A Simple Method of Solution For Multi-label Feature Selection

Jayaraman K Valadi
Shiv Nadar University, Gautum Budha Nagar, U.P. ,201314
Pune -411007
Pune, India
jayaraman.valadi@snu.edu.in

Prasad T Ovhal
Centre for Modeling And Simulation, Savitribai Phule Pune University, Pune -411007
Pune, India
prasadovhal99@gmail.com

Kunal J Rathore
Centre for Modeling And Simulation, Savitribai Phule Pune University,
Pune -411007
Pune, India
kunalrathore326@gmail

***Abstract*—Multi-label learning has been a topic of research interest in multimedia, text & speech recognitions, music, image processing, information retrieval etc. In Multi-label classification (MLC) each instance is associated with a set of multiple class labels. Like other machine learning algorithms, data preprocessing plays an key role in MLC. Feature selection is an important preprocessing step in MLC, due to high dimensionality of datasets and associated computational costs.**

Extracting the most informative features considerably reduces the computational loads of MLC. Most of the Multi-label feature selection algorithms available in literature involve conversions to multiple single labeled feature selection problems. We proposed an efficient modification of a recent multi-label feature selection algorithm [1] available in literature. Our algorithm consists of two steps: in the first step we decompose the output label space into lower dimensions using simple matrix factorization method; subsequently we employ feature selection process in the decoupled reduced space. Our simulations with real world datasets reveal the efficiency of proposed framework.

Keywords—Feature selection, Attribute ranking, Multi-label learning, Dimensionality reduction, Matrix Factorization, Multi-label Classification.

I. INTRODUCTION

MLC problems have been found to be useful in diverse fields like gene & protein function classification [15], music categorization [16], Text classification [12], identification of drug side effects [14] and semantic scene classification [6]. In MLC each instance is associated with a set of labels $Y \subseteq L$; where L is set of disjoint labels. For example, a peptide sequence can simultaneously have anti-viral, anti-fungal & anti-bacterial activities, another sequence may have only anti-fungal and anti-bacterial activities and so on. Also a drug can have multiple side effects or an image can have multiple scenes.

Currently for MLC two major methods are available as discussed below:

Problem transformation methods: This approach employs conversion of multi-label problems to known learning methods [10]. These methods include Binary Relevance [8], Label Powerset, Classifier Chains, Calibrated Label Ranking and Random k-label sets [9].

Algorithm adaptation methods: While the earlier method converts data to suit available algorithms, this approach converts algorithms to suit data. This method adapts popular learning techniques to deal with multi-label data directly. Some of the algorithms are first-order approaches such as ML-kNN, ML-DT and second-order approach like Rank-SVM [13].

Multi-label learning problems are usually high dimensional and computationally very intensive. So it is necessary to extract the most informative features for multi-label classification and use the selected subsets. Conventional methods of feature selection convert the problem into single label sub problems. It has been found that most of these methods do not perform well on MLC because they are not able to pick up label correlations. Original algorithm [1] employed a novel method of multi-label feature selection in which they first convert the output space into reduced dimension and subsequently use feature selection.

Their final algorithm results in an objective function which has to iteratively evolve three matrices simultaneously. In our work we have employed a simple matrix factorization method which decomposes the original output label matrix into two matrices. Subsequently the feature selection is done on the reduced decoupled lower dimensional label space.

Such a simplification results in an objective function which requires iterative convergence of final feature weight matrix. The method is simpler and computationally faster. In the following section we describe our method in detail.

The proposed improvements are outlined as follows:

- New computationally efficient attribute selection modification of original method, which is able to provide high performance attribute subset valid for all labels.

- Successful validation through simulations with three available datasets.

978-1-5386-8158-9/19/$31.00©2019IEEE

II. MULTI-LABEL FEATURE SELECTION FRAMEWORK

A. Formulation

As we have modified the approach of original method [1]. We first describe their algorithm briefly before providing details of our algorithm:

Description of original algorithm [1]:

In their framework for Multi-label feature selection (MIFS) they perform decomposition of the multi-label output space **Y** into **V** and **B**, where $\mathbf{V} \in \mathbf{R}^{n \times c}$ and $\mathbf{B} \in \mathbf{R}^{c \times k}$; **V** denotes the latent semantics [2] of the multi-label information and **B** is a coefficient matrix. So this can be interpreted as grouping output labels into semantically similar clusters. Their algorithm results in the following function:

$$\Theta(\mathbf{W}, \mathbf{V}, \mathbf{B}) = \|\mathbf{XW} - \mathbf{V}\|_F^2 + \alpha \|\mathbf{Y} - \mathbf{VB}\|_F^2 + \beta \, Tr(\mathbf{V}^T \mathbf{LV}) + \|\mathbf{W}\|_{2,1} \quad (1)$$

Where α, β terms are the regularization parameters and **W** in the given term represents the attribute coefficient matrix and every row in this matrix provide attribute importance of i^{th} feature capturing hidden information **V**. Also γ is the regularization parameter for representing the importance of $l_{2,1}$ – norm regularization [4]. The term $\boldsymbol{\beta * Tr(V^T LV)}$ constrains the neighbourhood similarity in **V** space to preserve local geometric structure in input space **X**. So minimizing the above objective function will evolve a weight matrix **W** preserving local geometry structure, capturing label correlation and accomplishing attribute selection simultaneously. They have observed that the objective function is not convex [5].

In addition, due to non smoothness of $\|\mathbf{W}\|_{2,1}$ term, $\mathbf{2Tr(W^T DW)}$ replaces $\|\mathbf{W}\|_{2,1}$; where **D** being diagonal matrix with each diagonal element represented as $\mathbf{1 / (2 * \sqrt{(W^T_{i:} W_{i:} + \varepsilon)})}$, where $\boldsymbol{\varepsilon}$ is a small positive constant.

Their final algorithm results in the following function:

$$\Theta(\mathbf{W}, \mathbf{V}, \mathbf{B}) = \|\mathbf{XW} - \mathbf{V}\|_F^2 + \alpha \|\mathbf{Y} - \mathbf{VB}\|_F^2 + \beta \, Tr(\mathbf{V}^T \mathbf{LV}) + 2\gamma \, Tr(\mathbf{W}^T \mathbf{DW}) \quad (2)$$

B. Our Simplified Approach

If constraining the neighbourhood similarity in **V** space, to the neighbourhood similarity in **X** space is relaxed, then the decomposition of output label space Y~ VB can be done (*a priori*) separately and optimization of weight matrix can be decoupled.

C. Simple Matrix Factorization

Simple matrix factorization (SMF) [17] methods have been used in many field successfully and in collaborative filtering several such algorithm are used routinely. In one such method, **Y** is decomposed in into two matrices **V** and **B**. Where **V** and **B** are initialized with some trial values and difference between the approximated value of **V, B** and **Y** is minimized using gradient descent method.

With this modification objective function for weight matrix optimization simplifies to:

$$\Theta(\mathbf{W}, \mathbf{V}, \mathbf{B}) = \|\mathbf{XW} - \mathbf{V}\|_F^2 + 2\gamma \, Tr(\mathbf{W}^T \mathbf{DW}) \quad (3)$$

Differentiating w.r.t. **W**, we have :

$$\partial\Theta / \partial\mathbf{W} = 2[(\mathbf{X}^T(\mathbf{XW} - \mathbf{V}) + \gamma \mathbf{DW})] \quad (4)$$

With this, the update rule can be written as :

$$\mathbf{W} = \mathbf{W} - \lambda_W (\partial\Theta / \partial\mathbf{W}) \quad (5)$$

Where λ_W is step size for the gradient descent update rules. Gradient descent can be used for numerically solving the above mentioned objective function.

Proposed method decouples output space decomposition from weight matrix optimization procedure. This does not require computation of additional variables, constraints and eliminates additional parameters, and this ultimately simplifies the attribute selection process. The initial decomposition of output space effectively reduces time and computational requirements due to elimination of redundancy and noisy labels.

III. ALGORITHM MULTI-LABEL FEATURE SELECTION

Input: Initialize W, Parameter γ,
Output: Top ranked attributes

1. Decompose Y ~ V B using SMF method.
2. **Repeat :**
3. $\partial\Theta / \partial W = 2[(X^T(XW - V) + \gamma DW)]$;
4. Determine step sizes λ_W with Armijo rule;
5. $W = W - \lambda_W (\partial\Theta / \partial W)$;
6. Update D ;
7. Until Convergence;
8. Return W*, Optimum weight matrix;
9. Rank attributes in decreasing order of norm of W* and return the top ranked attributes.

IV. EXPERIMENTAL STUDIES

We have made extensive simulations with three different datasets and recorded the performance of Our Modified Attribute Selection procedure (OMFS) .

A. Data sets description:

Experiments are conducted on three publicly available bench mark datasets including Scene, Yeast and Emotions datasets. Scenes dataset contains characteristics about images and their labels. A given image can belong to one or more labels. Emotions dataset contains 72 music attributes for 593 songs categorized into 6 labels of emotions. Yeast dataset contains information about a set of Yeast cells. The task in each dataset is to determine the localization compartment of each cell. Details of the datasets are described in Table 1.

TABLE I. DETAILS OF THREE BENCHMARK DATASETS:

Dataset	Instances	Attributes	Labels
Scenes	2407	294	6
Emotions	593	72	6
Yeast	2417	103	14

B. Experimental settings

To provide rational comparison, we have applied Binary relevance [8] as it has been used in earlier approach [1]. Three regularization parameters α, β, γ in MIFS were tuned with grid search strategy varying the range in { 10^{-4}, 10^{-3},10^{-2}, 0.1, 0.2, 0.4, 0.6, 0.8, 1, 10}. Support Vector Machine (SVM) algorithm is applied to train MIFS and OMFS models on the datasets. One widely accepted evaluation criteria for multi-label learning based on micro-average, F-measure [11] is used as performance measure for MLC. To get an unbiased estimate of algorithm parameters, we computed micro-average cross validation F-measure averaged over ten different random splits of five-folds. Experiments were performed with different kernels and kernel parameters and the grand average of five-fold micro-average F-measure was computed. Performance comparison was then done with existing other methods of multi-label feature selection using the best tuned parameters.

For computing five-fold CV F-measure, every time one fold is kept as test and remaining four folds to train the model. This is done for each of k-labels separately for Binary Relevance algorithm [8]. Trained model with four folds is applied on test fold each time and each example in test fold is categorised as TP, TN, FP, FN and formula given below is applied on test fold.

F-measure is defined as:

$$F-measure=\frac{2\,TP}{2\,TP+FP+FN} \tag{6}$$

Where, TP = Number of examples originally of positive class and correctly classified as positive class; FP = Number of examples originally of negative class and wrongly classified as positive class; FN = Number of examples originally of positive class and wrongly classified as negative class. Defining F-Measure averaging over all k-labels.

We used this for our experiments, defined as follows:

$$Micro-average=\frac{\sum_{i=1}^{k} 2\,TP_i}{\sum_{i=1}^{k}\left(2\,TP_i+\mathrm{FP}_i+\mathrm{FN}_i\right)} \tag{7}$$

The higher micro-average indicates the better classification performance.

Our modified method is compared with following feature selection methods along with MIFS for MLC problems:

1. **F-Score**: In Fisher Score [7] method, attributes are ranked based on higher similarity in their values of examples belonging to the same class in the dataset.

2. **RFS**: Robust Feature Selection [3] method minimizes a combined $l_{2,1}$–norm on the regularized objective function which is insensitive to outliers.

In our Experiments we have evaluated performance for various subsets of selected features ranging from 5% of total number of attributes. Fig.1- 3 shows percentage of selected features against Micro-average F-measure for Scene, Emotions and Yeast in the original method all the three matrices V, B, W have to be simultaneously optimized; in this new attribute selection procedure, due to relaxation of constrained optimization of V, B matrices problem is decoupled with optimization of weight matrix problem, and hence OMFS method becomes very much less intensive computationally, faster and easier to handle.

C. Results on datasets using discussed methods:

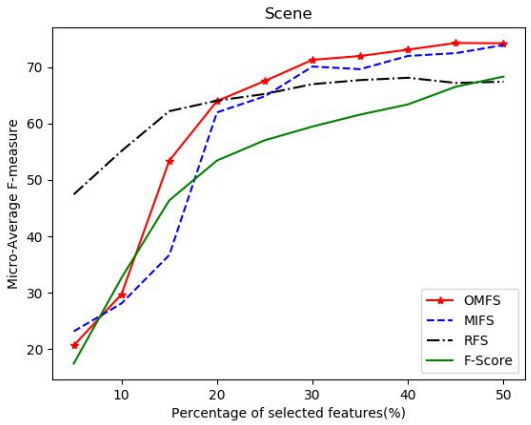

Fig. 1. Results for Scene dataset

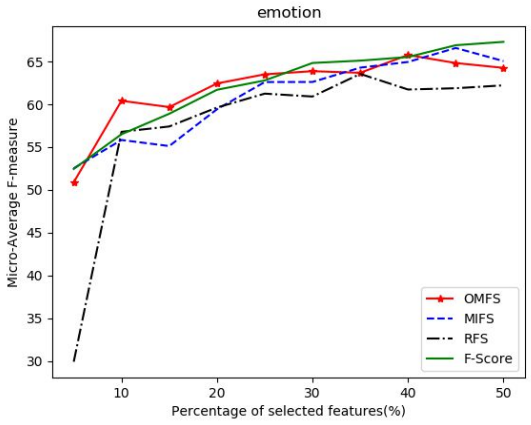

Fig. 2. Results for Emotions dataset

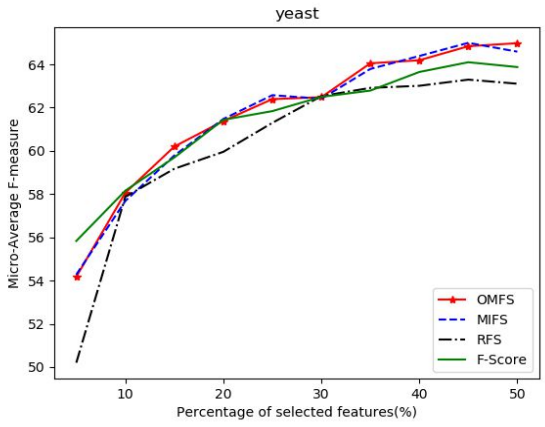

Fig.3. Results for Yeast dataset

D. *Results interpretations:*

It can be seen from figure 1. for Scene dataset OMFS method performs best in terms of micro-average F-measure. MIFS method is comparable to OMFS method and performance for all other methods is inferior. Similar results are obtained for emotions dataset with both OMFS and MIFS performing equally well again and superior to other methods. In Yeast dataset the same results hold good; performance of other method is also not lagging behind.

With the increase in number of selected attributes, it is observed that, the classification performance first tends to increase upto 50% of selected attributes and then becomes stable (not shown in the figures). OMFS works better than RFS and F-Score feature selection methods employed for multi-label attribute selection on these datasets. OMFS performs as good as original MIFS with less computation time.

V. CONCLUSION

We have presented a new modification of attribute selection with multiple labels. Due to relaxations of constraints imposed in the earlier method, our method is solving two decoupled problems separately to produce optimal weight matrix. This reduces complexity and computational run time. Additionally, the proposed method performs as good as the MIFS and superior to other mentioned methods. It can be advantageously used for handling high dimensional multi-label datasets.

VI. REFERENCES

[1] Jian, Ling, Jundong Li, Kai Shu, and Huan Liu. "Multi-Label Informed Feature Selection." In IJCAI, pp. 1627-1633. 2016.

[2] Dumais, Susan T. "Latent semantic analysis." Annual review of information science and technology 38, no. 1 (2004): 188-230.

[3] Nie, Feiping, Heng Huang, Xiao Cai, and Chris H. Ding. "Efficient and robust feature selection via joint $\ell_{2,1}$-norms minimization." In Advances in neural information processing systems, pp. 1813-1821. 2010.

[4] Zhao, Zheng Alan, and Huan Liu. Spectral feature selection for data mining. Chapman and Hall/CRC, 2011.

[5] Chang, Xiaojun, Feiping Nie, Yi Yang, and Heng Huang. "A Convex Formulation for Semi-Supervised Multi-Label Feature Selection." In AAAI, pp. 1171-1177. 2014.

[6] Boutell, Matthew R., Jiebo Luo, Xipeng Shen, and Christopher M. Brown. "Learning multi-label scene classification." Pattern recognition 37, no. 9 (2004): 1757-1771.

[7] Gu, Quanquan, Zhenhui Li, and Jiawei Han. "Generalized fisher score for feature selection." arXiv preprint arXiv:1202.3725 (2012)..

[8] Zhang, Min-Ling, Yu-Kun Li, Xu-Ying Liu, and Xin Geng. "Binary relevance for multi-label learning: an overview." Frontiers of Computer Science (2018): 1-12.

[9] Tsoumakas, Grigorios, Ioannis Katakis, and Ioannis Vlahavas. "Random k-labelsets for multilabel classification." IEEE Transactions on Knowledge and Data Engineering 23, no. 7 (2011): 1079-1089.

[10] SpolaôR, Newton, Everton Alvares Cherman, Maria Carolina Monard, and Huei Diana Lee. "A comparison of multi-label feature selection methods using the problem transformation approach." Electronic Notes in Theoretical Computer Science 292 (2013): 135-151.

[11] Wu, Xi-Zhu, and Zhi-Hua Zhou. "A unified view of multi-label performance measures." arXiv preprint arXiv:1609.00288 (2016).

[12] Zhang, Min-Ling, and Zhi-Hua Zhou. "Multilabel neural networks with applications to functional genomics and text categorization." IEEE transactions on Knowledge and Data Engineering 18, no. 10 (2006): 1338-1351.

[13] Zhang, Min-Ling, and Zhi-Hua Zhou. "A review on multi-label learning algorithms." IEEE transactions on knowledge and data engineering 26, no. 8 (2014): 1819-1837.

[14] Zhang, Wen, Feng Liu, Longqiang Luo, and Jingxia Zhang. "Predicting drug side effects by multi-label learning and ensemble learning." BMC bioinformatics 16, no. 1 (2015): 365.

[15] Schietgat, Leander, Celine Vens, Jan Struyf, Hendrik Blockeel, Dragi Kocev, and Sašo Džeroski. "Predicting gene function using hierarchical multi-label decision tree ensembles." BMC bioinformatics 11, no. 1 (2010): 2.

[16] Trohidis, Konstantinos, Grigorios Tsoumakas, George Kalliris, and Ioannis P. Vlahavas. "Multi-label classification of music into emotions." In ISMIR, vol. 8, pp. 325-330. 2008.

[17] Lee, Daniel D., and H. Sebastian Seung. "Learning the parts of objects by non-negative matrix factorization." *Nature* 401, no. 6755 (1999): 788.

[18] http://www.quuxlabs.com/blog/2010/09/matrix-factorization-a-simple-tutorial-and-implementation-in-python/

[19] Dataset source:https://sci2s.ugr.es/keel/multilabel.php