Kernel Multi Label Vector Optimization (kMLVO): A Unified Multi-Label Classification Formalism

Gilad Liberman[1], Tal Vider-Shalit[2], and Yoram Louzoun[2(✉)]

[1] Gonda Multidisciplinary Brain Research Center,
Bar Ilan University, Ramat Gan, Israel
[2] Department of Mathematics, Bar Ilan University, Ramat Gan, Israel
louzouy@math.biu.ac.il

Abstract. We here propose the kMLVO (kernel Multi-Label Vector Optimization) framework designed to handle the common case in binary classification problems, where the observations, at least in part, are not given as an explicit class label, but rather as several scores which relate to the binary classification. Rather than handling each of the scores and the labeling data as separate problems, the kMLVO framework seeks a classifier which will satisfy all the corresponding constraints simultaneously. The framework can naturally handle problems where each of the scores is related differently to the classifying problem, optimizing both the classification, the regressions and the transformations into the different scores. Results from simulations and a protein docking problem in immunology are discussed, and the suggested method is shown to outperform both the corresponding SVM and SVR.

1 Introduction

Classic supervised learning problems are formulated as a set of (x_i, y_i) pairs, where x_i lies in the problem domain (typically $\mathbb{R}^n$, but may be more complex), and $y_i \in \{0, 1, \ldots, n\}$ for classification problems (with $n = 1$ for decision problems) or $y_i \in \mathbb{R}$ for regression problems. Naturally, not all problems fall into these categories and several generalization have been suggested, where each instance belongs to more than one class or where multiple instances have multiple labels [1,2]. In this study, we keep the assumption that each instance either belongs to some target class or does not; however, the available data might not contain this labeling but rather some indirect measurements. This formulation is related to many real life problems. For example, in the medical domain, the decision whether a subject is ill or not is made not just based on past subjects' data along with their diagnoses, but also on past subjects' data along with their physiological condition scores, appetite and happiness scores, etc.

This notion can be especially helpful in areas where the classification is difficult to obtain, with limited data sets, or where the data suffers from high variation in measurement modality and protocol. The application of a first solution,

G. Nicosia and P. Pardalos (Eds.): LION 7, LNCS 7997, pp. 131–137, 2013.
DOI: 10.1007/978-3-642-44973-4_15,

the MLVO method, was briefly introduced in a recent publication [3]. Here, we present the extention of this fomalism to a kernel machine - the kMLVO, along with some useful extensions.

The remainder of this manuscript is organized as follows. In Sect. 2, we present the kMLVO framework. Extensions of the kMLVO are shown in Sect. 3. In Sect. 4 we discuss the results on simulated and experimental data sets and compare the performance of different classifiers, and we conclude with a summary in Sect. 5.

2 The kMLVO Framework

We use the SVR formalization for the regression part. This implies linear penalties on regression errors, which improves the robustness to outliers. When using a non-linear kernel, the direct relation to the original problem dimension is lost, and the contribution of w_0 must be indirect (we will return to this issue in the kMLVO extensions).

2.1 Formalism

We note by X the training set, which is now not restricted to be a subset of $\mathbb{R}^n$ but can be of any input space $\mathcal{X}$. Let each instance $x_i \in \mathcal{X}$ of the training set be associated with a class $y_i \in \{-1, 1, \phi\}$, where ϕ represents an uknown value (i.e. we do not know the classification for this point) and with L target values $s_{i,l} \in \mathbb{R} \bigcup \{\phi\}$. The classifier is a function $f : \mathcal{X} \rightarrow \{-1, 1\}$. In a manner analogous to the SVM and SVR primal formulation, the classifier f is a weight vector w in the problem space (or a vector space in which the prolem is transformed into using explicit transformation or a multiplication kernel) and a bias b. The optimality criteria for the classifier is the (soft) maximum margin both for the classification and the regression tasks, with different costs. We then note by C the cost vector (or scalar) for each misclassified sample, and by D the cost matrix for the regression tasks, i.e. $D_{i,l}$ is the cost for the ith sample, for the lth measurement. The problem can be written, ignoring w_0, as:

$$\begin{aligned} \underset{w,b,\xi,\alpha,\beta,\zeta,\zeta^*}{\text{minimize}} \quad & \frac{1}{2}\|w\|^2 + C^T\xi + \sum_{i=1}^{L} D_i^T\zeta_i^+ \\ \text{subject to} \quad & Y(Xw + 1b) \geq 1 - \xi,\ \xi \geq 0 \\ & (\alpha_i s_i + 1\beta_i) - Xw - 1b \leq 1\epsilon_i + \zeta_i,\ \zeta_i \geq 0, \forall_i, 1 \leq i \leq L \\ & Xw + 1b - (\alpha_i s_i + 1\beta_i) \leq 1\epsilon_i + \zeta_i^*,\ \zeta_i^* \geq 0, \forall_i, 1 \leq i \leq L \end{aligned} \tag{1}$$

Where s_i is the vector of scores for the ith modality, for all of the samples. Note that the units of the continuous scores s_i may differ from the units of the optimal separating hyper-plane of the binary data, thus we added the vectors α, β of length L, with the linear transformations for the corresponding modalities. ϵ is a vector of length L containing the insensitive loss parameters for the different

modalities (as in [4]). Note that since for each modality this parameter is applied after the linear transformation, its value is normalized and searching for the optimal value is easier and indicative of the regression fit (which is otherwise hidden).

2.2 Transition to the Dual Problem

We first transform the primal problem into its dual problem and then apply a KKT formalism to it [5,6]. Then, using the standard SVM technique, the problem can be written in a Lagrangian formalism and using the fact that the appropriate partial derivatives equal 0 at the optimum, we get the final formulation as a quadratic problem:

$$\begin{aligned}
\underset{\mu,\delta,\delta^*}{\text{maximize}} \quad & -\frac{1}{2}(\mu^T Y^T X X^T Y \mu + (\sum_{i=1}^{L} \delta_i^{-T}) X X^T (\sum_{i=1}^{L} \delta_i^{-})) + \mu^T 1 - \sum_{i=1}^{L} \delta_i^{+T} 1 \epsilon_i \\
& - \sum_{i=1}^{l} \delta_i^{-T} X X^T Y \mu \\
\text{subject to} \quad & \mu, \lambda, \delta_i, \delta_i^*, \eta_i, \eta_i^* \geq 0, \forall_i, 1 \leq i \leq L \\
& \mu^T y = 0; 1^T \delta_i^- = 0; s_i^T \delta_i^- = 0, 1 \leq i \leq L \\
& C = \mu + \lambda; D_i = \delta_i^{(*)} + \eta_i^{(*)}, \forall_i, 1 \leq i \leq L
\end{aligned} \tag{2}$$

where $\delta_i^- = \delta_i - \delta_i^*$ and $\delta_i^+ = \delta_i + \delta_i^*$, with μ, λ being the Lagrange multipliers corresponding to the classification problem (as in the SVM fromalism) and $\delta_i, \delta_i^*, \eta_i, \eta_i^*$ the Lagrange multipliers corresponding to the i-th modality of the regression problem (as in the SVR formalism, following [4,7]). Using such a formalism, we enjoy the advantages of a kernel machine, i.e. the optimization is on the support vectors coefficients μ, δ, δ^* and the (possibly high-dimensional) product XX^T can be replaced using any kernel function K. Here the decision function becomes $g(q) = \sum_{j=1}^{n}(\mu_j y_j + \sum_{i=1}^{L} \delta_{i,j}^-) K(x_j, q) + b$, i.e. a weighted sum on the contributions of the kernel function of the support vectors with the classified sample point, where the weight on each support vector considers both the classification and the various regression constraints.

2.3 Handling Missing Values

The method can handle any combination of inputs, by simply setting the corresponding element of the classification cost vector C and/or of the cost matrix D to 0 where a value is missing. This constraints the corresponding support vectors coefficient to be fixed (box constaint of 0) and effectively removes the element from the optimization problem, while keeping all the other information intact.

3 Extending kMLVO

Two possible extensions of the kMLVO framework handle the incorporation of w_0, as in the MLVO, and non-linear regression.

3.1 Incorporation of w_0

Given w_0, we would like to introduce a penalty when diverging from it which is simiar to the one used in the MLVO, i.e. $E_2 = \frac{1}{2}\|w_0 - w\|_2^2$. This however will automatically lead to terms which are not quadratic in the values of x_i. This can be solved by projecting w and w_0 on any basis of the feature space. Given a base $B = \{B_i, \cdots, B_n\}$ of the feature space, $\|w_0 - w\|_2^2 = \sum_{i=1}^{n}(< w_0, B_i > - < w, B_i >)^2 = \sum_{i=1}^{n}(\mu_i - < w, B_i >)^2$ where $< w_0, B_i >= \mu_i, \forall_i$ is the score induced by w_0 for the base vectors. We would like the scores induced by w to be similar, i.e. the problem is transformed into a regression problem. Thus it suffices to add the base vectors $\{B_i, \cdots, B_n\}$ as additional input samples along with their w_0 induced scores. As before, while the MLVO uses squared penalty, the kMLVO uses L_1 penalty.

3.2 Non-linear Regression

Suppose that the scores for the i-th modality are not linear with the optimal (or real) separating hyperplane for the classification problem, but follow $s_{i,k} = f(w^T x_k + b) = f(\sum_{j=1}^{n} \mu_j y_j K(x_j, x_k) + b)$ for some function f. In this case we would like to linearize the scores, i.e. applying f^{-1} before performing the kMLVO. If f (and thus f^{-1}) is unknown, we may let the kMLVO approximate it as a linear combination of score vectors. This can be performed by replacing the term $(\alpha_i s_i + 1\beta_i)$ in the equations with $(\alpha_{i,1} s_{i,1} + \cdots + \alpha_{i,p_i} s_{i,p_i} + 1\beta_i)$. The only additional constraints added to the final dual problem is:

$$s_{i,1}^T \delta_i^- = 0, \ldots, s_{i,p_i}^T \delta_i^- = 0 \tag{3}$$

These different scores can be, but not limited to, the original scores s_i in different powers, etc.

4 Simulations

In order to test whether the proposed formalism outperforms existing methods, we have compared the precision obtained using four different formalisms: SVR, SVM, MLVO and kMLVO on artificial datasets of different dimentionality, noise levels, and sample sizes. Additinaly, 10 % of the points were randomly chosen to be "outliers". For these points, the standard deviation of the added noise was 3 times the sandard deviation of all scores (instead of 0.03 or 0.6).

4.1 kMLVO Results on Simulations

The kMLVO outperforms the other classifiers in the presence of outliers. Such outliers can significantly affect the SVM and SVR formalisms, and we have here tested their effect on the kMLVO formalism. The average performace scores on the different data sets can be seen in Fig. 1. While for weak noise levels the MLVO is dominant, in the stronge noise level, a clear dominancy of the kMLVO can be seen, especially in the region of high number of samples with continous scores and a low number of binary samples. This can be explained by the fact that kMLVO (as SVR) is has L_1 regularization term, while the regression part of the MLVO (as LS-SVR) is regularized with an L_2 term.

In another application, regarding the binding to an immune system molecule, the transporter associated with antigen processing (TAP), the kMLVO outperformed the MLVO and the uni-label classifiers SVM (using binding/non binding data) and SVR (using affinity score), of the commonly used package LibSVM [8] on our data.

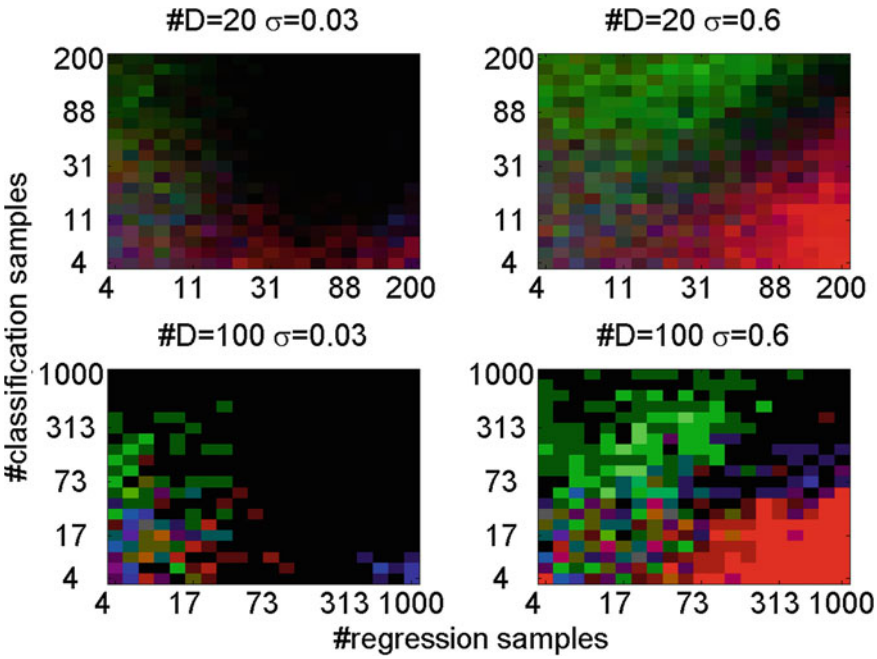

Fig. 1. Simulation results with outliers. The relative number of winners (that is, most accurate estimation of the direction vector) is coded to RGB by red - kMLVO, green - SVM, blue - SVR, and black - MLVO. The titles refer to the dimensionality of the data set and the noise's standard deviation.

5 Discussion

The approach presented can be used as a general supervised learning method when multiple labels of data are available. Such a situation often emerges in biological interactions, such as transcription factor binding or protein-protein interactions. In such cases, observations can either be binary (the presence or absence of an interaction) or continuous (the affinity).

Several extensions of the kMLVO have been proposed. The simplest expansion is the use of multiple continuous scores. Assume samples having continuous scores that are derived from several unit scales (e.g. IC50 and EC50 affinity related measurements). As part of the solution (as described above), we simultaneously fit between the predicted to the continuous score by linear regression. Thus, actually all the available measurements can be merged together, and the problem will be transformed to a set of linear regressions with multiple values of α and β. The algorithm can also be improved if the validity of the different dimensions of the samples in the n dimensional space or the validity of the samples themselves can be guessed. In such a case, the weight given to the similarity to the a priori guess in each dimension or the error of each classified data point (ξ_i) can be varied.

The use of kernels, along with extension for handling multiple measurements types, with different non-linear relations, and inherent consideration of missing values gives the suggested approach a higher flexibility and applicability for real-life problems.

The main limitations of the proposed methodology is that it mainly applies to cases where the number of observations is limited. When the number of observations is very large and biased toward one type of observations, the MLVO performes worse than the appropriate SVM or SVR. Another important caveat is the need to determine three constants, instead of the single box constraint constant in the standrad SVM formalism. In the presented cases, we have prededefined the constant to be used, or used an internal test set to determine the optimal constants, and then applied the results to an external test set. Even when these caveats are taken into consideration, the MLVO can be an important methodological approach in many biological cases, where the number of observations is highly limited.

Acknowledgment. We would like to thank M. Beller for editing this manuscript.

References

1. Zhou, Z., Zhang, M., Huang, S., Li, Y.: Multi-instance multi-label learning. Artif. Intell. **176**, 2291–2320 (2011)
2. Boutell, M., Luo, J., Shen, X., Brown, C.: Learning multi-label scene classification. Pattern Recogn. **37**, 1757–1771 (2004)
3. Vider-Shalit, T., Louzoun, Y.: Mhc-i prediction using a combination of t cell epitopes and mhc-i binding peptides. J. Immunol. Methods **374**, 43–46 (2010)

4. Farag, A., Mohamed, R.M: Regression using support vector machines: basic foundations. Technical Report, CVIP Laboratory, University of Louisville (2004)
5. Karush, W.: Minima of functions of several variables with inequalities as side constraints. Master's thesis, Department of Mathematics, University of Chicago (1939)
6. Kuhn, H., Tucker, A.: Nonlinear programming. In: Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability, California, vol. 5 (1951)
7. Cortes, C., Vapnik, V.: Support-vector networks. Mach. Learn. **20**, 273–297 (1995)
8. Chang, C.C., Lin, C.J.: LIBSVM: a library for support vector machines. ACM Trans. Intell. Sys. Technol. **2**, 27:1–27:27 (2011)