On Model Complexity Reduction in Instance-Based Learners

Saulo A. F. Oliveira*
MDCC – UFC
Federal University of Ceará (UFC)
Fortaleza, Brazil
saulo.freitas.oliveira@gmail.com

Ajalmar R. Rocha Neto
PPGCC – IFCE
Federal Institute of Ceará (IFCE)
Fortaleza, Brazil
ajalmar@ifce.edu.br

João P. P. Gomes
MDCC – UFC
Federal University of Ceará (UFC)
Fortaleza, Brazil
jpaulo@dc.ufc.br

***Abstract*—Instance-based learners habitually adopt instance selection techniques to reduce complexity and avoid overfitting. Such learners' most recent and well-known formulations seek to impose some sparsity in their training and prediction structure alongside regularization to meet such a result. Due to the variety of such instance-based learners, we will draw attention to the Least-Squares Support Vector Machines and Minimal Learning Machines because they embody additional information beyond the stored instances to perform predictions. Later, this thesis proposes variants constraining candidate solutions within a specific functional space where we avoid overfitting and reduce model complexity. The central core of such variants is related to penalizing samples with a specific condition during learning. For regressors, we adopted strategies based on random and observed linearity conditions related to the data. At the same time, we borrowed definitions from the computer vision field for classification tasks to derive a concept we call the class-corner relationship (in which we designed an instance selection algorithm). In the Least-Squares Support Vector Machines context, this thesis follows the pruning fashion by adopting the samples that share such a class-corner relationship. As for the Minimal Learning Machine model, this thesis introduces a new proposal called the Lightweight Minimal Learning Machine, a faster model for out-of-sample prediction due to the reduced number of computations inherent in the original proposal's multilateration process. Another remarkable feature is that it derives a unique solution when other formulations rely on overdetermined systems.**

I. Introduction

Instance-based learners are computational models that, instead of making explicit generalizations, compare instances of new problems with instances seen in the learning process (previously stored in memory) [1]. Such models build hypotheses directly from the training instances themselves, thus implying that the hypotheses' complexity can grow with the data. For complex enough models with a large number of free parameters, a perfect fit to the training data is possible [2].

In this case, reducing complexity means restricting the amount of data used in the learning process. By this reduction, we also restrict the space of hypotheses that the model can generalize. Following the principle of Occam's razor [3]: balancing both the complexity of the induced model and the ability to generalize ends up being a challenging task, while it is also highly desired.

Nevertheless, sparsity in LSSVMs and MLMs is associated with restricting the hypothesis space since it accounts for (and is used for restricting) the number of parameters to later measure the model complexity. In short, one can account for the degree of freedom via sparsity, but sparsity is not directly related to generalization.

This thesis follows a different path from most works in current literature. Its main object of study and contribution is related to using an instance selection algorithm as a form of regularization in the complexity term. In doing so, we achieved a direct method. From that, we investigate two **hypotheses**, listed as research questions:

1) Can one control the model's complexity without sacrificing the model's generalization by directly incorporating an instance selection algorithm into an instance-based learner?
2) Since such a selection gets rid of some data, can some speedup process incur an out-of-sample prediction?

Both hypotheses deal with two existing problems in instance-based models: (i) the lack of an instance selection mechanism wrapped up in the learning algorithm; (ii) and control of the model's complexity. This thesis unfolds four contributions. The first two were evaluated together in [4], while the third one is presented in [5]. The last part has not been published yet. Thus, only available here and in the final thesis document.

II. Contribution 1: Class-Corner Instance Selection

Our first contribution is mainly based on FAST [6], an image corner detector. The main idea is to use the definition of what is a corner[1] in FAST and then apply the same reasoning as the Instance Selection algorithm. However, it turns out that FAST formulation only deals with image data, i.e., two dimensional samples that are uniformly spaced in a grid. To overcome such limitations, [4] extended FAST so that we can apply it to high-dimensional inputs in a straightforward way. Although, the authors did not give it a proper name in [4], here we call it Class-Corner Instance Selection (CCIS).

[*] Ph.D. Thesis.

[1] A corner can be defined as the intersection of two edges. An edge (usually a step change in intensity) in an image corresponds to the boundary between two regions [6].

In FAST, all pixels are evaluated as corners or not. However, before such a classification, two subsets are derived: the candidate, and the actual corner set, respectively. CCIS adopts a similar analogy but for general data points. First, CCIS performs a greedy filtering approach to identify the input samples that somehow lie in the class-corner regions of each class to use them as the selected subset.

Since in real-world problems, one can not assume the data $\{\mathbf{x}_i\}_{i=1}^{N}$ is equally spaced as the pixels in an image grid, nor do they share a piece of neighboring information. CCIS emulates such an information it by employing the R-ball neighborhood of a query sample $\mathbf{x}$ as follows:

$$\mathcal{N}_R(\mathbf{x}) = \Big\{\mathbf{x}_i \in \mathcal{NN}_K(\mathbf{x}) \;\Big|\; 0 < \|\mathbf{x}_i - \mathbf{x}\|_2 \leq R\Big\}, \qquad (1)$$

where $R \in \mathbb{R}_+$ is the radius of the circle mask and $\mathcal{NN}_K(\cdot)$ yields the set with K nearest neighbors.

To identify the corner candidates and actual corner samples,

$$\Gamma(\mathbf{x}) = \sum_{\mathbf{x}_i} \mathbb{1}[\mathbf{y} \neq \mathbf{y}_i],\;\; \mathbf{x}_i \in \mathcal{N}_R(\mathbf{x}), \qquad (2)$$

where $\mathbb{1}[\cdot]$ is the indicator function equal to 1 if its argument is true and 0 otherwise. Note that, $\Gamma(\mathbf{x})$ is a simple counting function that yields the number of neighbors with different class labels than the query sample $\mathbf{x}$ inside the R-ball. Finally, such an identification function can be employed to select subsets according to a threshold P as follows:

$$\mathcal{PS} = \Big\{(\mathbf{x}_i, \mathbf{y}_i) \in \mathcal{D} \;\Big|\; \Gamma(\mathbf{x}_i) > P\Big\}, \qquad (3)$$

where $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{N}$. We discussed in [4] that by adopting $P = 0$ and default values for K and P, we derived feasible subsets that share this class-corner feature. **We validated our first contribution alongside the following one**.

III. Contribution 2: Class-Corner Least-Squares Support Vector Machine

Our second contribution is a well-succeded attempt to reduce the LSSVM complexity by directly employing CCIS. We named such a proposal Class-Corner Least-Squares Support Vector Machine (CC-LSSVM). Here, the $\mathcal{SV} = \text{CCIS}(\mathcal{D}, 0)$ is the set of support vectors while $\mathcal{PS} = \text{CCIS}(\mathcal{D}, P)$ represents the constraints in the model with threshold P, thus, keeping the link between the variables and constraints since $\mathcal{SV} \subset \mathcal{PS} \subset \mathcal{D}$. Then, we formulate the linear system in CC-LSVM as $\boldsymbol{\Lambda}\boldsymbol{\omega} = \boldsymbol{v}$ so that

$$\underbrace{\left[\begin{array}{c|c} 0 & \mathbf{1}^\intercal \\ \hline \mathbf{1} & \boldsymbol{\Psi} \end{array}\right]}_{\boldsymbol{\Lambda}} \underbrace{\begin{bmatrix} b^\star \\ \boldsymbol{\alpha}^\star \end{bmatrix}}_{\boldsymbol{\omega}} = \underbrace{\begin{bmatrix} 0 \\ \mathbf{Y}_{\text{SV}} \end{bmatrix}}_{\boldsymbol{v}}, \qquad (4)$$

where $\mathbf{Y}_{\text{SV}} = [y_1, y_2, \ldots, y_M]^\intercal$ is a matrix with labels from $\mathcal{SV}$ and

$$\Psi_{i,j} = \begin{cases} \mathcal{K}(\mathbf{x}_i, \mathbf{x}_j), & \text{if } \mathbf{x}_i \neq \mathbf{x}_j; \\ \mathcal{K}(\mathbf{x}_i, \mathbf{x}_j) + \gamma^{-1}, & \text{otherwise.} \end{cases}$$

with $\mathbf{x}_i \in \mathcal{PS}$ and $\mathbf{x}_j \in \mathcal{SV}$ and $\gamma \in \mathbb{R}_+$ is the same cost parameter in the original (LS)SVM.

A. Validation

In the first part, we validate CCIS and its ability to reduce the original datasets to get both $\mathcal{PS}$ and $\mathcal{SV}$, and later, we employ both on CC-LSSVM. We present in Fig. 1 bar plots showing the scaled and actual training set sizes for further analysis concerning how CCIS works. From that, one can see that in small data sets, CCIS not always reduce much from the training set $\mathcal{D}$ to the prototype set $\mathcal{PS}$. However, the size of $\mathcal{SV}$ is shown to be very small compared to the training set $\mathcal{D}$, i.e., $M \ll N$, thus, suggesting that our class-corner selection considers both the model size and model generalization capability since discarding too much data can be harmfull for generalization.

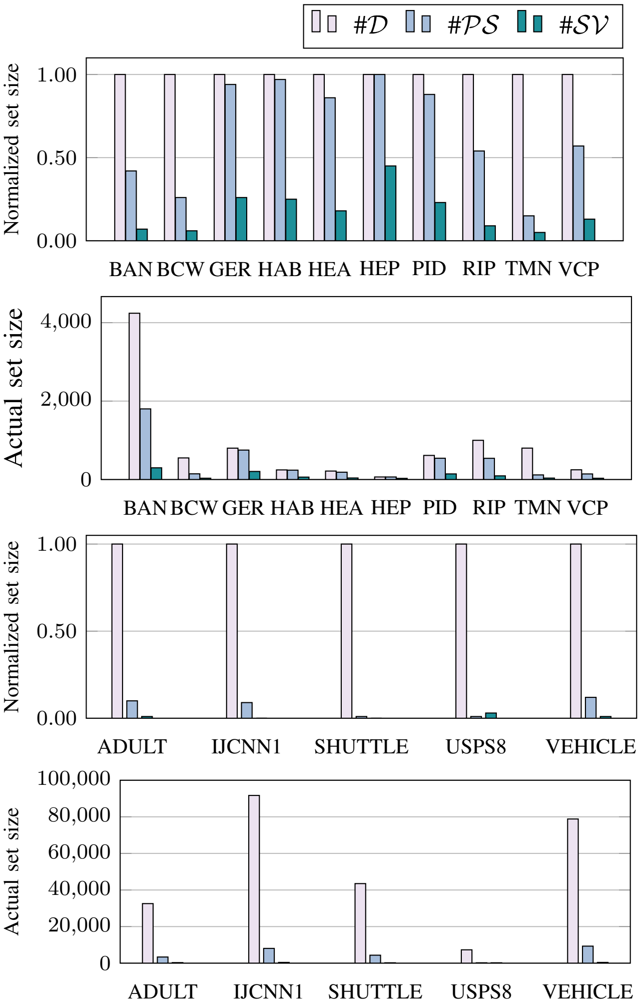
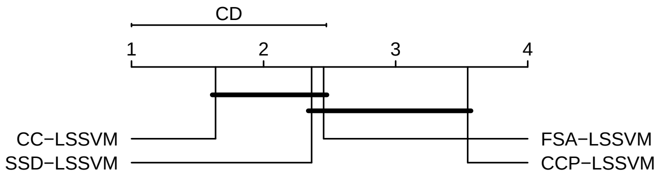


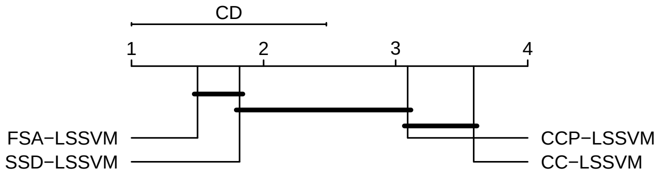
Fig. 1. Bar plots adopted from [4] showing the original training set $\mathcal{D}$, the number of elements in $\mathcal{PS}$, and $\mathcal{SV}$ for CC-LSSVM via CCIS. We show both the scaled and the actual sizes for each *toy size* and large size datasets.

Such a finding is also reinforced when analyzing Fig. 1, where one can see a dramatical reduction in each step, especially from $\mathcal{D}$ to $\mathcal{PS}$.

Next, we investigated how CC-LSSVM behaves concerning the following aspects: accuracy, sparseness, support vector selection, and hyperparameter sensitivity. Although not presented in this paper, we highlight that we originated the CD plots from the tables where we reported the metric results. Such tables are shown in the original thesis and [4], [5]. We hide them for space restrictions.



(a) Accuracy rankings.



(b) Sparsity rankings.

Fig. 2. Critical difference plots for toy problems. We highlight that those models which are not joined by a bold line can be regarded as different.

CC-LSSVM performed the best rank for the accuracy criteria, thus, showing high generalization performance against other variants. However, CC-LSSVM performed the last for the sparsity criteria, see the Critical difference plots[2] in Fig. 2. Such a finding indicates that our class-corner support vector selection via CCIS can balance between a smaller model and a less constrained one without sacrificing the generalization performance.

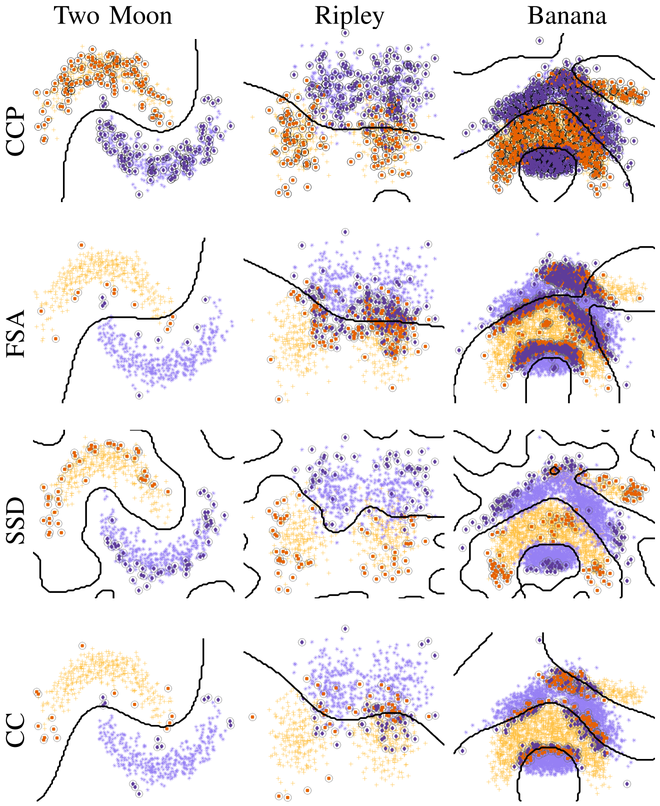
IV. Proposal 3: The Lightweight Minimal Learning Machine

A. The original Minimal Learning Machine

The Minimal Learning Machine (MLM) [8] is a supervised method used for pattern recognition and regression tasks. From the general framework for supervised learning, MLM estimates $h(\cdot)$ for the target function $f(\cdot)$ from the data $\mathcal{D}$ through the distance domain. For that, the problem is stated by employing pairwise distance matrices of each point of $\mathcal{D}$, namely, $\mathbf{D}$ and $\mathbf{\Delta}$, both representing the Euclidean distance – in the notation of $\mathrm{d}(\cdot,\cdot)$ – of each point from $\mathcal{D}$ to the i-th reference point of $\mathcal{D}$, i.e., $D_{i,j} = \mathrm{d}(\mathbf{x}_i, \mathbf{x}_j)$ and $\mathbf{\Delta}_{i,j} = \mathrm{d}(\mathbf{y}_i, \mathbf{y}_j)$, then they have $N \times N$ dimensions.

For the sake of simplicity, in the following description and notation of MLM, consider the two distance mapping functions $\Phi : \mathbb{R}^D \to \mathbb{R}^N$ and $\Psi : \mathbb{R}^S \to \mathbb{R}^N$. Here $\Phi(\mathbf{x}) = \left[\mathrm{d}(\mathbf{x},\mathbf{x}_1), \mathrm{d}(\mathbf{x},\mathbf{x}_2), \ldots, \mathrm{d}(\mathbf{x},\mathbf{x}_N)\right]^\intercal$ while $\Psi(\mathbf{y}) = \left[\mathrm{d}(\mathbf{y},\mathbf{y}_1), \mathrm{d}(\mathbf{y},\mathbf{y}_2), \ldots, \mathrm{d}(\mathbf{y},\mathbf{y}_N)\right]^\intercal$. Furthermore, we call $\mathbf{D}$ and $\mathbf{\Delta}$ the input and output spaces, respectively. Stated that, by assuming that the mapping between the distance matrices has a linear structure for each response, the MLM model can be rewritten in the form $\mathbf{\Delta} = \mathbf{DB} + \mathbf{E}$, where $\mathbf{E}$ being the residuals matrix.

Fig. 3. Decision boundaries and support vector selections by LSSVM variants in some 2D datasets. Adapted from [4].



The learning process consists of finding the mapping between the distances in the input and output space. Since we assumed such a mapping has a linear structure for each response, the regression model can be rewritten in the form:

$$\min_{\mathbf{B}} \;\; \mathcal{J}(\mathbf{B}) = \| \, \mathbf{DB} - \mathbf{\Delta} \, \|_{\mathcal{F}}^2, \tag{5}$$

and it can be estimated by:

$$\hat{\mathbf{B}} = \mathbf{D}^{-1}\mathbf{\Delta}. \tag{6}$$

Predicting the outputs for new input data mainly refers to project the new data point through the mapping and estimate the image of such a projection. Therefore, it is necessary that the pattern $\mathbf{x}$ be also represented in the domain of distances so that we can represent it in the output space. Such a representation is achieved by $\Phi(\mathbf{x})\,\mathbf{B}$.

From this, the problem is to estimate the image $\hat{\mathbf{y}} = h(\mathbf{x})$, from $\Phi(\mathbf{x})\,\mathbf{B}$ and the images of reference points. This problem can be treated as a multilateration [9]. In a geometric viewpoint, estimate $\hat{\mathbf{y}}$ belonging to the set $\mathbb{R}^S$ is equivalent to solving the determined set of N non-linear equations corresponding to the S-dimensional hyper-spheres centered on the images of the reference points, denoted by $\{\mathbf{y}_i\}_{i=1}^N$. The

[2] In this plot, the top line is the axis on which each model's average ranks, while the connected group of models indicated that they are not significantly different. Moreover, the critical difference (CD) is presented above the plot and by adopting the significance level of $\alpha = 0.05$ [7].

location of $\hat{\mathbf{y}}$ can be estimated by minimizing the objective function below:

$$\hat{\mathbf{y}} = h(\mathbf{x}) = \arg\min_{\mathbf{y}} \parallel \Psi(\mathbf{y}) - \Phi(\mathbf{x})\,\mathbf{B} \parallel_2 . \tag{7}$$

B. Our lightweight formulation

The "Lightweight" MLM (LW-MLM) builds a regularized system by pattern to impose sparseness, not by selection but by using weighted information in the model. Unlike other MLM variants, LW-MLM does not work at the error but in the complexity term. LW-MLM has the following cost function:

$$\min_{\mathbf{B}} \ \mathcal{J}_{\text{LW}}(\mathbf{B}, \mathbf{P}) = \ \parallel \mathbf{DB} - \mathbf{\Delta} \parallel_{\mathcal{F}}^2 + \parallel \mathbf{PB} \parallel_{\mathcal{F}}^2 \tag{8}$$

which yields the following solution:

$$\hat{\mathbf{B}}_{\text{LW}} = (\mathbf{D}^{\intercal}\mathbf{D} + \mathbf{P}^{\intercal}\mathbf{P})^{-1}\mathbf{D}^{\intercal}\mathbf{\Delta}, \tag{9}$$

where $\mathbf{P}$ is a regularization matrix based on the sample regularization factor. The role of $\mathbf{P}$ here is the main proposal of our work. Although, it sounds inappropriate having a hyper-parameter $\mathbf{P} \in \mathbb{R}^{N \times N}$, one can derived it by a vector $\mathbf{p} \in \mathbb{R}^N$ by adopting $\mathbf{P}$ as a diagonal matrix, i.e., $\mathbf{P} = \text{diag}(\mathbf{p})$.

C. Speeding up the out-of-sample prediction

The "lightweight" in LW-MLM is not just related to the smaller coefficient values in $\mathbf{B}$ but also the speedup procedure we adopt in the out-of-sample prediction. Since we employ all samples in the learning algorithm, we believe that LW-MLM learns the whole known geometric structure, i.e., the domain knowledge is "fully" represented in $\mathbf{B}$. Such an assumption encourages us to discard some components (RPs) in the out-of-sample prediction procedure, providing a more compact one since most of the RP projections will be close to zero. First, let us define the discard function $\kappa : \mathbb{R}^N \to \mathbb{R}^K$ as

$$\kappa(\mathbf{a}) = (a_{i_1}, a_{i_2}, \ldots, a_{i_K})^{\intercal}, \tag{10}$$

where $i_1, i_2, \ldots, i_K, \ldots, i_N$ form a random permutation of $\{1, \ldots, N\}$. Now, the location of $\hat{\mathbf{y}}$ can be estimated by minimizing the following objective function

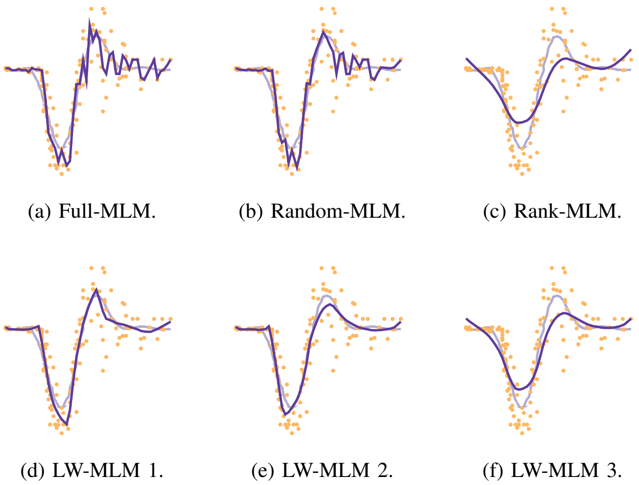
$$\hat{\mathbf{y}} = h(\mathbf{x}) = \arg\min_{\mathbf{y}} || \, \kappa\left(\Psi(\mathbf{y}) - \Phi(\mathbf{x})\,\mathbf{B}\right) ||_2 . \tag{11}$$

We discussed details concerning such speedup in the experiments where we showed the relationship between the number of components employed and the prediction error. In the end, by employing such a fashion, we achieved up to 5% (i.e., $K = \lfloor 0.05 \times N \rfloor$) of dataset reduction.

D. Validation

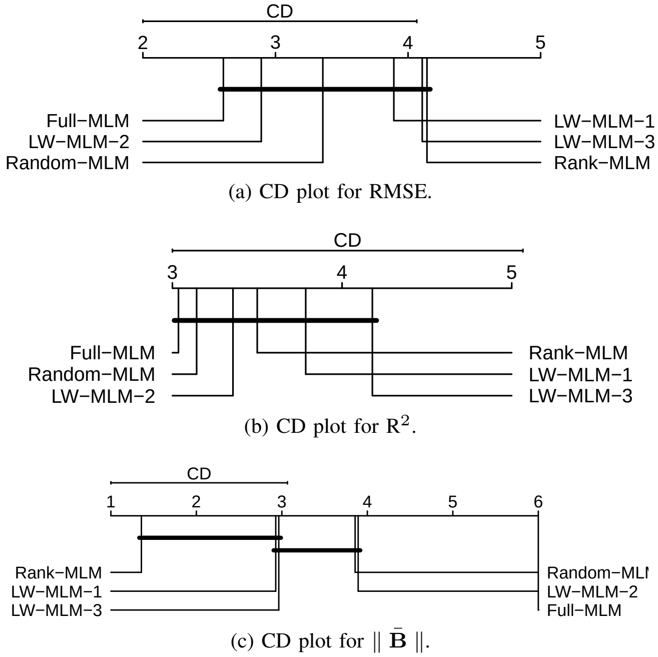
We assessed LW-MLM's performance alongside three other MLM variants: Full-MLM, Random-MLM, and Rank-MLM [10]. We investigated how LW-MLM behaves by adopting three different mechanisms of generating $\mathbf{P}$, deriving three other LW-MLM versions: LW-MLM-1, LW-MLM-2, and LW-MLM-3.

One can notice in Fig. 4 a case where regularization and overfit add difficulties to the learning process. Both Full-MLM and Random-MLM presented aspects of overfitting because of the heteroscedasticity scenario. Thus, adopting all (Full-MLM) or some (Random-MLM) RPs is insufficient, indicating the need for regularization. Concerning the regularized variants, namely, Rank-MLM and LW-MLM, one can notice smoother functions. However, the homoscedasticity behavior embedded into Rank-MLM did not achieve a proper fit in contrast to LW-MLM variants. From such a finding, we genuinely believe that regularization and heteroscedasticity are beneficial to the model.


(a) Full-MLM. (b) Random-MLM. (c) Rank-MLM.

(d) LW-MLM 1. (e) LW-MLM 2. (f) LW-MLM 3.

Fig. 4. MLM variants for Mcycle dataset.



(a) CD plot for RMSE.

(b) CD plot for R^2.

(c) CD plot for $\parallel \bar{\mathbf{B}} \parallel$.

Fig. 5. Critical Difference plots for the black-box experiments.

The $\parallel \bar{\mathbf{B}} \parallel$ stands for values were scaled between 0 and 1 by the solution of Full-MLM. In this case, $\mathbf{B}$ from (5) acts as an upper bound because in other models either $\mathbf{D}$ and/or $\mathbf{\Delta}$

are not squared matrices nor they present any regularization factor.

Moreover, to assess the *lightweight* pillar in LW-MLM, we conducted an experiment discarding some RPs in the out-of-sample prediction, thus, analyzing how such a discard influences the error. For that, we vary the quantity of RPs from 2 points and then increase it by multiples of 5% of the actual dataset size, see Fig. 6.

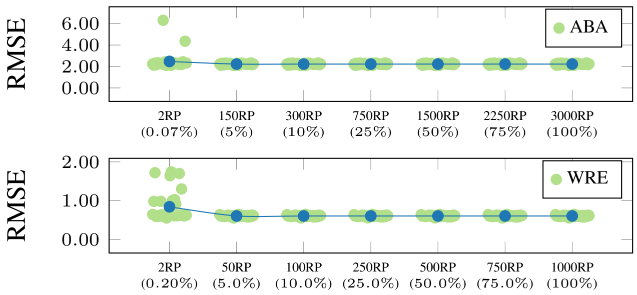


Fig. 6. RMSE for ABA and WRE datasets. Adapted from [5].

V. Contribution 4: Class-Corner Lightweight Minimal Learning Machine

Our last contribution is the LW-MLM formulation for classification tasks. To fulfill the LW-MLM requirement of providing $\mathbf{P}$ before training, we computed a distance factor as regularization cost for each sample concerning the closest class corner yielded by CCIS. Such a proposition resulted in the class-corner samples with higher regularization costs, while the samples far from the corners will be penalized less.

Let $\mathcal{PS} = \mathrm{CCIS}(\mathcal{D}, 0)$ be the set of sample pairs yielded by CCIS, then, we defined the Nearest Corner Distance $\mathrm{NCD}(\cdot)$ of a given sample $\mathbf{x}$ as:

$$\mathrm{NCD}_{\mathcal{PS}}(\mathbf{x}) = \min \left\{ ||\mathbf{x} - \mathbf{x}_j||_2 \right\}, \forall \mathbf{x}_j \in \mathcal{PS}. \qquad (12)$$

Then, we defined the maximum cost by class-corner nearness as the maximum distance of a query sample to the class-corners for all samples in $\mathcal{D}$:

$$\zeta = \max \left\{ \mathrm{NCD}_{\mathcal{PS}}(\mathbf{x}_i) \right\}, \forall \mathbf{x}_i \in \mathcal{D}, \qquad (13)$$

so that finally, we can derive the cost by class-corner nearness of a given sample as:

$$\varsigma(\mathbf{x}) = \zeta - \mathrm{NCD}_{\mathcal{PS}}(\mathbf{x}). \qquad (14)$$

To achieve a "Lightweight" fashion in classification tasks, one must present a way to produce a $\mathbf{P}$ for LW-MLMs. Here, we chose to regularize each sample by the complement of its closeness to the corners. To do it so, we execute CCIS, as usual, to get $\mathcal{PS}$ and employ the cost by class-corner nearness as the regularization values in $\mathbf{P}$, i.e.,

$$P_{i,j} = \begin{cases} \varsigma(\mathbf{x}_i) & \text{if } i = j; \\ 0 & \text{otherwise.} \end{cases} \qquad (15)$$

A. Learning algorithm and out-of-sample prediction

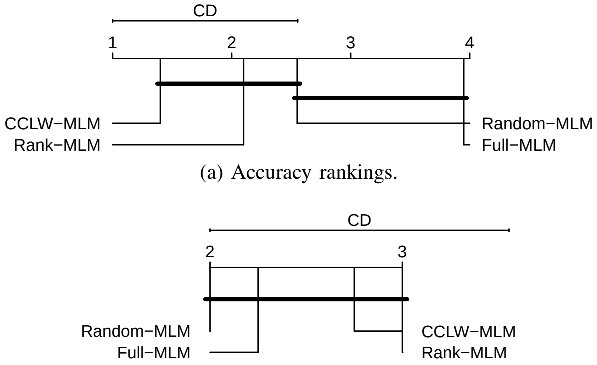
By taking advantage of the known output space, i.e., the already known labels, we avoided the multilateration procedure during prediction by just replacing it by directly applying the known labels to the cost function. Firstly, let us define $\mathcal{Y}^\star$ as the set with labels (being One-Hot-Encoded) from C classes, then we rewrite the out-of-sample prediction as:

$$\hat{\mathbf{y}} = h(\mathbf{x}) = \underset{\mathbf{y}_c \in \mathcal{Y}^\star}{\arg\min} \, || \, \Psi(\mathbf{y}_c) - \Phi(\mathbf{x}) \, \mathbf{B}||_2. \qquad (16)$$

B. Validation

We investigated how CCLW-MLM behaves with respect to accuracy and sparseness. Again, we reported the average accuracy (ACC) and sparseness via the scaled $||\mathbf{B}||_\mathcal{F}$ over 30 independent realizations in such a comparison, see Fig. 7.

Regarding accuracy, all models are seen as equivalents. Moreover, we support that model regularization is advantageous for MLM accuracy since Rank-MLM and CCLW-MLM ranked best. On the other hand, when analyzing the norm rank, we see a different perspective. There, we noticed two groups of equivalence: one with Full-MLM, Random-MLM, and Rank-MLM, and the second with only Rank-MLM and CCLW-MLM. Thus, the way we define regularization impacts the final result.



(a) Accuracy rankings.

(b) Normalized norm rankings.

Fig. 7. Critical Difference plots for the black-box experiments.

Next, we empirically assessed the decision boundary quality. In this set of experiments, we chose three datasets from the toy problems $\in \mathbb{R}^2$, see Fig. 8.

In Fig. 8, one can see all variants produced proper decision boundaries able to separate the data at their best. In the cases where there are slight class overlapping and some outliers in the data, we noticed some overfitting. However, CCLW-MLM is the one with less complex boundaries. Therefore, we support it did not sacrifice the generalization performance while keeping higher sparsity scores (less complexity), achieving higher ranks than other variants that also employ some regularization. Thus, becoming a desirable formulation to deal with classification tasks.

VI. Shortcomings

The class-corner concept highly relies upon distance computations. Therefore, we might deal with the effects of the curse

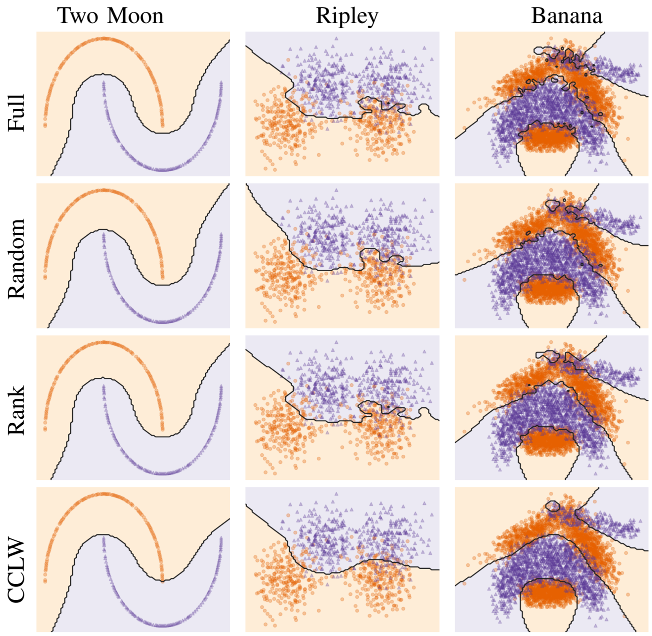


Fig. 8. Results for 2D datasets *vs*. MLM variants.

of dimensionality because the Euclidean distance becomes meaningless as the data dimension increases significantly. Also, because we do not discard any samples during training, the memory and training costs might be prohibitive for some problems. Thus, some other strategies regarding instance selection might take place in the formulation.

VII. Concluding remarks

The contribution presented in this thesis covers four contributions and investigated the model complexity reduction in two Instance-based learners: The Least-Squares Support Vector Machine and the Minimal Learning Machine.

The common idea behind all the solutions is to reduce the complexity in Instance-based learners from instance selection, treating it as a regularization task. Thus, excluding our first contribution, an instance selection algorithm, we modified the design of such LSSVM and MLM algorithms to embed such a complexity reduction.

We carried out some experiments to evaluate each contribution's different aspects, investigating how they behave concerning the following aspects: the prediction error, the goodness-of-fit of estimated vs. measured values, the model complexity, the influence of the parameters, and the learned models' empirical visual analysis.

Even though our contributions strongly rely on distance computations, thus suffering from the Dimensionality curse, they consistently outperformed the other models in artificial and real-world scenarios. This thesis's apparent unfolding is to directly apply metric learning methods to derive more algorithms with consistent hypotheses.

VIII. Publications

We published the results of this thesis in *Neurocomputing* [4], [5]. We also would like to highlight that Saulo A. F. Oliveira also contributed as the main author in *Computer Vision and Image Understanding* [11] and co-authored two other works in *Applied Soft Computing* [12] and *Soft Computing* [13]. All above mentioned journals have Qualis A1.

Acknowledgments

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) – Finance Code 001.

References

[1] D. AHA, D. KIBLER, and M. ALBERT, "Instance-Based Learning Algorithms," *MACHINE LEARNING*, vol. 6, no. 1, pp. 37–66, JAN 1991.

[2] P. Norvig and S. Russell, *Inteligência Artificial: Tradução da 3a Edição*. Elsevier Brasil, 2017. [Online]. Available: https://books.google.com.br/books?id=BsNeAwAAQBAJ

[3] D. MACKAY, "Bayesian Interpolation," *Neural Computation*, vol. 4, no. 3, pp. 415–447, MAY 1992.

[4] S. A. F. Oliveira, J. P. P. Gomes, and A. R. Rocha Neto, "Sparse Least-Squares Support Vector Machines via Accelerated Segmented Test: A dual approach," *NEUROCOMPUTING*, vol. 321, pp. 308–320, DEC 10 2018.

[5] J. A. Florencio, V, S. A. F. Oliveira, J. P. P. Gomes, and A. R. Rocha Neto, "A new perspective for Minimal Learning Machines: A lightweight approach," *NEUROCOMPUTING*, vol. 401, pp. 308–319, AUG 11 2020.

[6] E. Rosten and T. Drummond, "Machine learning for high-speed corner detection," in *COMPUTER VISION - ECCV 2006 , PT 1, PROCEEDINGS*, ser. LECTURE NOTES IN COMPUTER SCIENCE, Leonardis, A and Bischof, H and Pinz, A, Ed., vol. 3951, no. 1. Graz, Austria: Adv Comp Vis; Graz Univ Technol; Univ Ljubljana, 2006, pp. 430–443.

[7] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *JOURNAL OF MACHINE LEARNING RESEARCH*, vol. 7, pp. 1–30, JAN 2006.

[8] A. H. de Souza Junior, F. Corona, G. A. Barreto, Y. Miche, and A. Lendasse, "Minimal Learning Machine: A novel supervised distance-based approach for regression and classification," *NEUROCOMPUTING*, vol. 164, pp. 34–44, SEP 21 2015, 12th International Work-Conference on Artificial Neural Networks (IWANN), Puerto de la Cruz, SPAIN, JUN 12-14, 2013.

[9] E. Niewiadomska-Szynkiewicz and M. Marks, "Optimization schemes for wireless sensor network localization," *IntINTERNATIONAL JOURNAL OF APPLIED MATHEMATICS AND COMPUTER SCIENCE*, vol. 19, no. 2, pp. 291–302, Jun. 2009. [Online]. Available: https://doi.org/10.2478/v10006-009-0025-3

[10] A. S. C. Alencar, W. L. Caldas, J. P. P. Gomes, A. H. de Souza Junior, P. A. C. Aguilar, C. Rodrigues, W. Franco, M. F. de Castro, and R. M. C. Andrade, "MLM-Rank: A Ranking algorithm based on the Minimal Learning Machine," in *2015 BRAZILIAN CONFERENCE ON INTELLIGENT SYSTEMS (BRACIS 2015)*. Rio Grande do Norte: Soc Brasileira Comp SBC; Univ Federal do Rio Grande do Norte, 2015, pp. 305–309, 4th Brazilian Conference on Intelligent Systems (BRACIS), Natal, BRAZIL, NOV 04-07, 2015.

[11] S. A. F. Oliveira, S. S. A. Alves, J. P. P. Gomes, and A. R. Rocha Neto, "A bi-directional evaluation-based approach for image retargeting quality assessment," *COMPUTER VISION AND IMAGE UNDERSTANDING*, vol. 168, no. SI, pp. 172–181, MAR 2018.

[12] D. P. P. Mesquita, J. P. P. Gomes, L. R. Rodrigues, S. A. F. Oliveira, and R. K. H. Galvao, "Building selective ensembles of randomization based neural networks with the successive projections algorithm," *APPLIED SOFT COMPUTING*, vol. 70, pp. 1135–1145, SEP 2018.

[13] E. d. S. Reboucas, R. C. P. Marques, A. M. Braga, S. A. F. Oliveira, V. H. C. de Albuquerque, and P. P. Reboucas Filho, "New level set approach based on parzen estimation for stroke segmentation in skull ct images," *SOFT COMPUTING*, vol. 23, no. 19, SI, pp. 9265–9286, OCT 2019.