Locally Optimized Kernels

Tomasz Maszczyk and Włodzisław Duch

Nicolaus Copernicus University

Toruń, Poland

{tmaszczyk, wduch}@is.umk.pl

Introduction

The type of solution offered by a given data model obtained by SVM with a specific kernel [1] may not be the most appropriate for particular data. Each data model defines a hypotheses space, that is a set of functions that this model may easily learn. Linear methods work best when decision borders are flat, but they are obviously not suitable for spherical distributions of data. For some problems (for example, high-dimensional parity and similar functions), neither linear nor radial decision borders are sufficient.

Kernel methods implicitly provide new, useful features $z_i(\vec{x}) = k(\vec{x}, \vec{x}_i)$ constructed around support vectors $\vec{x}_i$, a subset of input vectors relevant to the training objective. Prediction is supported by new features, most often distance functions from selected training vectors, weighted by a Gaussian function, making the decision borders flat in the kernel space. Multiple kernels may be used to construct new features, as shown in our Support Feature Machine algorithm [2]. Linear models defined in the enhanced space are equivalent to kernel-based SVMs. In particular, one can use linear SVM to find discriminant in the enhanced space, preserving the wide margins.

Various kernels may be used to create enhanced feature space. Here an approach that combines Gaussian kernels with selection of pure clusters, adopted from our Almost Random Projection Machines [3] algorithm, is investigated.

SFM vs SVM

The advantages of explicitly defined feature space:

- different types of features
- multiresolution
- feature selection
- different set of features for each class
- easier interpretation
- other methods than LDA

References

1. Boser, B.E., Guyon, I.M., Vapnik, V.N.: A training algorithm for optimal margin classifiers. In: 5th Annual ACM Workshop on COLT, 144–152, Pittsburgh, ACM Press. (1992)
2. Maszczyk, T., Duch, W.: Support feature machines: Support vectors are not enough. In: World Congress on Computational Intelligence, 3852-3859, IEEE Press (2010)
3. Duch, W., Maszczyk, T.: Almost random projection machine. Lecture Notes in Computer Science 5768, (2009)

Locally Optimized Kernels

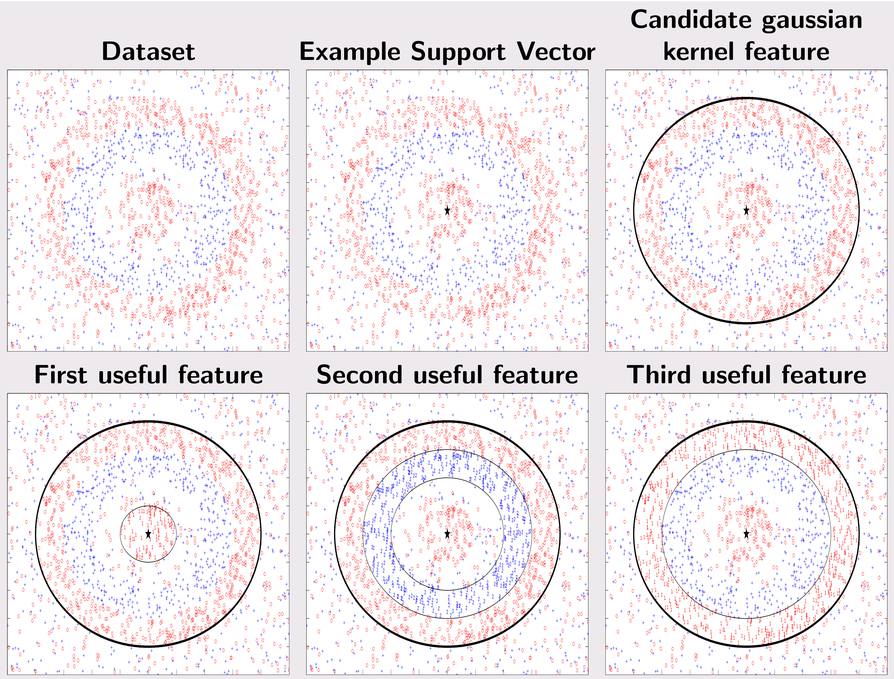
The Locally Optimized Kernels algorithm is based on generation of new features using restricted gaussian kernels followed by the Winner Takes All (WTA) mechanism or by the linear discrimination (LDA).

Algorithm: Locally Optimized Kernels

Require: Fix the values of internal parameters: η (minimum covering) and σ (dispersion).

1: Standardize the dataset, m vectors, d features.
2: Create candidate kernel features $g_i(\vec{x}) = \exp(-||\vec{x}_i - \vec{x}||^2/2\sigma^2)$.
3: Sort $g_i(\vec{x})$ values in descending order, with associated class labels.
4: Analyze $p(g_i|C)$ distribution to find all intervals with pure clusters defining binary features $B_{iab}(\vec{x}; C)$.
5: **if** the number of vectors covered by the feature $B_{iab}(\vec{x}; C) > \eta$ **then**
6: accept this binary feature creating class-labeled hidden network node.
7: **end if**
8: Classify test data mapped into the enhanced space:
9: Sum the activity of hidden node subsets for each class to calculate network outputs (WTA).
10: Build linear model on the enhanced feature space (LDA).

Clean clusters are found either in the local neighborhood of the support vector in the interval $[0, b]$, or if the support vector is surrounded by vectors from another class they may be quite far, with large values of both $a < b$ using interval $[a, b]$.



Results

10x10CV accuracy results

Dataset	SVM(L)	SVM(G)	LOKWTA	LOKLDA
arrhythmia	**50.92±17.31**	43.36±21.47	42.00±24.19	39.10±12.98
autos	54.48±13.75	74.29±12.58	58.69±11.03	**74.36±10.40**
balance-scale	84.47±3.17	89.83±2.09	90.71±2.38	**96.46±2.62**
breast-cancer	73.27±6.10	75.67±5.35	**76.58±6.37**	75.09±1.99
breast-w	96.60±2.07	96.77±1.84	96.93±1.62	**97.21±2.13**
car	67.99±2.61	**98.90±0.90**	84.72±3.44	93.57±1.81
cmc	19.14±2.14	34.09±3.67	48.54±2.52	**51.06±4.30**
credit-a	**86.36±2.86**	86.21±2.90	82.67±4.01	84.70±4.91
credit-g	73.95±4.69	**74.72±4.03**	73.10±2.38	72.70±3.86
cylinder-bands	74.58±5.23	76.89±7.57	74.32±6.41	**80.11±7.53**
dermatology	94.01±3.54	94.49±3.88	87.97±5.64	**94.71±3.02**
diabetes	76.88±4.94	76.41±4.22	74.88±3.88	**76.95±4.47**
ecoli	78.48±5.90	84.17±5.82	82.47±3.66	**85.66±5.40**
glass	42.61±10.05	62.43±8.70	64.96±7.72	**71.08±8.13**
haberman	72.54±1.96	72.91±5.93	**76.46±4.34**	73.53±0.72
heart-c	**82.62±6.36**	80.67±7.96	81.07±7.56	81.04±5.17
heart-statlog	**83.48±7.17**	83.40±6.56	81.48±8.73	83.33±7.46
hepatitis	83.25±11.54	84.87±11.98	**89.88±10.14**	84.05±4.40
ionosphere	87.72±4.63	94.61±3.68	85.18±6.28	**95.16±2.72**
iris	72.20±7.59	**94.86±5.75**	94.67±6.89	93.33±5.46
kr-vs-kp	96.03±0.86	**99.35±0.42**	83.73±2.58	98.25±0.45
liver-disorders	68.46±7.36	**70.30±7.90**	57.40±5.72	69.72±6.57
lymph	81.26±9.79	**83.61±9.82**	76.96±13.07	80.52±7.91
sonar	73.71±9.62	86.42±7.65	**86.57±7.01**	86.52±8.39
vote	96.12±3.85	**96.89±3.11**	92.57±7.52	93.95±4.18
vowel	23.73±3.13	**98.05±1.90**	92.49±3.37	97.58±1.52
zoo	91.61±6.67	93.27±7.53	88.47±5.35	**94.07±6.97**

Conclusions

- Locally Optimized Kernels algorithm is focused on generation of new useful kernel features.
- LOK approach despite its simplicity generates enhanced feature space (using locally optimized gaussian kernels) that improves construction of the original SVM kernel space.
- Methods that extract information from data, making the linear discrimination simpler and more accurate.
- In the case of gaussian kernel features there is no reason why the same dispersion should be used for all support vectors. Those support vectors that are far from decision border on the correct side should have large dispersions, and vectors closer to decision borders should have smaller dispersions.