Support Vector Random Fields for Spatial Classification

Chi-Hoon Lee, Russell Greiner, and Mark Schmidt

Department of Computing Science
University of Alberta
Edmonton AB, Canada
{chihoon,greiner,schmidt}@cs.ualberta.ca

Abstract. In this paper we propose Support Vector Random Fields (SVRFs), an extension of Support Vector Machines (SVMs) that explicitly models spatial correlations in multi-dimensional data. SVRFs are derived as Conditional Random Fields that take advantage of the generalization properties of SVMs. We examine computing posterior probability distributions from SVMs, and present a local-consistency potential measure that encourages spatial continuity. SVRFs can be efficiently trained, converge quickly during inference, and can be trivially augmented with kernel functions. SVRFs are more robust to class imbalance than Discriminative Random Fields (DRFs), and are more accurate near edges. Our results on synthetic data and a real-world tumor detection task show the superiority of SVRFs over both SVMs and DRFs.

1 Introduction

The task of classification has traditionally focused on data that is "independent and identically distributed" (iid), in particular assuming that the class labels for different data points are conditionally independent (ie. knowing that one patient has cancer does not mean another one will). However, real-world classification problems often deal with data points whose labels are correlated, and thus the data violates the iid assumption. There is extensive literature focusing on the 1-dimensional 'sequential' case (see [1]), where correlations in the labels of data points in a linear sequence exist, such as in strings, sequences, and language. This paper focuses on the more general 'spatial' case, where these correlations exist in data with two-dimensional (or higher-dimensional) structure, such as in images, volumes, graphs, and video.

Classifiers that make the iid assumption often produce undesirable results when applied to data with spatial dependencies in the labels. For example, in the task of image labeling, a classifier could classify a pixel as 'face', even if all adjacent pixels were classified as 'non-face'. This problem motivates the use of Markov Random Fields (MRFs) and more recently Conditional Random Fields (CRFs) for spatial data. These classification techniques augment the performance of an iid classification technique (often a Mixture Model for MRFs, and Logistic Regression for CRFs) by taking into account spatial class dependencies.

Support Vector Machines (SVMs) are classifiers that have appealing theoretical properties [2], and have shown impressive empirical results in a wide variety of tasks. However, this technique makes the critical iid assumption. This paper proposed an extension to SVMs that considers spatial correlations among data instances (as in Random Field models), while still taking advantage of the powerful discriminative properties of SVMs. We refer to this technique as Support Vector Random Fields (SVRFs)

The remaining sections of this paper are organized as follows. Section 2 formalizes the task and reviews related methods for modeling dependencies in the labels of spatial data. Section 3 reviews Support Vector Machines, and presents our Support Vector Random Field extension. Experimental results on synthetic and real data sets are given in Sect. 4, while a summary of our contribution is presented in Sect. 5.

2 Related Work

The challenge of performing classification while modeling class dependencies is often divided into two perspectives: Generative and Discriminative models [1]. Generative classifiers learn a model of the joint probability, $p(x, y) = p(x|y)p(y)$, of the features x and corresponding labels y. Predictions are made using Bayes rule to compute $p(y|x)$, and finding an assignment of labels maximizing this probability. In contrast, discriminative classifiers model the posterior $p(y|x)$ directly without generating any prior distributions over the classes. Thus, discriminative models solely focus on maximizing the conditional probability of the labels, given the features. For many applications, discriminative classifiers often achieve higher accuracy than generative classifiers [1]. There has been much related work on using random field theory to model class dependencies in generative and more recently discriminative contexts [3, 4]. Hence, we will first review *Markov Random Fields* (typically formulated as a generative classifier), followed by *Conditional Random Fields* (a state-of-the-art discriminative classifier built upon the foundations of Markov Random Fields).

2.1 Problem Formulation

In this work, we will focus on the task of classifying elements (pixels or regions) of a two-dimensional image, although the methods discussed also apply to higher-dimensional data. An image is represented with an M by N matrix of elements. For an instance $X = (x_{11}, x_{12}, \ldots, x_{1N}, \ldots, x_{M1}, x_{M2}, \ldots, x_{MN})$, we seek to infer the most likely joint class labels:

$$Y^* = (y^*_{11}, y^*_{12}, \ldots, y^*_{1N}, \ldots, y^*_{M1}, y^*_{M2}, \ldots, y^*_{MN})$$

Without loss of generality, we will write y_{ij} simply as y_k, where the site index k can range from 1 to $S = MN$. For L class labels, the cardinality of the labeling search space is L^S. If we assume that the labels assigned to elements are independent, the following joint probability can be formulated:

$$P(Y) = \prod_{i=1}^{S} P(y_i).$$

However, conditional independency does not hold for image data, since spatially adjacent elements are likely to receive the same labels. We therefore need to explicitly consider this local dependency. This involves addressing three important issues: How should the optimal solution be defined, how are spatial dependencies considered, and how should we search the (exponential size) configuration space.

2.2 Markov Random Fields (MRFs)

Markov Random Fields (MRFs) provide a mathematical formulation for modeling local dependencies, and are defined as follows [3]:

Definition 1. *A set of random variables Y is called a Markov Random Field on S with respect to a neighborhood N, if and only if the following two conditions are satisfied, where $S - \{i\}$ denotes the set difference, $y_{S-\{i\}}$ denotes random variables in $S-\{i\}$, and N_i denotes the neighboring random variables of random variable i:*

1. $P(Y) > 0$
2. $P(y_i|y_{S-\{i\}}) = P(y_i|y_{N_i})$

Condition 1 is called *Positivity*, which allows the joint probability of any random field to be uniquely determined by its conditional probability. Condition 2 (Markovianity) states that the conditional distribution of an element y_i is dependent only on its neighbors. Markov Random Fields have traditionally sought to maximize the joint probability $P(Y^*)$ (a generative approach). In this formulation, the posterior over the labels given the observations is formulated using Bayes' rule as:

$$P(Y|X) \propto P(X|Y)P(Y) = P(Y)\prod_{i=1}^{S} P(x_i|y_i) \tag{1}$$

In (1), the equivalence between MRFs and Gibbs Distributions [5] provides an efficient way to factor the prior $P(Y)$ over cliques defined in the neighborhood Graph G. The prior $P(Y)$ is written as

$$P(Y) = \frac{\exp(\sum_{c\in C} V_c(Y))}{\sum_{Y'\in\Omega} exp(\sum_{c\in C} V_c(Y'))} \tag{2}$$

where $V_c(Y)$ is a clique potential function of labels for clique $c \in C$, C is a set of cliques in G, and Ω is the space of all possible labelings (consisting of L^{MN} terms). From (1) and (2), the target configuration Y^* is a realization of a locally dependent Markov Random Field with a specified prior distribution. Based on (1) and (2) and using Z to denote the (normalizing) "partition function", if we assume Gaussian likelihoods then the posterior distribution can be factored as:

$$P(Y|X) = \frac{1}{Z}\exp\Big[\sum_{i\in S}\log(P(x_i|y_i)) + \sum_{c\in C} V_c(Y_c)\Big] \tag{3}$$

The Gaussian assumption for $P(X|Y)$ in (1) allows straightforward Maximum Likelihood parameter estimation. Although there have been many approximation algorithms designed to find the optimal Y^*, we will focus on a local method called *Iterated Conditional Modes* [5], written as:

$$y_i^* = \arg\max_{y_i \in L} P(y_i|y_{N_i} x_i) \tag{4}$$

Assuming Gaussians for the likelihood and a pairwise neighborhood system for the prior over labels, (4) can be restated as:

$$y_i^* = \arg\max_{y_i \in L} \frac{1}{Z_i} \exp\Big[\log(P(x_i|y_i)) + \sum_{j \in N_i} \beta y_i y_j\Big] \tag{5}$$

where β is a constant and L is a set of class labels.

This concept has proved to be applicable in a wide variety of domains where there exists correlations among neighboring instances. However, the generative nature of the model and the assumption that the likelihood is Gaussian can be too restrictive to capture complex dependencies between neighboring elements or between observations and labels. In addition, the prior over labels is completely independent from the observations, thus the interactions between neighbors are not proportional to their similarity.

2.3 Conditional Random Fields (CRFs)

CRFs avoid the Gaussian assumption by using a model that seeks to maximize the conditional probability of the labels given the observations $P(Y^*|X)$ (a discriminative model), and are defined as follows [1]:

Definition 2. *Let $G = (S, E)$ be a graph such that Y is indexed by the vertices S of G. Then (X, Y) is said to be a CRF if, when conditioned on Y, the random variables y_i obey the Markov property with respect to the graph: $P(y_i|X, y_{s\setminus i}) = P(y_i|X, y_{N_i})$.*

This model alleviates the need to model the observations $P(X)$, allowing the use of arbitrary attributes of the observations without explicitly modeling them. CRFs assume a 1-dimensional chain-structure where only adjacent elements are neighbors. This allows the factorization of the joint probability over labels. Discriminative Random Fields (DRFs) extend 1-dimensional CRFs to 2-dimensional structures [6]. The conditional probability of the labels Y in the Discriminative Random Field framework is defined as:

$$P(Y|X) = \frac{1}{Z} \exp\Big(\sum_{i \in S} A_i(y_i, X) + \sum_{i \in S} \sum_{j \in N_i} I_{ij}(y_i, y_{j,X})\Big) \tag{6}$$

A_i is the 'Association' potential that models dependencies between the observations and the class labels, while I_i is the 'Interaction' potential that models dependencies between the labels of neighboring elements (and the observations).

Note that this is a much more powerful model than the assumed Gaussian Association potential and the indicator function used for the Interaction potential (that doesn't consider the observations) in MRFs. Parameter learning in DRFs involves maximizing the log likelihood of (6), while inference uses ICM [6].

DRFs are a powerful method for modeling dependencies in spatial data. However, several problems associated with this method include the fact that it is hard to find a good initial labeling and stopping criteria during inference, and it is sensitive to issues of class imbalance. Furthermore, for some real-world tasks the use of logistic regression as a discriminative method in DRFs often does not produce results that are as accurate as powerful classification models such as Support Vector Machines (that make the iid assumption).

3 Support Vector Random Fields (SVRFs)

This section presents Support Vector Random Fields (SVRFs), our extension of SVMs that allows the modelling of non-trivial two-dimensional (or higher) spatial dependencies using a CRF framework. This model has two major components: The *observation-matching* potential function and the *local-consistency* potential function. The *observation-matching* function captures relationships between the observations and the class labels, while the *local-consistency* function models relationships between the labels of neighboring data points and the observations at data points. Since the selection of the observation-matching potential is critical to the performance of the model, the Support Vector Random Field model employs SVMs for this potential, providing a theoretical and empirical advantage over the logistic model used in DRFs and the Gaussian model used in MRFs, that produce unsatisfactory results for many tasks. SVRFs can be formulated as follows:

$$P(Y|X) = \frac{1}{Z} \exp\Big\{ \sum_{i \in S} \log(O(y_i, \Upsilon_i(X))) + \sum_{i \in S} \sum_{j \in N_i} V(y_i, y_j, X) \Big\} \qquad (7)$$

In this formulation, $\Upsilon_i(X)$ is a function that computes features from the observations X for location i, $O(y_i, \Upsilon_i(X))$ is the observation-potential, and $V(y_i, y_j, X)$ is the local-consistency potential. The pair-wise neighborhood system is defined as a local dependency structure. In this work, interactions between pixels with a Euclidean distance of 1 were considered (the von Neumann neighborhood of radius 1). We will now examine these potentials in more detail.

3.1 Observation-Matching

The observation-matching potential seeks to find a posterior probability distribution that maps from the observations to corresponding class labels. DRFs employ a Generalized Linear Models (GLM) for this potential. However, GLMs often do not estimate appropriate parameters. This is especially true in image data where feature sets may have a high number of dimensions and/or several

features have a high degree of correlation. This can cause problems in parameter estimation and approximations to resolve these issues may not produce optimal parameters [7].

Fortunately, the CRF framework allows a flexible choice of the observation-matching potential function. We overcome the disadvantages of the GLM by employing a Support Vector Machine classifier, seeking to find the margin maximizing hyperplane between the classes. This classifier has appealing properties in high-dimensional spaces and is less sensitive to class imbalance. Furthermore, due to the properties of error bounds, SVMs tends to outperform GLMs, especially when the classes overlap in the feature space (often the case with image data). Parameter estimation for SVMs involves optimizing the following Quadratic Programming problem for the training data x_i (where C is a constant that bounds the misclassification error):

$$\max \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i}^{N} \sum_{j}^{N} \alpha_i \alpha_j y_i y_j x_i^T x_j$$

$$subject \;\; to \;\; 0 \leq \alpha_i \leq C \; and \; \sum_{i=1}^{N} \alpha_i y_i = 0 \tag{8}$$

Consequently, the decision function, given the parameters α_i for the l training instances and bias term b, is (for a more thorough discussion of SVMs, we refer to [2]):

$$f(x) = \sum_{i=1}^{l} (\alpha_i y_i x \cdot x_i) + b \tag{9}$$

Unfortunately, the decision function $f(x)$ produced by SVMs measures distances to the decision boundary, while we require a posterior probability function. We adopted the approach of [8] to convert the decision function to a posterior probability function. This approach is efficient and minimizes the risk of overfitting during the conversion, but has some ambiguities and potential difficulties in numerical computation. We have addressed these issues in our approach, which will be briefly outlined here.

We estimate a posterior probability from the Support Vector Machine decision function using the sigmoid function:

$$O(y_i = 1, \Upsilon_i(X)) = \frac{1}{1 + \exp(Af(\Upsilon_i(X)) + B)} \tag{10}$$

The parameters A and B are estimated from training data represented as pairs $(f(\Upsilon_i(X)), t_i)$, where $f(\cdot)$ is the Support Vector Machine decision function, and t_i denotes a relaxed probability that $y_i = 1$ as in (10). We could set $t_i = 1$, if the class label at i is 1(ie. $y_i = 1$). However, this ignores the possibility that $\Upsilon_i(X)$ has the opposite class label (ie. -1). Thus, we employed the relaxed probability: $t_i = \frac{N_+ + 1}{N_+ + 2}$, if $y_i = 1$, and $t_i = \frac{1}{N_- + 2}$, if $y_i = -1$ (N_+ and N_- being the number of positive and negative class instances). By producing the new forms of training instances, we can solve the following optimization problem to estimate parameters:

$$\min -\sum_{i=1}^{l}\Big[t_i \log p(\Upsilon_i(X)) + (1-t_i)\log(1-p(\Upsilon_i(X)))\Big] \tag{11}$$

where

$$p(\Upsilon_i(X)) = \frac{1}{1+exp(Af(\Upsilon_i(X))+B)}$$

[8] adopted a Levenberg-Marquardt approach to solve the optimization problem, finding an approximation of the Hessian matrix. However, this may cause incorrect computations of the Hessian matrix (especially for unconstrained optimizations [7]). Hence, as in [9], we employed Newton's method with backtracking line search to solve the optimization.

In [8], the measure of goodness of fit is computed from the last training instances in conjunction with (11). As opposed to [8], our measure assesses goodness of fit by incorporating all training instances. In addition, similar to [9], in order to avoid overflows and underflows of *exp* and *log* functions, we reformulate (11) as follows (see Appendix 1):

$$-\Big(t_i \log p(\Upsilon_i(X)) + (1-t_i)\log(1-p(\Upsilon_i(X)))\Big)$$

$$= t_i(Af(\Upsilon_i(X))+B) + \log(1+\exp(-Af(\Upsilon_i(X))-B)) \tag{12}$$

3.2 Local-Consistency

In MRFs, local-consistency considers correlations between neighboring data points, and is considered to be observation independent. CRFs provide more powerful modelling of local-consistency by removing the assumption of observation independence. In order to use the principles of CRFs for local-consistency, an approach is needed that penalizes discontinuity between pairwise sites. For this, we use a linear function of pairwise continuity:

$$V(y_i, y_j, X) = y_i y_j \nu^T \Phi_{ij}(X) \tag{13}$$

$\Phi_{ij}(X)$ is a function that computes features for sites i and j based on observations X. As opposed to DRFs, which penalize discontinuity by considering the absolute difference between pairwise observations [6], our approach introduces a new mapping function $\Phi(\cdot)$ that encourages continuity in addition to penalizing discontinuity (using $\max(\Upsilon(X))$ to denote the vector of max values for each feature):

$$\Phi_{ij}(X) = \frac{\max(\Upsilon(X)) - \mid \Upsilon_i(X) - \Upsilon_j(X) \mid}{\max(\Upsilon(X))} \tag{14}$$

3.3 Learning and Inference

The proposed model needs to estimate the parameters of the observation-matching function and the local-consistency function. Although we estimate these parameters sequentially, our model outperforms the simultaneous learning approach of DRFs and significantly increases its computational efficiency.

The parameters of the Support Vector Machine decision function are first estimated by solving the Quadratic Programming problem in (8) (using SVM-light [10]). We then convert the decision function to a posterior function using (11) and the new training instances. Finally, we adopted pseudolikelihood [3] to estimate the local consistency parameters ν, due to its simplicity and fast computation. For training on l pixels from K images, pseudolikehood is formulated as:

$$\widehat{\nu} = \arg\max_{\nu} \prod_{k=1}^{K} \prod_{i=1}^{l} P(y_i^k | y_{N_i}^k, X^k, \nu) \tag{15}$$

As in [6], to ensure that the log-likelihood is convex and to prevent over-smoothing due to the pseudolikelihood approximation we assume a Gaussian prior on ν and compute the local-consistency parameters using its penalized log likelihood $l(\widehat{\nu})$:

$$l(\widehat{\nu}) = \arg\max_{\nu} \sum_{k=1}^{K} \sum_{i=1}^{l} \left\{ O_i^n + \sum_{j \in N_i} V(y_i^k, y_j^k, X^k) - \log(z_i^k) \right\} - \frac{1}{2\tau} \nu^T \nu \tag{16}$$

In this model, z_i^k is a partition function for each site i in image k, and τ is a regularizing constant. We optimize (16) using gradient descent, and note that the observation matching function acts as a constant during this process. Due to the employment of SVMs, the time complexity of learning is $O(S^2)$, where S is the number of pixels to be trained, although in practice it is much faster.

The inference problem is to infer an optimal labeling Y^* given a new instance X and the estimated model parameters. We herein adopted the Iterated Conditional Modes (ICM) approach described in Section 2.2, that maximizes the local conditional probability iteratively. For our proposed model and [6], ICM is expressed as,

$$y_i^* = \arg\max_{y_i \in L} P(y_i | y_{N_i}, X) \tag{17}$$

Although ICM is based on iterative principles, it often converges quickly to a high quality configuration, and each iteration has time complexity $O(S)$.

We close by noting that the M^3N [11] framework resembles SVRFs at a high-level, since it also incorporates label dependencies and uses a max-margin approach. One of the key differences is that the M^3N approach uses a margin that magnifies the difference between the target labels and the best runner-up, while we use the 'traditional' binary SVM approach of maximizing the distance from the classes to a separating hyperplane. An efficient approach for training and inference in a special case of M^3Ns was presented in [12]. The restrictive associative edge potentials used in this special case allow M^3Ns to be tractably applied to spatial data. In contrast, SVRFs do not require the restrictive associative assumption, and training SVRFs has a much lower computational cost.

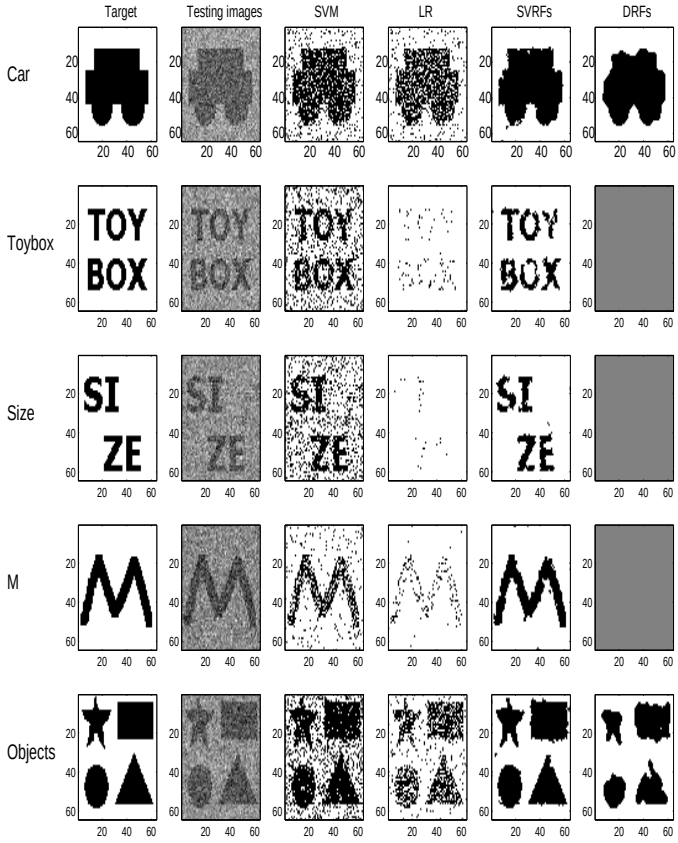


Fig. 1. Example Data and Results for the Different Classifiers

4 Experiments

We have evaluated our proposed model on synthetic and real-world binary image labeling tasks, comparing our approach to Logistic Regression, SVMs, and DRFs for these problems. Since class imbalance was present in many of the data sets, we used the Jaccard measure to quantify performance: $f = \frac{TP}{TP+FP+FN}$, where TP is the number of true positives, FP denotes the number of false positives, and FN tallies false negatives.

4.1 Experiments on Synthetic data

We evaluated the four techniques over 5 synthetic binary image sets. These binary images were corrupted by zero mean Gaussian noise with unit standard deviation, and the task was to label the foreground objects (see the first and second columns in Fig. 2). Two of the sets contained balanced class labels (*Car* and *Objects*), while the other three contained imbalanced classes. The five 150

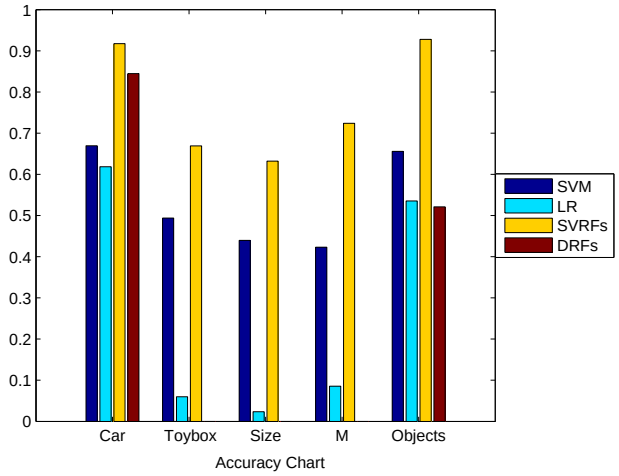


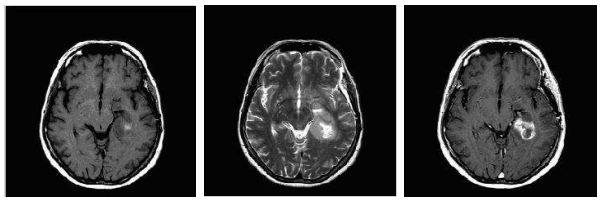
Fig. 2. Average scores on synthetic data sets

image sets were divided into 100 images for training and 50 for testing. Example results and aggregate scores are shown in Fig. 2. Note that the last 4 columns illustrate the outcomes from each technique– SVMs, Logistic Regression (LR), SVRFs, and DRFs.

Logistic Regression and subsequently DRFs performed poorly in all three imbalanced data sets (*Toybox*, *Size*, and *M*). In these cases, SVMs outperformed these methods and consequently our proposed SVRFs outperformed SVMs. In the first balanced data set (*Car*), DRFs and SVRFs both significantly outperformed SVMs and Logistic Regression (the iid classifiers). However, DRFs performed poorly on the second balanced data set (*Objects*). This is due to DRFs simultaneous parameter learning, that tends to overestimate the local-consistency potential. Since the observation-matching is underweighted, edges become degraded during inference (there are more edge areas in the *Objects* data). Terminating inference before convergence could reduce this, but this is not highly desirable for automatic classification. Overall, our Support Vector Random Field model demonstrated the best performance on all data sets, in particular those with imbalanced data and a greater proportion of edge areas.

4.2 Experiments on Real data

We applied our model to the real-world problem of tumor segmentation in medical imaging. We focused on the task of brain tumor segmentation in MRI, an important task in surgical planning and radiation therapy currently being laboriously done by human medical experts. There has been significant research focusing on automating this challenging task (see [13]). Markov Random Fields have been explored previously for this task (see [13]), but recently SVMs have



(a) T1 (b) T2 (c) T1-Contrast

Fig. 3. A multi-spectral MRI

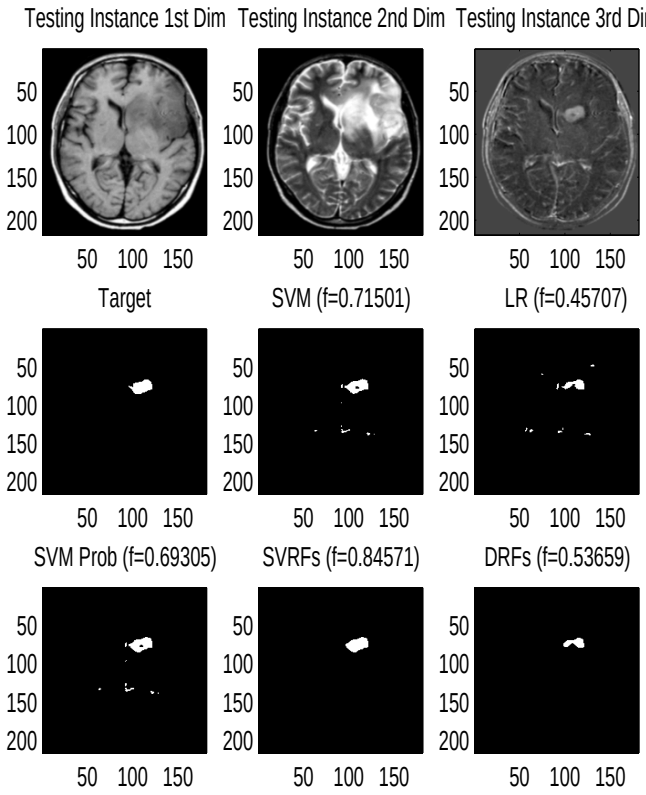


Fig. 4. Example 1 of the classification results

shown impressive performance [14, 15]. This represents a scenario where our proposed Support Vector Random Field model could have a major impact.

We evaluated the four classifiers from the previous section over 7 brain tumor patients. For each patient, three MRI 'modalities' were available: T1 (visualizing fat locations), T2 (visualizing water locations), and an additional T1 image

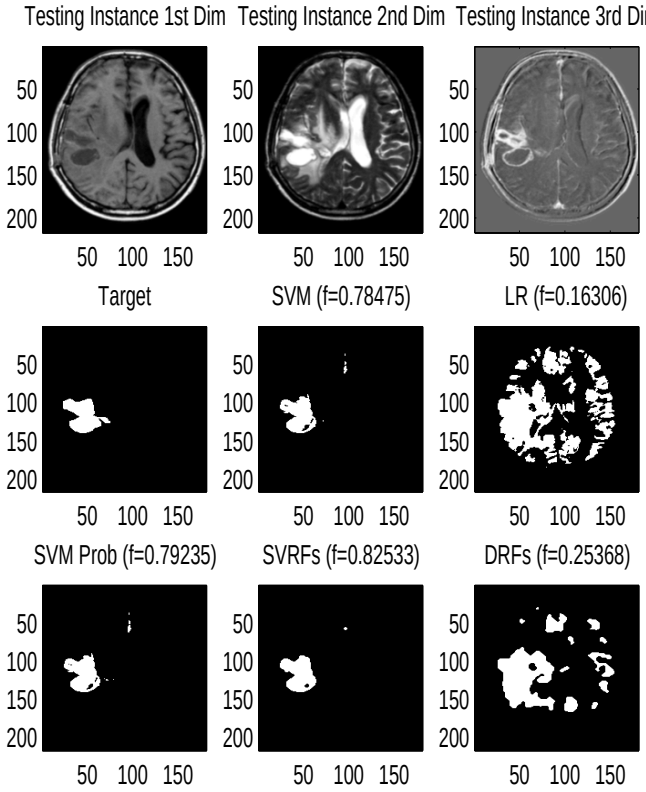


Fig. 5. Example 2 of the classification result

with a 'contrast agent' added to enhance the visualization of metabolically active tumor areas (refer to Fig. 3). The data was preprocessed with the Statistical Parametric Mapping software [16] to non-linearly align the images with a template in a standard coordinate system, and remove intensity inhomogeneity field effects. This non-linear template alignment approach was quantified to be highly effective in [17], and the inhomogeneity correction step computes a smooth corrective field that seeks to minimize the residual entropy after transformation of the log-intensity value's probability distribution [18]. We used 12 features that incorporate image information and domain knowledge (the raw intensities, spatial expected intensities within the coordinate system, spatial priors for the brain area and normal tissue types within the coordinate system, the template image information, and left-to-right symmetry), each measured as features at 3 scales by using 3 different sizes of Gaussian kernel filters. We used a 'patient-specific' training scenario similar to [14, 15].

Results for two of the patients are shown in Fig. 5, while average scores over the 7 patients are shown in Fig. 4.2. Note that 'SVM+prob' in Fig. 5

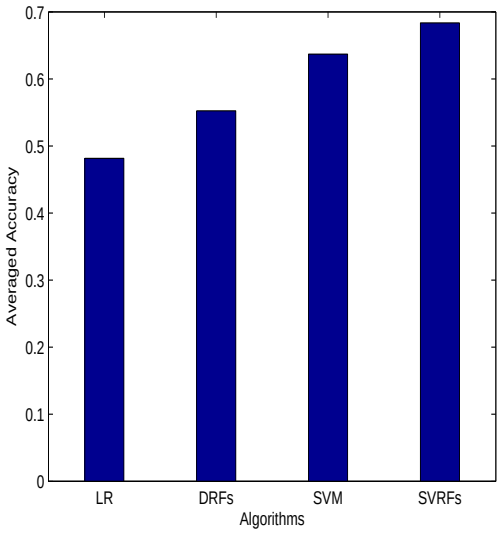


Fig. 6. Average Accuracy

denotes the classification results from the Support Vector Machine posterior probability estimate. The Logistic Regression model performs poorly at this task, but DRFs perform significantly better. As with the synthetic data in cases of class imbalance, SVMs outperform both Logistic Regression and the DRFs. Finally, SVRFs improve the scores obtained by the SVMs by almost 5% (a significant improvement).

We compared convergence of the DRFs and SVRFs by measuring how many label changes occured between inference iterations averaged over 21 trials (Fig. 4.2). These results show that DRFs on average require almost 3 times as many iterations to converge, due to the overestimation of the local-consistency potential.

5 Conclusion

We have proposed a novel model for classification of data with spatial dependencies. The Support Vector Random Field combines ideas from SVMs and CRFs, and outperforms SVMs and DRFs on both synthetic data sets and an important real-world application. We also examined improvements to computing posterior probability distributions from SVM decision functions, and a method to encourage continuity with local-consistency potentials. Our Support Vector Random Field model is robust to class imbalance, can be efficiently trained, converges quickly during inference, and can trivially be augmented with kernel functions to further improve results.

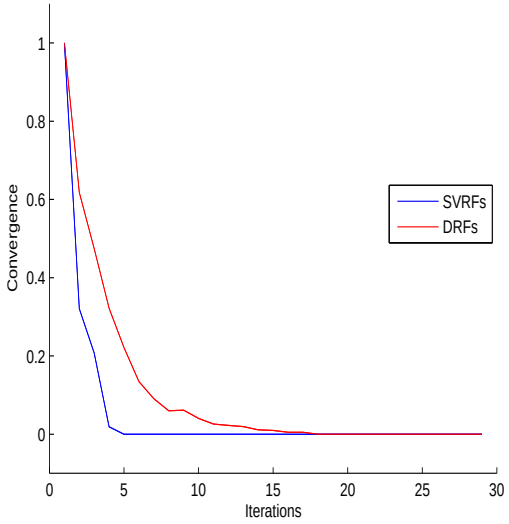


Fig. 7. Comparison of convergence in inference

Appendix

This section illustrates the derivations of (12) to improve the numerical issues that occur when programming with *exp* and *log* functions.

$$Let \quad a = Ax + B, \quad p = \frac{1}{1+e^a}, \quad 1-p = \frac{e^a}{1+e^a}$$

$$\begin{aligned}
-[t\log p + (1-t)\log(1-p)] &= -t\log p - log(1-p) + t\log(1-p) \\
&= -t\log(1+e^a)^{-1} - \log(\frac{e^a}{1+e^a}) + t\log(\frac{e^a}{1+e^a}) \\
&= t\log(1+e^a) - \log e^a + \log(1+e^a) + t\log(e^a) - t\log(1+e^a) \\
&= ta + \log(1+e^a) - \log e^a \\
&= ta + \log(\frac{1+e^a}{e^a}) \\
&= ta + \log(e^{-a}+1)
\end{aligned}$$

Acknowledgment

R. Greiner is supported by the National Science and Engineering Research Council of Canada (NSERC) and the Alberta Ingenuity Centre for Machine Learning (AICML). C.H. Lee is supported by NSERC, AICML, and iCORE. Our thanks to Dale Schuurmans for helpful discussions on optimization and parameter estimation, J. Sander for hlepful discussions for the classification issues, BTGP members for help in data processing, and Albert Murtha (M.D.) for domain knowledge on the tumor data set.

References

1. Lafferty, J., Pereira, F., McCallum, A.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. ICML (2001)
2. Shawe-Taylor, Cristianini: Kernel Methods for Pattern Analysis. Cambridge University Press, Cambridge, UK (2004)
3. Li, S.Z.: Markov Random Field Modeling in Image Analysis. Springer-Verlag, Tokyo (2001)
4. Kumar, S., Hebert, M.: Discriminative random fields: A discriminative framework for contextual interaction in classification. ICCV (2003) 1150–1157
5. Besag, J.: On the statistical analysis of dirty pictures. Journal of Royal Statistical Society. Series B **48** (1986) 3:259–302
6. Kumar, S., Hebert, M.: Discriminative fields for modeling spatial dependencies in natural images. NIPS (2003)
7. R.Fletcher: Practical Methods of Optimization. John Wiley & Sons (1987)
8. Platt, J.: Probabilistic outputs for support vector machines and comparison to regularized likelihood methods. MIT Press, Cambridge, MA (2000)
9. Lin, H.T., Lin, C.J., Weng, R.: A note on platt's probabilistic outputs for support vector machines. Technical report, National Taiwan University (2003)
10. Joachims, T.: Making large-scale svm learning practical. In Scholkopf, B., Burges, C., Smola, A., eds.: Advances in Kernel Methods - Support Vector Learning, MIT Press (1999)
11. Taskar, B., Guestrin, C., Koller, D.: Max margin markov networks. NIPS (2003)
12. Anguelov, D., Taskar, B., Chatalbashev, V., Koller, D., Gupta, D., Heitz, G., Ng, A.: Discriminative learning of markov random fields for segmentation of 3d scan data. CVPR (2005)
13. Gering, D.: Recognizing Deviations from Normalcy for Brain Tumor Segmentation. PhD thesis, MIT (2003)
14. Zhang, J., Ma, K., Er, M., Chong, V.: Tumor segmentation from magnetic resonance imaging by learning via one-class support vector machine. Int. Workshop on Advanced Image Technology (2004) 207–211
15. Garcia, C., Moreno, J.: Kernel based method for segmentation and modeling of magnetic resonance images. LNCS **3315** (2004) 636–645
16. : Statistical parametric mapping, http://www.fil.ion.bpmf.ac.uk/spm/ (Online)
17. Hellier, P., Ashburner, J., Corouge, I., Barillot, C., Friston, K.: Inter subject registration of functional and anatomical data using spm. In: MICCAI. Volume 587-590. (2002)
18. Ashburner, J.: Another mri bias correction approach. In: 8th Int. Conf. on Functional Mapping of the Human Brain, Sendai, Japan. (2002)