 2^m regions $\{\mathcal{R}_1, \mathcal{R}_2, \dots, \mathcal{R}_{2^m}\}$ through a binary encoding

[5] For a given ε, there is a maximum possible surface area; the constraint is no longer applicable when the constraint is larger than this maximum possible surface area. Also note that s and λ are inversely related.

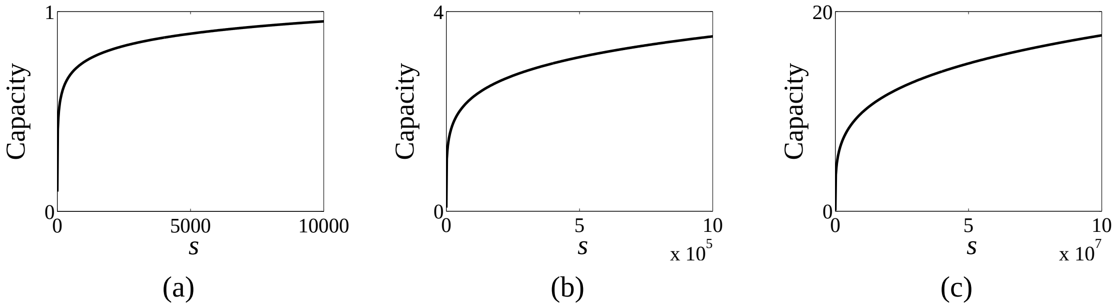

Figure 7: The Rademacher capacity term (10) as a function of s for signed distance functions on $\Omega = [0,1]^D$ with surface area less than s with (a) $D = 2$, (b) $D = 3$, and (c) $D = 4$. The values of ε and n are fixed at 0.01 and 1000 respectively.

(Vese and Chan, 2002). Thus, for $\mathbf{x} \in \mathcal{R}_1$, $(\varphi_1(\mathbf{x}) < 0) \wedge \cdots \wedge (\varphi_m(\mathbf{x}) < 0)$; for $\mathbf{x} \in \mathcal{R}_2$, $(\varphi_1(\mathbf{x}) < 0) \wedge \cdots \wedge (\varphi_{m-1}(\mathbf{x}) < 0) \wedge (\varphi_m > 0)$; and for $\mathbf{x} \in \mathcal{R}_{2^m}$, $(\varphi_1(\mathbf{x}) > 0) \wedge \cdots \wedge (\varphi_m(\mathbf{x}) > 0)$.

This binary encoding specifies the regions, but in order to apply margin-based loss functions, we also need a value for margin. In binary classification, the special encoding $y \in \{-1,+1\}$ allows $y\varphi(\mathbf{x})$ to be the argument to the loss function. For multicategory classification, the argument to the loss function is through functions $\psi_y(\mathbf{x})$, which are also specified through a binary encoding: $\psi_1(\mathbf{x}) = \max\{+\varphi_1(\mathbf{x}), \ldots, +\varphi_m(\mathbf{x})\}$, $\psi_2(\mathbf{x}) = \max\{+\varphi_1(\mathbf{x}), \ldots, +\varphi_{m-1}(\mathbf{x}), -\varphi_m(\mathbf{x})\}$, and $\psi_{2^m}(\mathbf{x}) = \max\{-\varphi_1(\mathbf{x}), \ldots, -\varphi_m(\mathbf{x})\}$. Then, the M-ary level set classification energy functional we propose is

$$\mathrm{E}(\varphi_1, \ldots, \varphi_m) = \sum_{i=1}^{n} \mathrm{L}(\psi_{y_i}(\mathbf{x}_i)) + \frac{\lambda}{m} \sum_{j=1}^{m} \oint_{\varphi_j = 0} d\mathbf{s}. \tag{11}$$

The same margin-based loss functions used in the binary case, such as the hinge and logistic loss functions, may be used in the multicategory case (Zou et al., 2006, 2008). The regularization term included in (11) is the sum of the surface areas of the zero level sets of the m signed distance functions.

The gradient descent flows for the m signed distance functions are

$$\left.\frac{\partial \varphi_1(\mathbf{x})}{\partial t}\right|_{\mathbf{x}=\mathbf{x}_i} = \begin{cases} \mathrm{L}(\psi_{y_i}(\mathbf{x}_i))\nabla\varphi_1(\mathbf{x}_i) - \frac{\lambda}{m}\nabla^2\varphi_1(\mathbf{x}_i)\nabla\varphi_1(\mathbf{x}_i), & \varphi_1(\mathbf{x}_i) < 0 \\ -\mathrm{L}(\psi_{y_i}(\mathbf{x}_i))\nabla\varphi_1(\mathbf{x}_i) - \frac{\lambda}{m}\nabla^2\varphi_1(\mathbf{x}_i)\nabla\varphi_1(\mathbf{x}_i), & \varphi_1(\mathbf{x}_i) > 0 \end{cases}$$

$$\vdots$$

$$\left.\frac{\partial \varphi_m(\mathbf{x})}{\partial t}\right|_{\mathbf{x}=\mathbf{x}_i} = \begin{cases} \mathrm{L}(\psi_{y_i}(\mathbf{x}_i))\nabla\varphi_m(\mathbf{x}_i) - \frac{\lambda}{m}\nabla^2\varphi_m(\mathbf{x}_i)\nabla\varphi_m(\mathbf{x}_i), & \varphi_m(\mathbf{x}_i) < 0 \\ -\mathrm{L}(\psi_{y_i}(\mathbf{x}_i))\nabla\varphi_m(\mathbf{x}_i) - \frac{\lambda}{m}\nabla^2\varphi_m(\mathbf{x}_i)\nabla\varphi_m(\mathbf{x}_i), & \varphi_m(\mathbf{x}_i) > 0 \end{cases}.$$

In the case $M = 2$ and $m = 1$, the energy functional and gradient flow revert back to binary level set classification described in Section 2.

The proposed multicategory classifier is different from the commonly used technique known as one-against-all (Hsu and Lin, 2002), which constructs an M-ary classifier from M binary classifiers, both because it treats all M classes simultaneously in the objective and because the decision regions are represented by a logarithmic rather than linear number of decision functions. Zou et al. (2006) also treat all M classes simultaneously in the objective, but their multicategory kernel machines use M decision functions. In fact, to the best of our knowledge, there is no M-ary classifier representation in the literature using as few as $\lceil \log_2 M \rceil$ decision functions. Methods that combine binary classifier outputs using error-correcting codes make use of a logarithmic number of binary classifiers with a larger multiplicative constant, such as $\lceil 10 \log M \rceil$ or $\lceil 15 \log M \rceil$ (Rifkin and Klautau, 2004; Allwein et al., 2000).

We give an example showing multicategory level set classification with $M = 4$ and $D = 2$. The data set has 250 points for each of the four class labels $y_i = 1$, $y_i = 2$, $y_i = 3$, and $y_i = 4$. The classes are not perfectly separable by simple boundaries. With four classes, we use $m = 2$ signed distance functions. Figure 8 shows the evolution of the two contours, the magenta and cyan curves. The final decision region for class $y = 1$ is the portion inside both the magenta and cyan curves, and coincides with the training examples with class label 1. The final decision region for class 2 is the region inside the magenta curve but outside the cyan curve, the final decision region for class 3 is the region inside the cyan curve, but outside the magenta curve, and the final decision region for class 4 is outside both curves. The final decision boundaries are fairly smooth and partition the space with small training error.

5. Implementation and Classification Results

In this section, we describe how to implement geometric level set classification practically using RBFs and give classification performance results when applied to several real binary and multicategory data sets.

5.1 Radial Basis Function Level Set Method

There have been many developments in level set methods since the original work of Osher and Sethian (1988). One development in particular is to represent the level set function by a superposition of RBFs instead of on a grid (Cecil et al., 2004; Slabaugh et al., 2007; Gelas et al., 2007). Grid-based representation of the level set function is not amenable to classification in high-dimensional input spaces because the memory and computational requirements are exponential in the dimension of the input space. A nonparametric RBF representation, however, is tractable for classification. The RBF level set method we use to minimize the energy functionals (7) and (11) for binary and multicategory margin-based classification is most similar to that described by Gelas et al. (2007) for image processing.

The starting point of the RBF level set approach is describing the level set function $\varphi(\mathbf{x})$ via a strictly positive definite[6] RBF $\mathrm{K}(\cdot)$ as follows:

$$\varphi(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i \,\mathrm{K}\left(\|\mathbf{x} - \mathbf{x}_i\|\right). \tag{12}$$

The zero level set of φ defined in this way is the contour C. For the classification problem, we take the centers $\mathbf{x}_i$ to be the data vectors of the training set. Then, constructing an $n \times n$ matrix $\mathbf{H}$ with

[6] A more complete discussion including conditionally positive definite RBFs would add a polynomial term to (12), to span the null space of the RBF (Wendland, 2005).

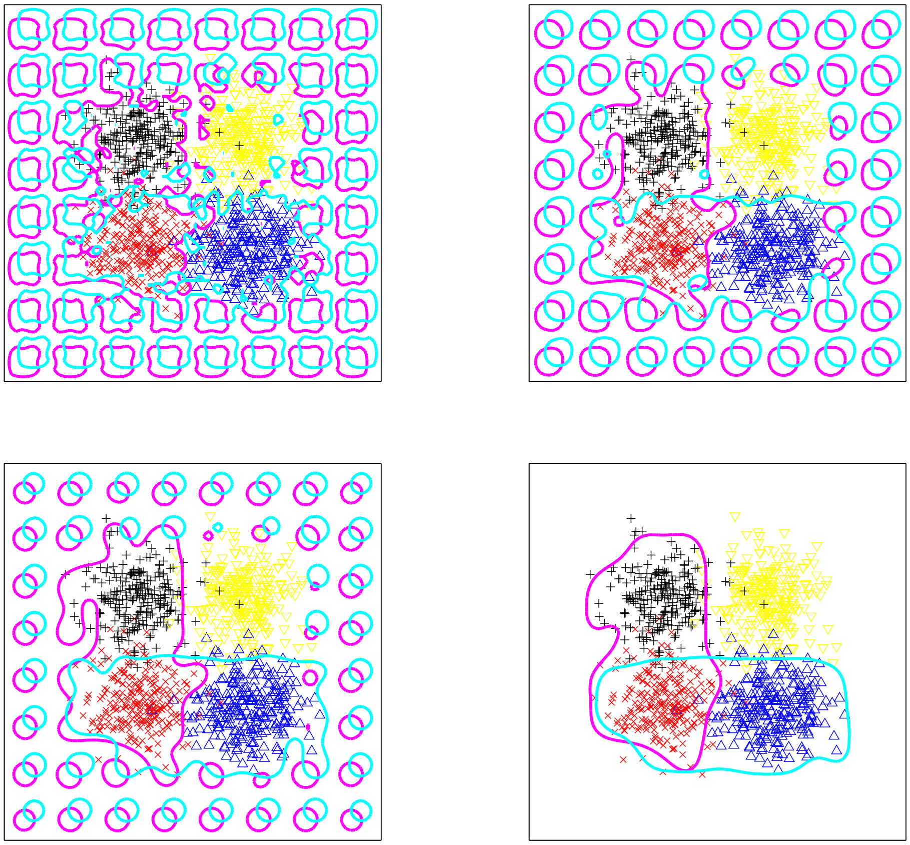

Figure 8: Curve evolution iterations with $\lambda = 0.5$ for multicategory classification proceeding in raster scan order. The red $\times$ markers indicate class label 1, the black $+$ markers indicate class label 2, the blue $\triangle$ markers indicate class label 3, and the yellow $\bigtriangledown$ markers indicate class label 4. The magenta and cyan lines are the zero level sets of the $m = 2$ signed distance functions and together make up the decision boundary.

elements $\{\mathbf{H}\}_{ij} = \mathrm{K}(\|\mathbf{x}_i - \mathbf{x}_j\|)$, and letting $\boldsymbol{\alpha}$ be the vector of coefficients in (12), we have:

$$\begin{bmatrix} \varphi(\mathbf{x}_1) \\ \vdots \\ \varphi(\mathbf{x}_n) \end{bmatrix} = \mathbf{H}\boldsymbol{\alpha}.$$

To minimize an energy functional of C, the level set optimization is over the coefficients $\boldsymbol{\alpha}$ with $\mathbf{H}$ fixed. In order to perform contour evolution with the RBF representation, a time parameter t is introduced like in Section 2, giving:

$$\mathbf{H}\frac{d\boldsymbol{\alpha}}{dt} = \begin{bmatrix} \left.\frac{\partial\varphi(\mathbf{x})}{\partial t}\right|_{\mathbf{x}=\mathbf{x}_1} \\ \vdots \\ \left.\frac{\partial\varphi(\mathbf{x})}{\partial t}\right|_{\mathbf{x}=\mathbf{x}_n} \end{bmatrix}. \tag{13}$$

For the binary margin-based classification problem with surface area regularization that we are interested in solving, we substitute the gradient flow (8) into the right side of (13). For the multicategory classification problem, we have m level set functions as discussed in Section 4 and each one has a gradient flow to be substituted into an expression like (13).

The iteration for the contour evolution is then:

$$\boldsymbol{\alpha}^{(k+1)} = \boldsymbol{\alpha}^{(k)} - \tau\mathbf{H}^{-1} \begin{bmatrix} \left.\frac{\partial\varphi^{(k)}(\mathbf{x})}{\partial t}\right|_{\mathbf{x}=\mathbf{x}_1} \\ \vdots \\ \left.\frac{\partial\varphi^{(k)}(\mathbf{x})}{\partial t}\right|_{\mathbf{x}=\mathbf{x}_n} \end{bmatrix}, \tag{14}$$

where τ is a small step size and $\varphi^{(k)}$ comes from $\boldsymbol{\alpha}^{(k)}$. We normalize $\boldsymbol{\alpha}$ according to the ℓ_1-norm after every iteration.

The RBF-represented level set function is not a signed distance function. However, as discussed by Gelas et al. (2007), normalizing the coefficient vector $\boldsymbol{\alpha}$ with respect to the ℓ_1-norm after every iteration of (14) has a similar effect as reinitializing the level set function as a signed distance function. The Lipschitz constant of the level set function is constrained by this normalization. The analysis of Section 3 applies with minor modification for level set functions with a given Lipschitz constant and surface area constraint. The RBF level set approach is similar to kernel machines with the RBF kernel in the sense that the decision function is represented by a linear combination of RBFs. However, kernel methods in the literature minimize a reproducing kernel Hilbert space squared norm for regularization, whereas the geometric level set classifier minimizes decision boundary surface area for regularization. The regularization term and consequently inductive bias of the geometric level set classifier is new and different compared to existing kernel methods. The solution decision boundary is the zero level set of a function of the form given in (12). Of course this representation does not capture all possible functions, but, given that we use a number of RBFs equal to the number of training examples, the granularity of this representation is well-matched to the data. This is similar to the situation found in other contexts such as kernel machines using RBFs.

We initialize the decision boundary with $\boldsymbol{\alpha} = n(\mathbf{H}^{-1}\mathbf{y})/\|\mathbf{H}^{-1}\mathbf{y}\|_1$, where $\mathbf{y}$ is a vector of the n class labels in the training set. Figure 9 shows this initialization and following RBF-implemented

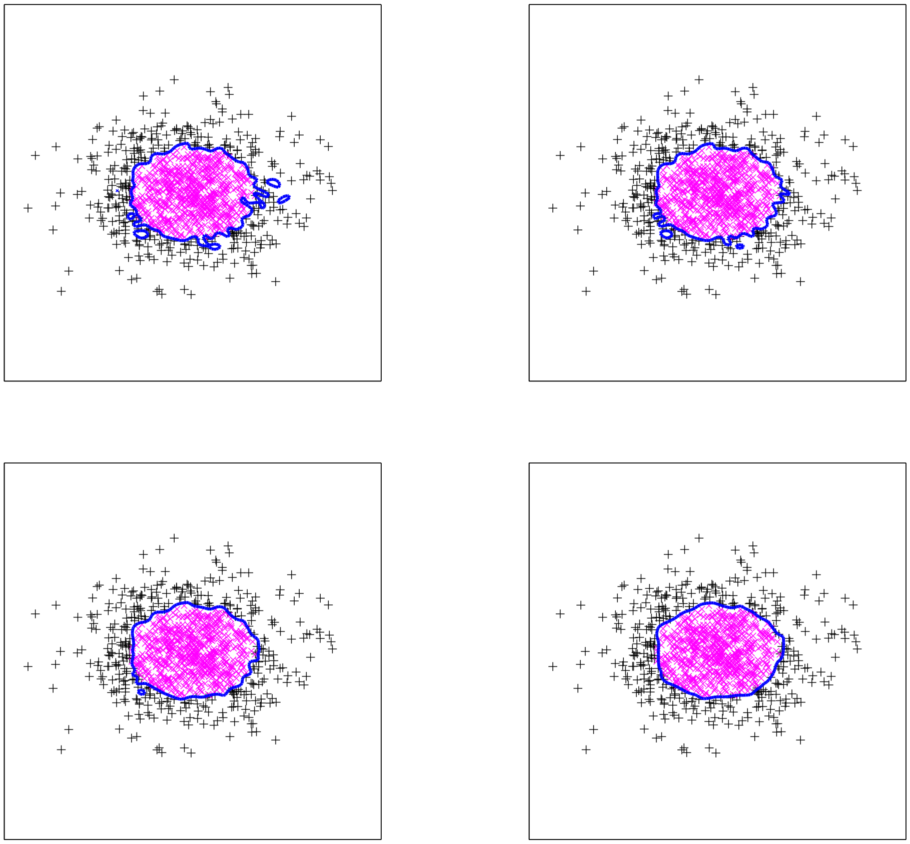

Figure 9: Curve evolution iterations with RBF implementation and $\lambda = 0.5$ for example training set proceeding in raster scan order. The magenta $\times$ markers indicate class label -1 and the black $+$ markers indicate class label $+1$. The blue line is the decision boundary.

contour evolution on the elliptically-separable data set presented in Section 2. The initial decision boundary is tortuous. It is smoothed out by the surface area penalty during the course of the contour evolution, thereby improving the generalization of the learned classifier as desired. To initialize the m vectors $\boldsymbol{\alpha}$ in M category classification, we use m length n vectors of positive and negative ones constructed from the binary encoding instead of $\mathbf{y}$.

5.2 Classification Results

We give classifier performance results on benchmark data sets from the UCI Machine Learning Repository (Asuncion and Newman, 2007) for geometric level set classification and compare them to the performance of several other classifiers, concluding that level set classification is a competitive technique. We present the tenfold cross-validation classification error performance with RBF level set implementation on four binary data sets: Pima Indians Diabetes ($n = 768$, $D = 8$), Wis-

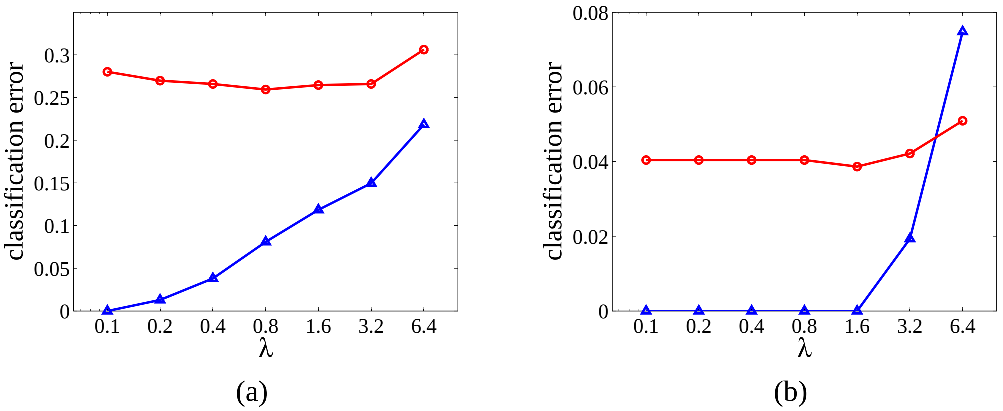

Figure 10: Tenfold cross-validation training error (blue line with triangle markers) and test error (red line with circle markers) for the (a) Pima, and (b) WDBC data sets as a function of the regularization parameter λ on a logarithmic scale.

consin Diagnostic Breast Cancer ($n = 569$, $D = 30$), BUPA Liver Disorders ($n = 345$, $D = 6$) and Johns Hopkins University Ionosphere ($n = 351$, $D = 34$), and four multicategory data sets: Wine Recognition ($n = 178$, $M = 3$, $D = 13$), Iris ($n = 150$, $M = 3$, $D = 4$), Glass Identification ($n = 214$, $M = 6$, $D = 9$), and Image Segmentation ($n = 2310$, $M = 7$, $D = 19$). For the binary data sets, there is $m = 1$ level set function, for the wine and iris data sets $m = 2$ level set functions, and for the glass and segmentation data sets $m = 3$ level set functions.

We shift and scale the data so that each of the input dimensions has zero mean and unit variance, use the RBF $\mathrm{K}(\|\mathbf{x} - \mathbf{x}_i\|) = e^{-\|\mathbf{x}-\mathbf{x}_i\|^2}$, the logistic loss function, $\tau = 1/m$, and the initialization $\boldsymbol{\alpha} = n(\mathbf{H}^{-1}\mathbf{y})/\|\mathbf{H}^{-1}\mathbf{y}\|_1$. First, we look at classification error as a function of λ. Figure 10 shows the tenfold cross-validation training and test errors for the Pima and WDBC data sets; other data sets yield similar plots. The plots show evidence of the structural risk minimization principle and complexity analysis given in Section 3.3. For small λ (corresponding to large surface area constraint s), the model class is too complex and we see that although the training error is zero, the test error is not minimal due to overfitting. For large λ, the model class is not complex enough; the training error is large and the test error is not minimal due to underfitting. There is an intermediate value of λ that achieves the minimal test error. However, we notice that the test error is fairly insensitive to the value of λ. The test error does not change much over the plotted range.

In Table 1, we report the tenfold cross-validation test error (as a percentage) on the eight data sets and compare the performance to nine other classifiers.[7] On each of the ten folds, we set λ using cross-validation. Specifically, we perform fivefold cross-validation on the nine tenths of the full data set that is the training data for that fold. We select the λ from the set of values $\{0.2, 0.4, 0.8, 1.6, 3.2\}$ that minimizes the fivefold cross-validation test error. The performance results of the nine other

[7] For lower-dimensional data sets (up to about $D = 12$), it is possible to use optimal dyadic decision trees (Scott and Nowak, 2006; Blanchard et al., 2007). We found that the results using such trees are not significantly better than those obtained using the C4.4 and C4.5 decision trees (which could be applied to all of the data sets without concern for dimensionality).

Data Set (M,D)	NB	BN	kNN	C4.4	C4.5	NBT	SVM	RBN	LLS	GLS
Pima $(2,8)$	23.69	25.64	27.86	27.33	26.17	25.64	22.66	24.60	29.94	25.94
WDBC $(2,30)$	7.02	4.92	3.68	7.20	6.85	7.21	2.28	5.79	6.50	4.04
Liver $(2,6)$	44.61	43.75	41.75	31.01	31.29	33.87	41.72	35.65	37.39	37.61
Ionos. $(2,34)$	17.38	10.54	17.38	8.54	8.54	10.27	11.40	7.38	13.11	13.67
Wine $(3,13)$	3.37	1.11	5.00	6.14	6.14	3.37	1.67	1.70	5.03	3.92
Iris $(3,4)$	4.00	7.33	4.67	4.00	4.00	6.00	4.00	4.67	3.33	6.00
Glass $(6,9)$	50.52	25.24	29.89	33.68	34.13	24.78	42.49	34.50	38.77	36.95
Segm. $(7,19)$	18.93	9.60	5.20	4.27	4.27	5.67	8.07	13.07	14.40	4.03

Table 1: Tenfold cross-validation error percentage of geometric level set classifier (GLS) with RBF level set implementation on several data sets compared to error percentages of various other classifiers reported in Cai and Sowmya (2007). The other classifiers are: naïve Bayes classifier (NB), Bayes net classifier (BN), *k*-nearest neighbor with inverse distance weighting (kNN), C4.4 decision tree (C4.4), C4.5 decision tree (C4.5), naïve Bayes tree classifier (NBT), SVM with polynomial kernel (SVM), radial basis function network (RBN), and learning level set classifier (LLS) of Cai and Sowmya (2007).

classifiers are as given by Cai and Sowmya (2007), who report the same tenfold cross-validation test error that we do for the geometric level set classifier. Details about parameter settings for the other nine classifiers may be found in Cai and Sowmya (2007).

The geometric level set classifier outperforms each of the other classifiers at least once among the four binary data sets, and is generally competitive overall. Level set classification is also competitive on the multicategory data sets. In fact, it gives the smallest error among all of the classifiers on the segmentation data set. The proposed classifier is competitive for data sets of both small and large dimensionality D; there is no apparent relationship between D and the performance of the geometric level set classifier in comparison to other methods.

6. Conclusion

Level set methods are powerful computational techniques that have not yet been widely adopted in machine learning. Our main goal with this contribution is to open a conduit between the application area of learning and the computational technique of level set methods. Towards that end, we have developed a nonlinear, nonparametric classifier based on level set methods that minimizes margin-based empirical risk in both the binary and multicategory cases, and is regularized by a geometric complexity penalty novel to classification. This approach is an alternative to kernel machines for learning nonlinear decision boundaries in the input space and is in some ways a more natural generalization of linear methods.

The variational level set formulation is flexible in allowing the inclusion of various geometric priors defined in the input space. The surface area regularization term is one such example, but others may also be included. Another example is an energy functional that measures feature relevance using the partial derivative of the signed distance function (Domeniconi et al., 2005), and can be used for ℓ_1-regularized feature subset selection as discussed in Varshney and Willsky (2008).

We have provided an analysis of the classifier by characterizing its ε-entropy. This characterization leads to results on consistency and complexity. We have described a multicategory level set classification procedure with a logarithmic number of decision functions, rather than the linear number that is typical in classification and decision making, through a binary encoding made possible by the level set representation.

It is a known fact that with finite training data, no one classification method is best for all data sets. Performance of classifiers may vary quite a bit depending on the data characteristics because of differing inductive biases. The classifier presented in this paper provides a new option when choosing a classifier. The results on standard data sets indicate that the level set classifier is competitive with other state-of-the-art classifiers. It would be interesting to systematically find domains in the space of data set characteristics for which the geometric level set classifier outperforms other classifiers (Ho and Basu, 2002).

Acknowledgments

This work was supported in part by an NSF Graduate Research Fellowship, by Shell International Exploration and Production, Inc., and by a MURI funded through ARO Grant W911NF-06-1-0076. The authors thank Justin H. G. Dauwels, John W. Fisher, III and Sujay R. Sanghavi for discussions. The level set methods toolbox by Barış Sümengen was used in producing figures in Section 2 and Section 4.

References

Shotaro Akaho. SVM that maximizes the margin in the input space. *Systems and Computers in Japan*, 35(14):78–86, December 2004.

Erin L. Allwein, Robert E. Schapire, and Yoram Singer. Reducing multiclass to binary: A unifying approach to margin classifiers. *Journal of Machine Learning Research*, 1:113–141, December 2000.

Arthur Asuncion and David J. Newman. UCI machine learning repository. Available at http://archive.ics.uci.edu/ml/, 2007.

Peter L. Bartlett and Shahar Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, November 2002.

Peter L. Bartlett, Michael I. Jordan, and Jon D. McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, March 2006.

Gilles Blanchard, Christin Schäfer, Yves Rozenholc, and Klaus-Robert Müller. Optimal dyadic decision trees. *Machine Learning*, 66(2–3):209–241, March 2007.

Erik M. Boczko, Todd R. Young, Minhui Xie, and Di Wu. Comparison of binary classification based on signed distance functions with support vector machines. In *Proceedings of the Ohio Collaborative Conference on Bioinformatics*, Athens, Ohio, June 2006.

Xiongcai Cai and Arcot Sowmya. Level learning set: A novel classifier based on active contour models. In Joost N. Kok, Jacek Koronacki, Ramon Lopez de Mantaras, Stan Matwin, Dunja Mladenič, and Andrzej Skowron, editors, *Proceedings of the 18th European Conference on Machine Learning*, pages 79–90, Warsaw, Poland, 2007.

Vicent Caselles, Ron Kimmel, and Guillermo Sapiro. Geodesic active contours. *International Journal of Computer Vision*, 22(1):61–79, February 1997.

Thomas Cecil, Jianliang Qian, and Stanley Osher. Numerical methods for high dimensional Hamilton–Jacobi equations using radial basis functions. *Journal of Computational Physics*, 196 (1):327–347, May 2004.

Daniel Cremers, Mikael Rousson, and Rachid Deriche. A review of statistical approaches to level set segmentation: Integrating color, texture, motion and shape. *International Journal of Computer Vision*, 72(2):195–215, April 2007.

Michel C. Delfour and Jean-Paul Zolésio. *Shapes and Geometries: Analysis, Differential Calculus, and Optimization*. Society for Industrial and Applied Mathematics, Philadelphia, Pennsylvania, 2001.

Carlotta Domeniconi, Dimitrios Gunopulos, and Jing Peng. Large margin nearest neighbor classifiers. *IEEE Transactions on Neural Networks*, 16(4):899–909, July 2005.

Richard M. Dudley. Metric entropy of some classes of sets with differentiable boundaries. *Journal of Approximation Theory*, 10(3):227–236, March 1974.

Richard M. Dudley. Correction to “metric entropy of some classes of sets with differentiable boundaries”. *Journal of Approximation Theory*, 26(2):192–193, June 1979.

Bradley Efron. The efficiency of logistic regression compared to normal discriminant analysis. *Journal of the American Statistical Association*, 70(352):892–898, December 1975.

Arnaud Gelas, Olivier Bernard, Denis Friboulet, and Rémy Prost. Compactly supported radial basis functions based collocation method for level-set evolution in image segmentation. *IEEE Transactions on Image Processing*, 16(7):1873–1887, July 2007.

Tin Kam Ho and Mitra Basu. Complexity measures of supervised classification problems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(3):289–300, March 2002.

Chih-Wei Hsu and Chih-Jen Lin. A comparison of methods for multiclass support vector machines. *IEEE Transactions on Neural Networks*, 13(2):415–425, March 2002.

Andrey N. Kolmogorov and Vladimir M. Tihomirov. ε-entropy and ε-capacity of sets in functional spaces. *American Mathematical Society Translations Series 2*, 17:277–364, 1961.

Vladimir Koltchinskii. Rademacher penalties and structural risk minimization. *IEEE Transactions on Information Theory*, 47(5):1902–1914, July 2001.

Sanjeev R. Kulkarni. On metric entropy, Vapnik–Chervonenkis dimension, and learnability for a class of distributions. Technical Report P-1910, Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, September 1989.

Yi Lin. A note on margin-based loss functions in classification. *Statistics & Probability Letters*, 68 (1):73–82, June 2004.

Ulrike von Luxburg and Olivier Bousquet. Distance-based classification with Lipschitz functions. *Journal of Machine Learning Research*, 5:669–695, June 2004.

Stanley Osher and Ronald Fedkiw. *Level Set Methods and Dynamic Implicit Surfaces*. Springer, New York, 2003.

Stanley Osher and James A. Sethian. Fronts propagating with curvature-dependent speed: Algorithms based on Hamilton–Jacobi formulations. *Journal of Computational Physics*, 79(1):12–49, November 1988.

Nikos Paragios and Rachid Deriche. Geodesic active regions and level set methods for supervised texture segmentation. *International Journal of Computer Vision*, 46(3):223–247, February 2002.

Georg Pölzlbauer, Thomas Lidy, and Andreas Rauber. Decision manifolds—a supervised learning algorithm based on self-organization. *IEEE Transactions on Neural Networks*, 19(9):1518–1530, September 2008.

Ryan Rifkin and Aldebaro Klautau. In defense of one-vs-all classification. *Journal of Machine Learning Research*, 5:101–141, January 2004.

Frank Rosenblatt. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6):386–408, 1958.

Christophe Samson, Laure Blanc-Féraud, Gilles Aubert, and Josiane Zerubia. A level set model for image classification. *International Journal of Computer Vision*, 40(3):187–197, December 2000.

Bernhard Schölkopf and Alexander J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, Massachusetts, 2002.

Clayton Scott and Robert D. Nowak. Minimax-optimal classification with dyadic decision trees. *IEEE Transactions on Information Theory*, 52(4):1335–1353, April 2006.

Xiaotong Shen and Wing Hung Wong. Convergence rate of sieve estimates. *The Annals of Statistics*, 22(2):580–615, June 1994.

Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, August 2000.

Gregory G. Slabaugh, H. Quynh Dinh, and Gözde B. Unal. A variational approach to the evolution of radial basis functions for image segmentation. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Minneapolis, Minnesota, June 2007.

Ingo Steinwart. Consistency of support vector machines and other regularized kernel classifiers. *IEEE Transactions on Information Theory*, 51(1):128–142, January 2005.

Mark Sussman, Peter Smereka, and Stanley Osher. A level set approach for computing solutions to incompressible two-phase flow. *Journal of Computational Physics*, 114(1):146–159, September 1994.

Arkadiusz Tomczyk. Active hypercontours and contextual classification. In Halina Kwasnicka and Marcin Paprzycki, editors, *Proceedings of the 5th International Conference on Intelligent Systems Design and Applications*, pages 256–261, Wroclaw, Poland, September 2005.

Arkadiusz Tomczyk and Piotr S. Szczepaniak. On the relationship between active contours and contextual classification. In Marek Kurzyński, Edward Puchała, Michał Woźniak, and Andrzej Żołnierek, editors, *Proceedings of the 4th International Conference on Computer Recognition Systems*, pages 303–310, Rydzyna, Poland, 2005.

Arkadiusz Tomczyk and Piotr S. Szczepaniak. Adaptive potential active hypercontours. In Leszek Rutkowski, Ryszard Tadeusiewicz, Lotfi A. Zadeh, and Jacek Zurada, editors, *Proceedings of the 8th International Conference on Artificial Intelligence and Soft Computing*, pages 692–701, Zakopane, Poland, June 2006.

Arkadiusz Tomczyk, Piotr S. Szczepaniak, and Michal Pryczek. Active contours as knowledge discovery methods. In Vincent Corruble, Masayuki Takeda, and Einoshin Suzuki, editors, *Proceedings of the 10th International Conference on Discovery Science*, pages 209–218, Sendai, Japan, 2007.

Aad W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, Cambridge, England, 1998.

Vladimir N. Vapnik. *The Nature of Statistical Learning Theory*. Springer, New York, 1995.

Kush R. Varshney and Alan S. Willsky. Supervised learning of classifiers via level set segmentation. In *Proceedings of the IEEE Workshop on Machine Learning for Signal Processing*, pages 115–120, Cancún, Mexico, October 2008.

Luminita A. Vese and Tony F. Chan. A multiphase level set framework for image segmentation using the Mumford and Shah model. *International Journal of Computer Vision*, 50(3):271–293, December 2002.

Holger Wendland. *Scattered Data Approximation*. Cambridge University Press, Cambridge, England, 2005.

Rebecca Willett and Robert D. Nowak. Minimax optimal level-set estimation. *IEEE Transactions on Image Processing*, 16(12):2965–2979, December 2007.

Robert C. Williamson, Alexander J. Smola, and Bernhard Schölkopf. Generalization performance of regularization networks and support vector machines via entropy numbers of compact operators. *IEEE Transactions on Information Theory*, 47(6):2516–2532, September 2001.

Andy M. Yip, Chris Ding, and Tony F. Chan. Dynamic cluster formation using level set methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(6):877–889, June 2006.

Hui Zou, Ji Zhu, and Trevor Hastie. The margin vector, admissible loss and multi-class margin-based classifiers. Technical report, University of Minnesota, 2006.

Hui Zou, Ji Zhu, and Trevor Hastie. New multicategory boosting algorithms based on multicategory Fisher-consistent losses. *Annals of Applied Statistics*, 2(4):1290–1306, December 2008.