ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

Learning Kernel Parameters for Support Vector Classification Using Similarity Embeddings

Murilo V. F. Menezes[1], Luiz C. B. Torres[2] and Antonio P. Braga[1] *

1-Department of Electronics Engineering, Federal University of Minas Gerais
Av. Antônio Carlos, 6627, Pampulha, 30161-970, Belo Horizonte, MG, Brazil.

2- Department of Computing and Systems, Federal University of Ouro Preto
35931-008, João Monlevade, MG, Brazil.

Abstract. In order to solve non-linear problems, kernel-based classifiers rely on implicit mappings to very high-dimensional spaces. These target spaces, although mathematically robust, often lack the property of visual interpretation, limiting the intuition of the problem at hand. In this work, the notion of a similarity space is presented, to which one can map input samples and visualize how they interact under a given kernel function. By exploring statistics in such space, a class separability measure is derived, which can be used to find optimal kernel parameters for binary classification. Experiments using support vector machines were conducted, showing the method's effectiveness when compared to grid-search approaches.

1 Introduction

Implicit mapping enabled the formalization of kernel methods, particularly kernel-based large margin classifiers, like SVMs [1]. Explicit mapping, however, may provide an intuitive notion of the problem and uncover properties that are not easily reachable through the implicit perspective. Similarities in relation to each class, considering all samples in the learning set, can be calculated directly from kernel matrices and represented as mappings into a similarity space. For binary classification problems, visualization of such kernel representations, which are in fact row or column summations of the kernel matrix, may add up intuition into the learning problem.

In a previous work [2], properties of kernel affinity mappings were explored in order to develop an approach for learning kernel parameters of SVMs. This resulted in the description of a function between samples from opposite classes whose maximum, obtained by tuning kernel parameters, results on models that are good approximations of the generator functions. The fundamental principle of this method is that it is based solely on the properties of the projections, does not require validation sets and does not rely on extensive computational methods to be accomplished.

Other functions that can be adopted as separability measures in the projected kernel similarity space can also be found in the literature. An example is MMD [3, 4], which considers distances between functions directly in the reproducing kernel

*The authors would like to thank CAPES, CNPq, and FAPEMIG for the support given to this work.

ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

Hilbert space. In the present work, an extension of [2] is presented, with further formalization by considering probability distributions in input space. Kernel similarities between samples are considered in order to build representations in a two-dimensional space, which embeds both global and local relations in very high-dimensional kernel spaces. Mean statistics are adopted to set Gaussian and Laplacian kernels prior to applying them in classification.

2 Kernel mapping into a similarity space

Kernels are essentially pairwise similarity measures that can be considered to assess spatial relations of data and to infer density functions [5, 6]. By combining pairwise kernel values, one can derive similarity measures between sets and distributions. The **kernel similarity** $\psi_{k,\theta}(\mathbf{x}, P)$ of a sample $\mathbf{x} \in \mathcal{X}$ to a probability distribution P over $\mathcal{X}$, given a kernel k parameterized by θ, is given by:

$$\psi_{k,\theta}(\mathbf{x}, P) = \mathbb{E}_{\mathbf{x}' \sim P}\left[k_\theta(\mathbf{x}, \mathbf{x}')\right]. \tag{1}$$

For a binary classification problem with classes C_1 and C_2, for example, considering that $P_1 = P(\mathbf{x}|C_1)$ and $P_2 = P(\mathbf{x}|C_2)$, kernel similarities $\psi_{k,\theta}(\mathbf{x}, P_1)$ and $\psi_{k,\theta}(\mathbf{x}, P_2)$ are measures of how $\mathbf{x}$ is likely to belong to C_1 or C_2. This is basically the notion of likelihood, however, since there is no normalization for unitary integration of densities in this case, in this paper it will be referred to as a similarity measure between $\mathbf{x}$ and P. For a training set $\mathbb{X} = \{\mathbf{x}_1, \cdots, \mathbf{x}_N\}$ sampled from P, kernel similarity can be empirically estimated as:

$$\hat{\psi}_{k,\theta}(\mathbf{x}, \mathbb{X}) = \begin{cases} \frac{1}{N-1} \sum_{\substack{i=1 \\ \mathbf{x}_i \neq \mathbf{x}}}^{N} k_\theta(\mathbf{x}, \mathbf{x}_i) & \text{if } \mathbf{x} \in \mathbb{X} \\ \frac{1}{N} \sum_{i=1}^{N} k_\theta(\mathbf{x}, \mathbf{x}_i) & \text{otherwise} \end{cases} \tag{2}$$

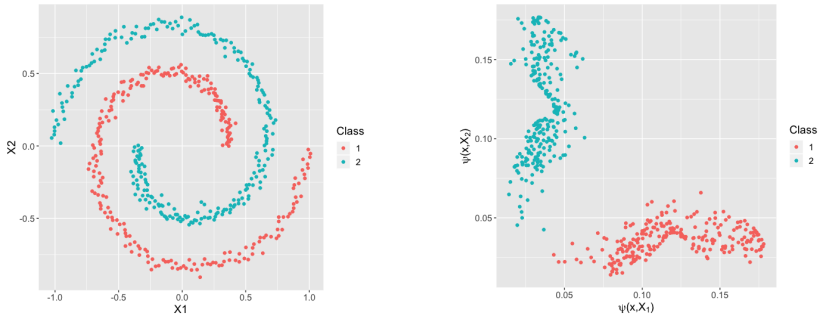
The average of kernel values $k(\mathbf{x}, \mathbf{x}')$ between $\mathbf{x}$ and all samples $\mathbf{x}'$ in a training set $\mathbb{X}$ is, then, a measure of similarity between $\mathbf{x}$ and $\mathbb{X}$. In binary classification problems, kernel similarities are estimated for the two conditional distributions P_1 and P_2 and can, therefore, be mapped into the two dimensional space representing $P_1 \times P_2$ by an operator Ψ as defined next:

$$\{\Psi_{k,\theta} : \mathcal{X} \to \mathbb{R}^2 \mid \Psi_{k,\theta}(x) = \left[\psi_{k,\theta}(\mathbf{x}, P_1), \psi_{k,\theta}(\mathbf{x}, P_2)\right]^T.\} \tag{3}$$

where Ψ is defined for a given kernel k with parameters θ.

As an illustrative example of such a two-dimensional mapping with a Gaussian kernel $k_\sigma(x, x') = \exp -\frac{1}{2}\frac{||x-x'||^2}{\sigma^2}$ with given width σ, consider Figure 1, where class samples are presented in both input and similarity spaces. In this example, the projection with kernel mapping was capable to linearize the problem, as can be observed in the figure. Such an explicit representation of the mapped samples provides intuition into the problem and paves the way to explore pattern interactions in the similarities space.

ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.



(a) Samples in a two-dimensional input space

(b) Mapped samples in similarity space. Each axis represents the similarity of the point to each class.

Fig. 1: Example of similarity mapping using a Gaussian kernel.

3 Methodology

In kernel inductive learning, the data set is generally fixed and given in advance so, in order to obtain the mapping, the type of kernel and its corresponding parameters should be set. Since the type of kernel is also given and usually not learned, similarity mapping relies solely on kernel parameters θ. As different values of θ yield different mappings, this work aims at a function to assess the effect of θ on pattern relations and on separability in the projected space.

Statistics of the projected data, e.g. mean and covariances, are capable to express the response of the projections with relation to kernel parameters, so here the average $\Psi_{k,\theta}(x)$ is considered to assess the central tendency of kernel projections as θ changes, as represented in Equation 4.

$$\Psi_{k,\theta}(\mathbb{X}_C) = \frac{1}{|\mathbb{X}_C|} \sum_{x \in \mathbb{X}_C} \Psi_{k,\theta}(x) \tag{4}$$

where $\mathbb{X}_C$ is the set of all samples belonging to class C. Since the mean vector is still in the same space, the notation Ψ is used to represent the similarity mapping of both a point and a set.

Projections tend to perform a trajectory in the plane $\Psi(\mathbf{x}, \mathbb{X}_1) \times \Psi(\mathbf{x}, \mathbb{X}_2)$ as a function of θ. Consider, for instance, the radius σ of an RBF kernel that is increased gradually while the corresponding projections are observed. Very small values of σ tend to map all samples close to the origin, since pairwise similarities get closer to zero. In such a situation, projected data has very small variances and little inter-class covariances due to the small receptive fields of kernel functions. As σ increases, the projections move from the origin and describe a trajectory in the space towards the point $[1, 1]^T$ to where all projections converge as $\sigma \to \infty$. Between the extreme points $[0, 0]^T$ and $[1, 1]^T$, the two clouds of projected points spread out, separate from each other and then get closer again as they

ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

approximate to $[1, 1]^T$ with the increase of σ. The Euclidean distance between vectors $\Psi_{k,\theta}(\mathbb{X}_{C_1})$ and $\Psi_{k,\theta}(\mathbb{X}_{C_2})$ is then considered as a measure of separability in similarity space, since it also quantifies proximity in relation to the linear separator $\Psi(\mathbf{x}, \mathbf{X}_1) = \Psi(\mathbf{x}, \mathbf{X}_2)$. Optimization of such a distance function is based on the prior assumption that covariances gained at the maximum embody the intricate pattern relations that will implicitly trade-off bias and variance of the resulting model. It is important to notice that it does not require validation sets and is not based on extensive search of parameters. The problem of optimizing kernel parameters θ on a dataset $\mathbb{X}$ with classes C_1 and C_2 is formalized as in Equation 5, as follows.

$$\theta^* = \arg\max_{\theta} ||\Psi_{k,\theta}(\mathbb{X}_{C_1}) - \Psi_{k,\theta}(\mathbb{X}_{C_2})||_2^2 \tag{5}$$

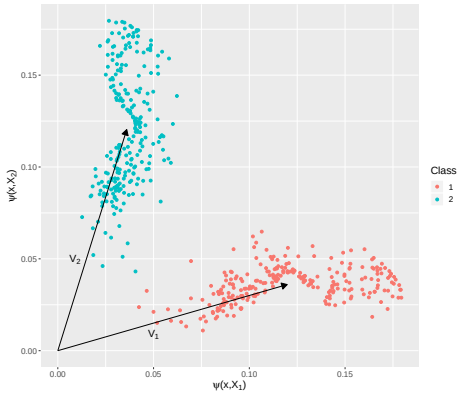


Fig. 2: Similarity mapping along with the class vectors.

4 Experiments

4.1 Setting

The proposed method was tested using support vector classification. Three different kernels were used, where their hyperparameters were chosen beforehand using the proposed criterion. Posteriorly, SVMs were trained with the chosen kernels, where only the flexible margin parameter C was tuned using grid-search. Baseline SVMs were trained using each kernel with both C and kernel hyperparameters being optimized jointly via grid-search.

Tests were conducted in 18 real-world datasets. The “Appendicitis" dataset (**appendicitis**) was obtained from the KEEL datasets repository [7]; “Breast Cancer Hess Probes" (**breastHess**) was obtained from [8]; and the dataset “Gene expression dataset" (**golub**) was obtained from [9]. All of the others come from the UCI Machine Learning Repository [10]. The dataset “Glass Identification" (**glass**), which has seven distinct classes, was modified to become

ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

a binary classification problem. Class 7 is considered the positive class, while all of the others are considered negatives.

4.2 Results

Accuracy values were measured for each dataset. Test metrics were obtained using 10-fold validation, and the confidence intervals use a significance of 0.05. Numerical results are on Table 1.

Table 1: Accuracy values on real-world datasets. Multiple kernel families were optimizing using grid-search (**GS**) and the proposed method (**SIM**).

Basename	Gaussian-GS	Gaussian-SIM	Laplacian-GS	Laplacian-SIM
fertility	0.870 ± 0.035	0.880 ± 0.030	0.880 ± 0.030	0.880 ± 0.030
appendicitis	0.885 ± 0.077	0.876 ± 0.079	0.838 ± 0.065	0.876 ± 0.072
australian	0.867 ± 0.037	0.861 ± 0.033	0.854 ± 0.044	0.865 ± 0.044
german	0.762 ± 0.031	0.760 ± 0.0281	0.719 ± 0.011	0.768 ± 0.021
golub	0.821 ± 0.042	0.795 ± 0.062	0.654 ± 0.048	0.795 ± 0.062
banknote	1.000 ± 0.000	1.000 ± 0.000	0.999 ± 0.002	0.999 ± 0.002
glass	0.972 ± 0.017	0.963 ± 0.014	0.963 ± 0.014	0.967 ± 0.016
ILPD	0.701 ± 0.032	0.715 ± 0.006	0.712 ± 0.011	0.706 ± 0.022
haberman	0.729 ± 0.043	0.725 ± 0.048	0.719 ± 0.036	0.725 ± 0.031
sonar	0.855 ± 0.057	0.870 ± 0.044	0.620 ± 0.032	0.817 ± 0.072
breastHess	0.805 ± 0.063	0.827 ± 0.062	0.746 ± 0.048	0.812 ± 0.069
breastcancer	0.968 ± 0.014	0.962 ± 0.015	0.963 ± 0.016	0.963 ± 0.017
parkinsons	0.933 ± 0.054	0.902 ± 0.040	0.907 ± 0.039	0.943 ± 0.047
heart	0.826 ± 0.042	0.844 ± 0.035	0.837 ± 0.047	0.833 ± 0.034
climate	0.954 ± 0.013	0.950 ± 0.013	0.915 ± 0.007	0.944 ± 0.009
diabetes	0.769 ± 0.025	0.773 ± 0.020	0.751 ± 0.027	0.775 ± 0.022
ionosphere	0.934 ± 0.031	0.946 ± 0.020	0.929 ± 0.022	0.943 ± 0.019
bupa	0.715 ± 0.045	0.698 ± 0.047	0.675 ± 0.048	0.696 ± 0.035
average	**0.854 ± 0.016**	**0.853 ± 0.015**	**0.816 ± 0.018**	**0.851 ± 0.016**

5 Conclusions and discussions

Trading-off bias and variance of models [11] is a basic principle of inductive machine learning. It is proved formally, so the designer will always need to choose how to balance two or more objective functions. However, the search for a global equilibrium point, without the need to select the importance of objective functions, has always been an important topic of research. Automatic selection of parameters that would allow the designer to fully rely on a prior decision strategy is hard to achieve and has been mainly based on exhaustive computational search in parameter's space in recent years. The strategy introduced in this paper presents a different perspective of the problem by focusing decision making on the choice of objective function, as represented in Equation 5. The set of parameters θ that maximize such a function are the ones selected to tune-up the model. The selection is direct, since the function can be maximized, but it is based on a prior understanding of the aimed properties of the projections in similarities space. Visualization of explicit mapping allows for further explorations of new objective functions that could express the aimed trade-off expressed by the projections.

ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

The method has proven to be competitive with grid-search, considered as a baseline, in many datasets.

References

[1] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[2] Murilo VF Menezes, Luiz CB Torres, and Antonio P Braga. Width optimization of rbf kernels for binary classification of support vector machines: A density estimation-based approach. *Pattern Recognition Letters*, 128:1–7, 2019.

[3] Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13:723–773, 2012.

[4] Bharath K. Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Bernhard Schölkopf, and Gert R. G. Lanckriet. On integral probability metrics, ϕ-divergences and binary classification. (1):1–18, 2009.

[5] Bernard W Silverman. *Density estimation for statistics and data analysis*, volume 26. CRC press, 1986.

[6] Bernhard Scholkopf and Alexander J Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2001.

[7] Jesús Alcalá-Fdez, Alberto Fernández, Julián Luengo, Joaquín Derrac, Salvador García, Luciano Sánchez, and Francisco Herrera. Keel data-mining software tool: data set repository, integration of algorithms and experimental analysis framework. *Journal of Multiple-Valued Logic & Soft Computing*, 17, 2011.

[8] Kenneth R Hess, Keith Anderson, W Fraser Symmans, Vicente Valero, Nuhad Ibrahim, Jaime A Mejia, Daniel Booser, Richard L Theriault, Aman U Buzdar, Peter J Dempsey, et al. Pharmacogenomic predictor of sensitivity to preoperative chemotherapy with paclitaxel and fluorouracil, doxorubicin, and cyclophosphamide in breast cancer. *Journal of clinical oncology*, 24(26):4236–4244, 2006.

[9] Todd R Golub, Donna K Slonim, Pablo Tamayo, Christine Huard, Michelle Gaasenbeek, Jill P Mesirov, Hilary Coller, Mignon L Loh, James R Downing, Mark A Caligiuri, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *science*, 286(5439):531–537, 1999.

[10] M. Lichman. UCI machine learning repository, 2013.

[11] Stuart Geman, Elie Bienenstock, and René Doursat. Neural networks and the bias/variance dilemma. *Neural computation*, 4(1):1–58, 1992.