A multilabel extension of LDA based on the Gram-Schmidt orthogonalization procedure

Juan Bekios-Calfa[1] and Brian Keith[1]

Departamento de Ingeniería de Sistemas y Computación
Universidad Católica del Norte
Av. Angamos 0610, Antofagasta, Chile
{juan.bekios, brian.keith}@ucn.cl

Abstract. Multilabel classification is a generalization of the traditional unidimensional classification problem, the goal of multilabel classification is to learn a function that maps instances into a set of relevant labels. This article proposes an extension to linear discriminant analysis in the context of multilabel classification. The new method is based on Gram-Schmidt orthogonalization procedure. The theoretical basis and underlying assumptions of the new model are described and the method is experimentally evaluated on the Emotions data set for multilabel classification. The analysis of the empirical results support that this new method is competitive and in some instances superior to the baseline.

Keywords: linear discriminant analysis, Gram-Schmidt orthogonalization, multilabel classification

1 Introduction

Traditionally the classification task focuses on the estimation of a single output variable. This is done through learning from a set of examples. These instances are associated with a unique tag λ_i from a set of disjoint tags $\mathcal{L} = \{\lambda_1, \lambda_2, \ldots, \lambda_n\}$ [13]. If $|\mathcal{L}| = 2$ the learning task is called a problem of binary classification, where generally $\mathcal{L} = \{\lambda, \neg\lambda\}$ indicates if an instance belongs or not to the class indicated by the label. For the case where $|\mathcal{L}| > 2$ the problem is called multiclass classification. The set of unique tags that can be associated with an instance is also known as class variable [13].

Consider the m-dimensional input space $\mathcal{X}$, with $\mathcal{X} = \mathcal{X}_1 \times \ldots \times \mathcal{X}_m$, $(\mathcal{X}_i; i = 1, \ldots, m)$, where $\mathcal{X}_i \in \mathbb{N}$ (nominal features), $\vee\ \Omega_{X\,i} \in \mathbb{R}$ (numeric features) and $\mathcal{L}$ as the set of output tags. The learning task consists of finding a function h

$$h : \mathcal{X}_1 \times \ldots \times \mathcal{X}_m \to \mathcal{L}$$
$$(x_1, \ldots, x_m) \mapsto \lambda_i \in \mathcal{L}$$

such that h must be able to generalize correctly, in the sense of minimizing the loss of expected prediction with respect to a specific loss function.

A natural generalization of the classical classification problem is the multi-dimensional classification problem. In this case, the classifier is associated with multiple output variables. These variables can either be binary or multiclass like in the classical problem. Depending on the *a priori* information that can be obtained from these variables and the output type the following taxonomy is used to classify these problems: multilabel, multi-dimensional or structured output. This article focuses on the multilabel problem.

A multilabel classifier can be seen as generalization of the single-output classifier. In these classifiers the instances are associated with a set of labels L, where $L \subset \mathcal{L}$. However, in contrast with the traditional approach, the labels assigned by the classifier are not mutually exclusive. Therefore, any instance in the data set can be associated with more than one label. Historically, these classifiers were motivated by problems in text classification and medical diagnosis.

In multilabel classification the learning task can be reformulated as finding the function h

$$\begin{aligned} h : \mathcal{X}_1 \times \ldots \times \mathcal{X}_m &\to \mathcal{P}(\mathcal{L}) \\ (x_1, \ldots, x_m) &\mapsto L = \{\lambda_r, \ldots, \lambda_q\} \subseteq \mathcal{L} \end{aligned}$$

where $\mathcal{X}_1 \times \cdots \times \mathcal{X}_m$ represents the input space and $\mathcal{P}(\mathcal{L})$ denotes the power set of $\mathcal{L}$ (i.e. the set of all possible subsets). This function associates each example with a set $L = \{\lambda_r, \ldots, \lambda_q\}$ of labels. The labels present in this subset of $\mathcal{L}$ are called the relevant labels. Thus, the learning task is to find the function h such that h minimizes the expected prediction loss for the label set [13].

An important assumption of this model is that the different class variables associated (either through manual or automatic labeling) for each of the instances serve to improve the efficiency of the classifiers. This is because they contain information that can be used to subdivide the space of original characteristics and enhance its discriminating power ever more [20, 19, 14].

The problem of multilabel classification is an active area of research, recent surveys demonstrate several challenges in multilabel classification. Some of these include dealing with high dimensionality, the exploration of label correlations in an efficient manner and the understanding of these correlations [6, 17].

This article proposes a new method for multilabel classification based on Linear Discriminant Analysis (LDA) [7] and Gram-Schmidt Orthogonalization (GS) [4] procedure through QR Factorization [1]. While the idea of applying GS in the context of LDA is not new (see for example: [18, 3, 2]), however none of those works have been focused on solving the multilabel problem, instead they deal with issues such as computational complexity and efficiency. On the other hand, there have been various proposals of multilabel extensions for LDA [15, 10, 8, 9]. None of the papers found in the reviewed literature used an approach similar to the one proposed in this article.

2 Proposed method: ML-GS-LDA

The classical LDA method is based on the computation of the scatter matrices S_w and S_b (the within class scatter matrix and the between class scatter matrix, respectively) and the optimization of the following objective function:

$$J(W) = \mathbf{tr}\left(\frac{W^T S_b W}{W^T S_w W}\right)$$

The projection W^* which maximizes $J(W)$ is calculated through Singular Value Decomposition (SVD) [7].

However, while this approach works for multiclass problems, it is still limited to a single label. One possible extension dealing with multilabel classification is through the definition of class-wise scatter matrices [15]. Another approach centered on multilabel classification efficiency can be made through the use of two successive QR factorization instead of conventional SVD and a clever application of null spaces [10]. Both extensions empirically demonstrate that classification performance can be improved by using multiple class labels together.

In this context, this article proposes a new method for multilabel classification based on LDA. The proposed method is arguably simpler and easy to implement, though it could be slightly more expensive in terms of computational time, basing itself on iterated LDA over different labels and Gram-Schmidt orthogonalization. The method is described in Algorithm 1, the details are given in the next paragraph.

The proposed method works as follows, for all the included labels the classical LDA projection matrix is built using the input data X and the corresponding column from the matrix Y. These projections all have the same dimensionality d, given by the number of eigenvectors selected, according to the parameter Dim. In this case, since there are only two possible classes for each label, a value of $Dim = 1$ was assigned. The conditional inside the loop considers the first iteration separately from the rest, indicating that when W is not initialized it corresponds to the classical LDA projection matrix of the first label. The matrices are concatenated in each iteration, forming the final matrix W of order $|A| \times d \cdot |C|$ where $|A|$ is the number of attributes in the original data, d is the dimensionality of each projection and $|C|$ is the number of classes given by the length of the *LabelOrder* parameter. Note that each of the individual projections is orthogonal, however, the resulting matrix W is not necessarily orthogonal. So, in order to guarantee the orthogonality (and therefore uncorrelatedness) of each axis, the Gram-Schmidt procedure is applied to the matrix W. The resulting matrix is returned and can be used to project the original data set onto the new subspace.

Some observations need to be made about the underlying assumptions in the proposed method. The first one is that orthogonality between a feature axis (in terms of the covariance dot product) implies uncorrelatedness of those same axis [11]. The second one is that the application of Gram-Schmidt orthogonalization procedure on the concatenated projection basis generates an adequate projection

Algorithm 1 ML-GS-LDA

Input: X, matrix containing the data in rows; Y, matrix containing the labels associated with each data point in X; LabelOrder, a list indicating the order in which the model must be trained (e.g. [0, 1, 2] means to predict the class using the information from labels zero, one and two); Target, the classification target of the algorithm, indicates the label that must be predicted; Dim, value indicating the number of dimensions of each LDA projection.

Output: W, the projection matrix generated after successive LDA computation and the application of the Gram-Schmidt procedure.

```
1: function ML-GS-LDA
2:     W = null
3:     for all Label in LabelOrder do
4:         W' = LDA(X, Y[Target, :], Dim)
5:         if W == null then
6:             W = W'
7:         else
8:             W = [W, W']
9:         end if
10:    end for
11:    return Gram-Schmidt(W)
12: end function
```

for this problem. And finally, the last assumption is that the generated projection will mainly encode the information of the first label while at the same time retaining the extra information from the auxiliary labels.

By using multiple information sources (i.e. a multilabel approach) and eliminating the correlation in this new subspace it is expected that there is improvement in terms of classification accuracy, similar to how the use of Gram-Schmidt in the context of classical linear regression produces some improvements in accuracy and uncorrelated regressors [5].

Table 1. Description of the labels contained in the Emotions data set.

Label	Description	#Examples
L1	amazed-surprised	173
L2	happy-pleased	166
L3	relaxing-calm	264
L4	quiet-still	148
L5	sad-lonely	168
L6	angry-fearful	189

3 Methodology

To evaluate the proposed method the Emotions [16, 12] data set was used. This data set contains 593 instances with 72 numerical attributes. These attributes represent 30 seconds from a musical piece fragment. The 72 attributes are obtained through the application of different filters and transformations to each musical fragment. Also, each instance can be associated with up to six labels corresponding to different emotional states, see Table 1.

3.1 Validation

The method was evaluated using a 5-*fold* cross validation approach. For each experiment PCA (Principal Component Analysis) was applied for a first stage dimensionality reduction. Afterwards, for further dimensionality reduction either the supervised method LDA or the proposed algorithm ML-GS-LDA were applied. To obtain the performance for each label a Naive Bayes classifier was used for each method. The parameters to define the best configuration for the PCA dimension and the label ordering selection for the ML-GS-LDA algorithm were performed using a *wrapper* method.

Two different approaches for ML-GS-LDA were evaluated with the purpose of verifying the significance of the orderings of the labels on the ML-GS-LDA algorithm and their impact on the performance of the classifier. The first one considers combinations of the labels (i.e. the ordering does not matter) and the second one considers permutations (i.e. the ordering matters). See line 3 of Algorithm 1.

4 Results and discussion

The obtained accuracy results are shown in Table 2. The analysis of these values shows a marginal increment in the classification accuracy for each label. Also, by analyzing the labels selected for each projection some relationships and dependencies between the different labels can be inferred.

Table 2. Results for the Emotions data set [13] for the PCA+LDA and PCA+ML-GS-LDA and a Naive Bayes classifier.

	Labels					
Classifier	L1	L2	L3	L4	L5	L6
PCA+LDA	80.95 ± 2.29 PCA: 15	73.20 ± 2.23 PCA: 30	73.68 ± 2.75 PCA: 45	88.87 ± 4.12 PCA: 35	83.31 ± 2.30 PCA: 55	79.26 ± 3.62 PCA: 20
PCA+ML+GS+LDA Combinations	82.13 ± 3.21 PCA: 20 L: (1, 5, 6)	74.38 ± 3.00 PCA: 30 L: (1, 3, 5)	76.37 ± 5.42 PCA: 20 L: (1, 2, 3, 4, 5)	90.06 ± 4.82 PCA: 35 L: (1, 4)	84.33 ± 2.80 PCA: 55 L: (2, 4, 5)	80.45 ± 4.15 PCA: 30 L: (1, 2, 4, 6)
PCA+ML+GS+LDA Permutations	82.13 ± 3.21 PCA: 20 L: (1, 5, 6)	76.24 ± 3.56 PCA: 50, L: (1, 3, 4, 2)	77.04 ± 5.27 PCA: 20 L: (1, 3, 4, 2, 6)	90.06 ± 4.02 PCA: 35 L: (5, 1, 4, 2)	84.33 ± 2.80 PCA: 55 L: (2, 4, 5)	82.64 ± 4.43 PCA: 30 L: (2, 4, 1, 6)

Marginal improvements in classification accuracy can be found for all the six labels, independently of the applied strategy (combinations or permutations). The analysis of these results suggests that the combination of different LDA projections, in conjunction with orthogonalization, could improve classification accuracy in the multilabel problem.

The proposed algorithm can find dependencies and relationships between the different labels, in the sense that the information obtained from their LDA projection and subsequent orthogonalization allows for prediction of another label.

A quick review of the relationships is described. However, an in-depth analysis of these dependencies is beyond the scope of this work. The following relationships were found:

- **Label 1**: In this case, both the combination and permutation approach have the same optimal label ordering. According to this, L1 would be related with L5 and L6. It is interesting to note that in both cases the target label is first in the ordering.
- **Label 2**: In this case, the combination and permutation approach have different optimal label orderings. However, in both cases L1 and L3 are present, while the difference between the two is the interchanging of L5 for L4 and L2, resulting in a higher classification accuracy for the permutation approach. It is interesting to note that the predicted label was found to give better results when placed last in the ordering instead of being in the first position.
- **Label 3**: In this case, the results are similar, however L5 in the combination approach has been replaced by L6 in the permutation approach. Also, the ordering of L2, L3 and L4 has been changed for the permutation case. This results in an increase in classification accuracy for this label.
- **Label 4**: In this case, the classification accuracy is the same for both cases, however the label orderings are different. It is interesting to note that the label ordering found by the combination approach is a subset of the one found by the permutation approach. Also, while the average value is the same, by using the permutation approach a lower variance was found.
- **Label 5**: In this case, both methods produce the same results. From this it can be found that L5 is related with L2 and L4.
- **Label 6**: In this case, classification accuracy is higher for the permutation case. However, the only difference between the methods is the label ordering, since both sets use the same labels, this is interesting because it reveals the importance of the ordering for finding the optimal classifier.

Based on the obtained results and the previous analysis, apart from the accuracy gain discussed before the benefits from this proposal are two-fold:

- The search for the optimal label ordering provides a new way to study the dependencies and relationships between the different labels. While this model provides a basis for this analysis, the theoretical implications and properties of these orderings must be further evaluated, both in terms of empirical analysis and the mathematical properties associated with them.

– The final method is fast in comparison with other approaches, since the results are obtained through a simple matrix multiplication and the application of Naive Bayes. It is important to note the training phase and the search for the optimal parametrization can require plenty of computational resources, however, this is true for all methods that require finding several hyper-parameters.

It is possible that through the application of another classifying scheme, such as some variant of SVM better results could be obtained. However, the use of NB has the aforementioned benefit of being more efficient in terms of computational resources compared to SVM.

5 Conclusions

This work has explored a novel generalization of the classical LDA method for multilabel problems, this proposal is based on the Gram-Schmidt orthogonalization procedure through QR factorization. The theoretical basis for this model and its underlying assumptions have been exposed and the proposal has been validated through experimental evaluation on the Emotions data set. The analysis of the empirical results support that this new method is competitive and in some instances superior to the baseline.

One of the main benefits of the proposed algorithm is that it serves as an exploration tool for label dependencies and relationships. An in-depth analysis, both theoretical and empirical, of the relationships that this algorithm finds is proposed as future work. Another important plus of this method is that it is fast once the classifier has been trained. Since the resulting projection can be obtained through the application of a simple matrix multiplication.

On the other hand, one of the key aspects of the method is the order of the decomposition, which affects the obtained results. The election of the right order is a challenge that must still be addressed. Also, experimental results show that the results vary from label to label. The task of correctly exploiting these results to provide the best global result is still pending and is considered as future work. Comparison with other state of the art methods, such as deep learning approaches for emotion recognition, is also considered as future work.

Finally, some limitations of the proposed method in its current form correspond with the combinatorial optimization required to determine the optimal multilabel combination. To approach this, a step-wise approach like the one used in the process of finding the best regressors for linear regression or a greedy strategy based on an easy to evaluate heuristic could be useful to find better solutions, although optimality of the combinations would still be an open problem. Future work plans on dealing with the implementation of a modified version of the algorithm that includes optimization routines and exhaustive evaluation on multilabel data sets with different metrics.

References

1. Björck, Å.: Numerics of gram-schmidt orthogonalization. Linear Algebra and Its Applications 197, 297–316 (1994)
2. Cai, D., He, X., Han, J.: Srda: An efficient algorithm for large-scale discriminant analysis. IEEE transactions on knowledge and data engineering 20(1), 1–12 (2008)
3. Cevikalp, H., Neamtu, M., Wilkes, M., Barkana, A.: Discriminative common vectors for face recognition. IEEE Transactions on pattern analysis and machine intelligence 27(1), 4–13 (2005)
4. Cohen, H.: A course in computational algebraic number theory, vol. 138. Springer Science & Business Media (2013)
5. Farebrother, R.: Algorithm as 79: Gram-schmidt regression. Journal of the Royal Statistical Society. Series C (Applied Statistics) 23(3), 470–476 (1974)
6. Gibaja, E., Ventura, S.: Multi-label learning: a review of the state of the art and ongoing research. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 4(6), 411–444 (2014)
7. Izenman, A.J.: Linear discriminant analysis. In: Modern multivariate statistical techniques, pp. 237–280. Springer (2013)
8. Ji, S., Ye, J.: Linear dimensionality reduction for multi-label classification. In: IJCAI. vol. 9, pp. 1077–1082 (2009)
9. Oikonomou, M., Tefas, A.: Direct multi-label linear discriminant analysis. In: International Conference on Engineering Applications of Neural Networks. pp. 414–423. Springer (2013)
10. Park, C.H., Lee, M.: On applying linear discriminant analysis for multi-labeled problems. Pattern recognition letters 29(7), 878–887 (2008)
11. Rodgers, J.L., Nicewander, W.A., Toothaker, L.: Linearly independent, orthogonal, and uncorrelated variables. The American Statistician 38(2), 133–134 (1984)
12. Trohidis, K., Tsoumakas, G., Kalliris, G., Vlahavas, I.P.: Multi-label classification of music into emotions. In: Bello, J.P., Chew, E., Turnbull, D. (eds.) ISMIR. pp. 325–330 (2008), http://dblp.uni-trier.de/db/conf/ismir/ismir2008.html#TrohidisTKV08
13. Tsoumakas, G., Katakis, I.: Multi-label classification: An overview. International Journal of Data Warehousing and Mining 3(3), 1 (2007)
14. Wan, H., Wang, H., Guo, G., Wei, X.: Separability-oriented subclass discriminant analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence (2017)
15. Wang, H., Ding, C., Huang, H.: Multi-label linear discriminant analysis. Computer Vision–ECCV 2010 pp. 126–139 (2010)
16. Wieczorkowska, A., Synak, P., Raś, Z.W.: Multi-Label Classification of Emotions in Music, pp. 307–315. Springer Berlin Heidelberg, Berlin, Heidelberg (2006), http://dx.doi.org/10.1007/3-540-33521-8_30
17. Zhang, M.L., Zhou, Z.H.: A review on multi-label learning algorithms. IEEE transactions on knowledge and data engineering 26(8), 1819–1837 (2014)
18. Zheng, W., Zou, C., Zhao, L.: Real-time face recognition using gram-schmidt orthogonalization for lda. In: Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on. vol. 2, pp. 403–406. IEEE (2004)
19. Zhu, M., Martinez, A.M.: Selecting principal components in a two-stage lda algorithm. In: Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on. vol. 1, pp. 132–137. IEEE (2006)
20. Zhu, M., Martinez, A.M.: Subclass discriminant analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence 28(8), 1274–1286 (2006)