ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

Feature Selection for Multi-label Classification with Minimal Learning Machine

Joakim Linja* and Joonas Hämäläinen* and Tommi Kärkkäinen*

University of Jyväskylä - Faculty of Information Technology
P.O. Box 35, FI-40014 University of Jyväskylä - Finland

Abstract. Multi-label classification problems, where more than one class can be active in a single instance, generalize the conventional single-label cases. In this article, we continue the research track documented in [1, 2], where the Minimal Learning Machine (MLM) was generalized into multi-label problems with competitive results compared to other state-of-the-art techniques. Our current interest is to consider whether we can reduce the complexity of the distance-based regression model in the MLM by performing feature selection. For this purpose, an existing feature selection filter technique is generalized to multi-label problems. Experimental results confirm that the proposed technique provides a useful ranking, which allows one to reduce the number of active features without jeopardizing the quality of the multi-label MLM classifier.

1 Introduction

In the multi-label classification (MLC) problems, the restriction of a single active class is relaxed to consider multiple active labels, which can be represented as a bipartition of the labels into relevant and irrelevant ones for each instance. The popularity of the MLC research field has been steadily increasing [2], and many methods and algorithms have been proposed over the years [3]. In principle, the MLC techniques can be divided into problem transformation (PT) and algorithm adaptation (AA) methods, where the MLC problem is transformed to a set of single-label problem(s) with PT and a direct modification of a learning algorithm is performed in AA [4]. The AA methods can be further categorized as first- (label-by-label style), second- (pairwise relations between labels), or higher-order (more complex linkages like one-to-many relations between labels) methods [5].

The minimal learning machine (MLM) is a supervised machine learning method that integrates the construction of a distance-regression model with a multilateration step to estimate the output from a set of predicted distances [6, 7]. Its extension to multi-label (ML) problems, ML-MLM, was preliminary proposed in ESANN 2021 [1] and, more recently, generalized and thoroughly experimented in [2]. In essence, the ML-MLM uses inverse distance weighting [8] to modify the second step of the basic MLM in order to establish a ranking of the predicted output labels with a threshold to identify the active ones. Experimental comparisons in [2] to a random forest-based ML method [9], which

*This work was supported by the Academy of Finland through the grant 351579. We acknowledge the work of Marko Niemelä regarding the nanclustering toolbox.

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

has been one the best-performing methods in extensive experimental evaluations like [3], concluded highly competitive performance of the ML-MLM.

Feature Selection (FS) refers to the identification and use of only a subset of the original features in the construction of a supervised model. Various suggestions for the multi-label FS were reviewed in [2, Section 2.1] and in [10, Section 2.6]. In general, FS is a search problem, and many FS algorithms perform an iterative identification of a reduced set of features (see the umbrella review on FS in [10, Section 2]). However, the filter FS techniques, which are independent of the predictive model, may provide direct scoring and ranking of the features with higher computational efficiency compared to the model-specific wrappers [10]. In this paper, we follow this line of development by augmenting the Kruskal-Wallis (KW) test statistics-based scoring of features [11, 12, 13] with a clustered multi-label output data. Use of clustering, which utilizes the toolbox from [14], is one way to perform a problem transformation of MLC to a single-label case, but here it is not used on the whole problem level but only to enable the application of the KW-based FS filter with the ML-MLM.

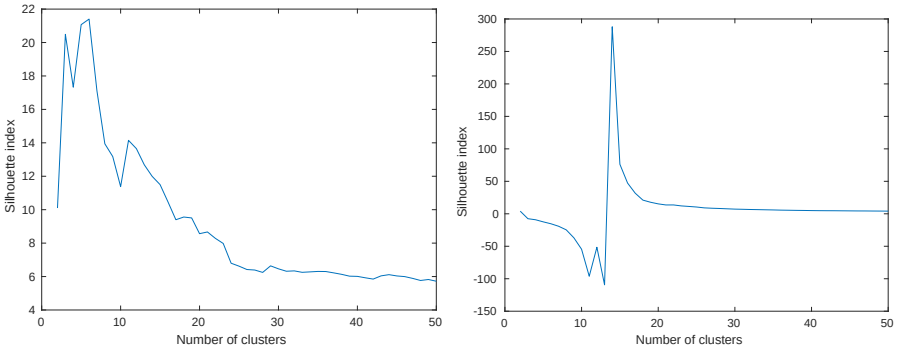
2 Methods

Define the set of instances as $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^N$, where $N = |\mathbf{X}|$ and $\mathbf{x}_i \in \mathbb{R}^M$. Let $\mathbf{Y} = \{\mathbf{y}_i\}_{i=1}^N$ be the set of label vectors so that $\mathbf{y}_i \in \{0,1\}^L$, where $\mathbf{y}_{i(j)} = 1$ associates that an instance $\mathbf{x}_i$ belongs to a class j.

The ML-MLM's [2] training phase applies the ordinary least squares (OLS) formulation to construct a distance mapping $\mathbf{B}$ between the input and the label space distance matrices $\mathbf{D_x}$ and $\mathbf{D_y}$. This solution is given by $\mathbf{B} = (\mathbf{D_x}^T\mathbf{D_x} + \alpha\mathbf{I})^{-1}\mathbf{D_x}^T\mathbf{D_y}$, where $\mathbf{D_x}$ is $N \times K$ matrix which holds distances between $\mathbf{X}$ and the input space reference points $\mathbf{R} \subset \mathbf{X}$ [7], $\mathbf{D_y}$ is $N \times N$ full distance matrix of $\mathbf{Y}$, and α is a small positive constant. In the ML-MLM's prediction phase, the label space distance estimates $[\rho_1, ..., \rho_N]$ for a new instance $\mathbf{x}^*$ are given by $[\ \|\mathbf{x}^* - \mathbf{r}_1\|, ..., \|\mathbf{x}^* - \mathbf{r}_K\|]\mathbf{B}$, where r_i refers to the reference point from $\mathbf{R}$. A scoring of the labels is then given by $\sum_{i=1}^N w_i\mathbf{y_i}$, where $w_i = \rho_i^{-P} / \sum_{i=1}^N \rho_i^{-P}$ when $\rho_i > 0$, and, $w_i = 1/\sum_{i=1}^N \rho_i^{-P}$ when $\rho_i = 0$. The power parameter $P > 0$ is selected based on the ranking loss statistic. Finally, for the selected power parameter value, the bipartition of the relevant and irrelevant labels is thresholded with the $t > 0$ parameter, which is selected by matching the out-of-sample label cardinality to the training set label cardinality.

In the basic ML-MLM, the input distance matrix $\mathbf{D_x}$ is computed using all features. However, a separability score of features with single-label labeling can be computed using the Kruskal-Wallis (KW) test [12, 13]. Hence, in order to generate a single-label approximation (categorization) of the multi-label output data, we apply the iterative kcentroids clustering algorithm with the Cityblock-distance and the Silhouette cluster validation index [15], otherwise using the default settings of the toolbox [14]. The search interval for the number of clusters C ranged from 2 to $min(50, Y_{\text{uniq}} - 1)$, where Y_{uniq} denotes the number of unique observations in Y. We used $Y_{\text{uniq}} - 1$ to ensure that some clustering error always

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.



(a) COREL5K (b) BIBTEX

Fig. 1: Silhouette index for two datasets

remained after the labeling. The data-specific C was determined by the smallest value of the Silhouette index. Examples for the determination of the number of clusters using the Silhouette index are presented in Figure 1 for datasets *COREL5K* and *BIBTEX*. As can be seen from Figure 1, the number of clusters for *COREL5K* was 50 clusters since the minima was found at the end, whereas for *BIBTEX*, the minima was found at 13 clusters.

The labels from clustering were used to score the features using the test statistics value of the Kruskal-Wallis H-test [11, 12, 13]. These values served as the ranking values of the Feature Importance Detector (FID) [16], which performed the actual feature selection (see Figure 2 for functionality illustration). Two cutoff values for the FID algorithms, 0.1 and 0.01, were selected for the experiments based on the experiments in [16].

3 Experiments and results

The experiments were conducted using Matlab by integrating and extending the existing methods and their implementations presented in [2, 14, 16] (links to Github-repositories are given in the corresponding reference entries). We evaluated the proposed method with two bipartition-based metrics, Accuracy (ACC) and Hamming Loss (HL), and two ranking-based metrics, Ranking Loss (RL) and Coverage (Cov) [2, 3]. In addition to these MLC metrics, we evaluated the effect of feature selection on the model complexity by assessing the relative change of the number of features (fN).

A set of MLC datasets that were previously used in [2] were used in the experiments. The used datasets are presented in Table 1 and were obtained from [17]. We removed duplicate observations and constant features and scaled the data to a range of $[0, 1]$ using MinMax-scaling.

The ML-MLM [2] was studied by comparing the baseline metrics to the values obtained after the proposed feature selection method. An example illustration

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

Dataset	N	N_{tst}	M	L	C	Dataset	N	N_{tst}	M	L	C
1. MEDICAL	333	645	1449	45	50	5. YEAST	1500	917	103	14	50
2. EMOTIONS	391	202	72	6	25	6. COREL5K	4500	500	499	374	50
3. ENRON	1123	579	1001	53	50	7. BIBTEX	4880	2515	1836	159	13
4. SCENE	1211	1196	294	6	13	8. TMC2007	21519	7077	500	22	50

Table 1: Dataset characteristics. N and N_{tst} are the number of observations in the training set and in the test set, respectively. M is the number of features, L is the number of outputs, and C is the number of clusters used to cluster the output.

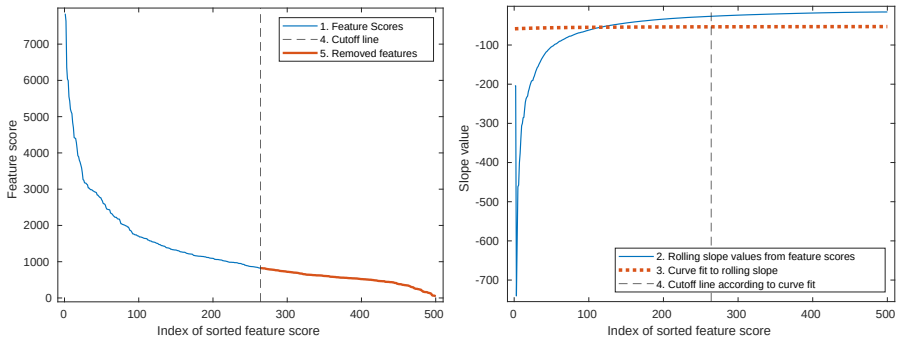


Fig. 2: An illustration of the function of the FID for dataset *TMC2007* with cutoff value 0.1. See the expanded illustration in [16].

of how the FID selects features is presented in Figure 2 for one of the datasets. The results for cutoff values 0.1 and 0.01 are presented in Table 2. The baseline vs. the feature selected results are presented using the four metrics and the number of features fN in the corresponding columns of Table 2. Each dataset is presented with the baseline values (BL), feature selected values (FS), and the difference between the two (Diff). The sign of Diff is set with the assumption that the difference is a positive number if the metric has been improved after the feature selection. The number is negative when the feature selection does not improve upon the baseline value. In the case of the feature number column, fN, the Diff-section is the ratio of the selected features to the original features.

Information regarding the used datasets can be gleaned from the results presented in Table 2. Altogether the proposed feature scoring, ranking, and selection technique showed very promising results. For majority of the datasets (MEDICAL, EMOTIONS, YEAST, BIBTEX, TMC2007), we were able to reduce the number of active features by 20%-60% and still maintain the same result quality than for the baseline with all features. With these datasets, the more instances we seem to have the more features we can drop without sacrificing quality. For two of the datasets (ENRON and SCENE), it was necessary to use the smaller cutoff value 0.01 to again end up with the similar result quality than with the baseline with all features. And this selection also yielded to a significant reduction on the number of features. For ENRON, apparently, one could have tested even a smaller cutoff value. Finally, for COREL5K, we obtained almost equal

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

c		cutoff = 0.1					cutoff = 0.01				
Set		ACC	HL	RL	Cov	fN	ACC	HL	RL	Cov	fN
1	BL	0.766	0.015	0.026	1.415	894	0.766	0.015	0.026	1.415	894
	FS	0.765	0.015	0.027	1.444	707	0.766	0.015	0.026	1.417	873
	Diff	-0.002	0.000	-0.001	-0.028	0.79	-0.001	-0.000	-0.000	-0.002	0.98
2	BL	0.604	0.191	0.139	1.733	72	0.604	0.191	0.139	1.733	72
	FS	0.601	0.189	0.137	1.738	59	0.601	0.191	0.142	1.738	70
	Diff	-0.002	0.002	0.002	-0.005	0.82	-0.002	-0.001	-0.002	-0.005	0.97
3	BL	0.404	0.046	0.122	203.806	642	0.404	0.046	0.122	203.806	642
	FS	0.231	0.061	0.159	236.425	72	0.356	0.049	0.126	207.016	373
	Diff	-0.174	-0.016	-0.037	-32.619	0.11	-0.048	-0.004	-0.004	-3.210	0.58
4	BL	0.768	0.082	0.063	0.418	294	0.768	0.082	0.063	0.418	294
	FS	0.587	0.142	0.129	0.755	49	0.743	0.089	0.073	0.463	202
	Diff	-0.181	-0.060	-0.066	-0.337	0.17	-0.025	-0.008	-0.009	-0.045	0.69
5	BL	0.573	0.194	0.165	5.985	103	0.573	0.194	0.165	5.985	103
	FS	0.566	0.198	0.169	6.069	59	0.568	0.196	0.166	6.021	96
	Diff	-0.007	-0.004	-0.004	-0.084	0.57	-0.005	-0.003	-0.001	-0.036	0.93
6	BL	0.199	0.014	0.115	99.743	499	0.199	0.014	0.115	99.743	499
	FS	0.191	0.014	0.122	106.016	180	0.195	0.014	0.116	100.705	429
	Diff	-0.008	-0.000	-0.007	-6.273	0.36	-0.004	-0.000	-0.001	-0.962	0.86
7	BL	0.407	0.018	0.084	25.640	1836	0.407	0.018	0.084	25.640	1836
	FS	0.397	0.018	0.088	26.465	888	0.406	0.019	0.083	25.590	1673
	Diff	-0.010	0.000	-0.004	-0.825	0.48	-0.000	-0.000	0.000	0.050	0.91
8	BL	0.994	0.001	0.000	1.207	500	0.994	0.001	0.000	1.207	500
	FS	0.992	0.001	0.000	1.211	263	0.995	0.001	0.000	1.207	462
	Diff	-0.003	-0.000	-0.000	-0.004	0.53	0.000	0.000	-0.000	-0.000	0.92

Table 2: Results

quality: visible increase in coverage means that we needed slightly more top-scored predicted labels to ensure a ground truth label with the reduced feature model compared to the baseline full-feature model.

4 Conclusions

In this paper, we studied the application of feature selection for the ML-MLM in the context of multi-label classification. We performed feature selection by integrating the use of Kruskal-Wallis test statistics, clustering-based single-label generation, and the Feature Importance Detector (FID) with two different cutoff values, 0.1 and 0.01. For all of the tested datasets, this feature selection strategy was able to reduce the number of features for the multi-label classifier without losing performance in multi-label ranking or classification.

Based on the results, the proposed approach can provide feasible feature importance scoring for multi-label classifiers. However, a varying performance regarding different FID cutoff values suggests that for keeping the classification performance on the same level as the full feature model, the thresholding of the proposed metalabeling-based scoring requires more thorough tailoring. For instance, the behaviour of the curve fit in Figure 2 suggests that the FID would benefit from normalization.

For the future work, another approach would be to leverage the single-label clustering based feature importance scoring as weights for the features in the ML-MLM to improve its classification performance. The ML-MLM uses a similar approach with inverse distance weighting for the label vector weighting, where the full training label data contributes to the final label scoring with varying contributions.

The feature selection algorithm, FID, was originally used with the Mean Absolute Sensitivity (MAS) -based feature scoring [10], which was applied as

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

wrapper for regression datasets. FID is quite dependent on the feature scoring, which brings the question of how the results would have changed with another scoring method. It is something to be further tested.

References

[1] J. Hämäläinen, P. Nieminen, and T. Kärkkäinen. Instance-based multi-label classification via multi-target distance regression. In *Proceedings of the European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning - ESANN 2021*, pages 653–658, 2021.

[2] J. Hämäläinen, A. H. de Souza Júnior, C. L. Mattos, J. P. Gomes, and T. Kärkkäinen. Minimal learning machine for multi-label learning. 2023. arXiv:2305.05518.

[3] J. Bogatinovski, L. Todorovski, S. Džeroski, and D. Kocev. Comprehensive comparative study of multi-label classification methods. *Expert Systems with Applications*, 203:117215, 2022.

[4] G. Tsoumakas and I. Katakis. Multi-label classification: An overview. *International Journal of Data Warehousing and Mining (IJDWM)*, 3(3):1–13, 2007.

[5] M.-L. Zhang and Z.-H. Zhou. A review on multi-label learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*, 26(8):1819–1837, 2013.

[6] 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*, 164:34–44, 2015.

[7] J. Hämäläinen, A. S. Alencar, T. Kärkkäinen, C. L. Mattos, A. H. Souza Júnior, and J. P. Gomes. Minimal learning machine: Theoretical results and clustering-based reference point selection. *The Journal of Machine Learning Research*, 21, 2020.

[8] D. Shepard. A two-dimensional interpolation function for irregularly-spaced data. In *Proceedings of the 1968 23rd ACM national conference*, pages 517–524, 1968.

[9] D. Kocev, C. Vens, J. Struyf, and S. Džeroski. Ensembles of multi-objective decision trees. In *European conference on machine learning*, pages 624–631. Springer, 2007.

[10] J. Linja, J. Hämäläinen, P. Nieminen, and T. Kärkkäinen. Feature selection for distance-based regression: An umbrella review and a one-shot wrapper. *Neurocomputing*, 518:344–359, 2023. Source codes available in `https://gitlab.jyu.fi/hnpai-public/extreme-minimal-learning-machine/`.

[11] W. H. Kruskal and W. A. Wallis. Use of ranks in one-criterion variance analysis. *Journal of the American statistical Association*, 47(260):583–621, 1952.

[12] A. Cord, C. Ambroise, and J.-P. Cocquerez. Feature selection in robust clustering based on laplace mixture. *Pattern Recognition Letters*, 27(6):627–635, 2006.

[13] M. Saarela, J. Hämäläinen, and T. Kärkkäinen. Feature ranking of large, robust, and weighted clustering result. In *Proceedings of the 21st Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining - PAKDD 2017*, pages 96–109, 2017.

[14] M. Niemelä, S. Äyrämö, and T. Kärkkäinen. Toolbox for distance estimation and cluster validation on data with missing values. *IEEE Access*, 10:352–367, 2021. Source codes available in `https://github.com/markoniem/nanclustering_toolbox`.

[15] P. J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987.

[16] J. Linja, J. Hämäläinen, A. Pihlajamäki, P. Nieminen, S. Malola, H. Häkkinen, and T. Kärkkäinen. Knowledge discovery from atomic structures using feature importances, 2023. arXiv:2303.09453.

[17] G. Tsoumakas, E. Spyromitros-Xioufis, J. Vilcek, and I. Vlahavas. Mulan: A java library for multi-label learning. *Journal of Machine Learning Research*, 12:2411–2414, 2011.