2015 IEEE 14th International Conference on Machine Learning and Applications

Multi-label Classification of Anemia Patients

Colin Bellinger*, Ali Amid†, Nathalie Japkowicz‡, Herna Victor§

*School of Computer Engineering and Electrical Engineering
University of Ottawa,
Email: colin.bellinger@gmail.com

†The Hospital for Sick Children
Toronto, Canada
Email: ali_amid_md@yahoo.com

‡School of Computer Engineering and Electrical Engineering
University of Ottawa,
Email: nat@eecs.uottawa.ca

§School of Computer Engineering and Electrical Engineering
University of Ottawa,
Email: hlviktor@eecs.uottawa.ca

***Abstract*—This work examines the application of machine learning to an important area of medicine which aims to diagnose paediatric patients with β-thalassemia minor, iron deficiency anemia or the co-occurrence of these ailments. Iron deficiency anemia is a major cause of microcytic anemia and is considered an important task in global health. Whilst existing methods, based on linear equations, are proficient at distinguishing between the two classes of anemia, they fail to identify the co-occurrence of this issues. Machine learning algorithms, however, can induce non-linear decision boundaries that enable accurate classification within complex domains. Through a multi-label classification technique, known as problem transformations, we convert the learning task to one that is appropriate for machine learning and examine the effectiveness of machine learning algorithms on this domain. Our results show that machine learning classifiers produce good overall accuracy and are able to identify instances of the co-occurrence class unlike the existing methods.**

I. Introduction

β-thalassemia minor (β-thal) and iron deficiency anemia (IDA) are the most common causes of anemia with small red blood cells (microcytic anemias) in paediatric population. They are considered global health challenges with a significant burden on health care systems of countries where IDA and β-thal are common. Differentiating between β-thal and IDA has important implications in β-thalassemia carrier screening as well as therapeutic intervention for iron deficiency anemia.

The confirmatory diagnosis of these conditions may be expensive, especially on a national-scale; thus, several linear complete blood count (CBC)-based equations have been proposed to differentiate between these two conditions. In particular, Mentzer observed that the so-called *Mentzer index* can be applied to distinguish between β-thal and IDA [1]. Subsequently, several other CBC-based equations have been developed [2, 3, 4, 5]. These methods, however, fail at distinguishing occurrences of β-thalassemia from the co-occurrence of IDA and β-thalassemia. Thus, the applicability of these equations in populations with high rates of iron deficiency and β-thalassemia, such as the Mediterranean and some developing countries, is limited [6]. Hence, a simple, easy to use test to differentiate these patients, and also those with both conditions on a large scale, is of significant benefit, as it may result in earlier diagnosis and a significant reduction of the cost on health care systems.

Machine learning offers great potential in this domain, as the data distributions are complex and there is a high degree of overlap between the β-thalassemia class and the co-occurrence class. Whilst methods based on linear equations are unable to sufficiently cope with such complexity, advanced learning algorithms induce non-linear decision boundaries that have proven to be effective on challenging classification tasks.

The existence of the co-occurrence class β-thal+IDA make this a particularly interesting problem. In the machine learning context, the co-occurrence problem fits into the domain of *multi-label classification* [7]. In this inaugural work, two popular methods for transforming the co-occurrence class to a standard learning form and five classification algorithms are tested. In general, our results show that machine learning increases our ability to discriminate between patients with β-thal and those with both β-thal and IDA.

II. Related Work

A. Machine Learning

Standard machine learning algorithms induce a function $h : \mathcal{X} \rightarrow \Omega$ mapping an input x to its corresponding class ω_i where $\omega_i \in \Omega = \omega_1, \omega_2, ..., \omega_l$. Each instance belongs to exactly one of the l classes in Ω. Alternatively, the objective of multi-label learning is to induce a function $h : \mathcal{X} \rightarrow \Omega$

978-1-5090-0287-0/15 $31.00 © 2015 IEEE
DOI 10.1109/ICMLA.2015.112

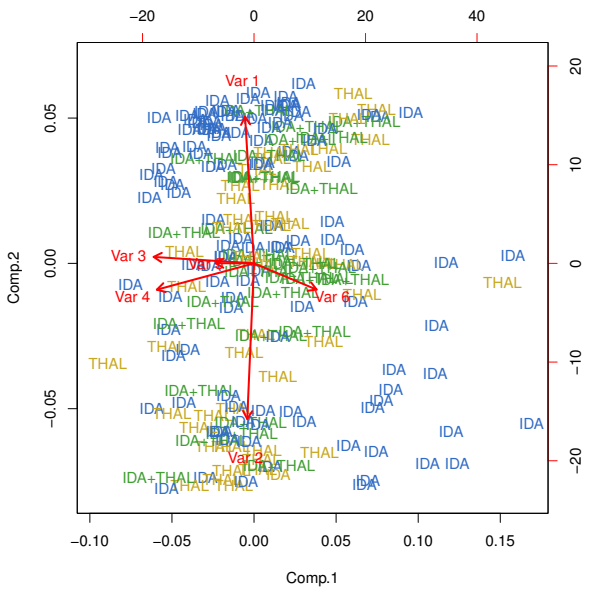

Fig. 1. PCA transformation of the microcytic anemia data

mapping an input x to its corresponding set of classes Ω_i where $\Omega_i \subset \Omega$. Thus, an instance may belong to one or more classes.

This works initiates the discussion of machine learning in the domain of microcytic anemia. Specifically, we examine the benefits of applying standard multi-class classifiers to distinguish between β-thal, IDA and their co-occurrence class (in order to do this, we first perform the multi-label classification technique known as problem transformation, which is discussed in Section II-B). The five classification algorithms employed are multilayer perceptron (MLP), support vector machines (SVM), decision tree (J48), naive Bayes' (NB) and k-nearest neighbours (kNN). As these are standard learning algorithms, we omit a discussion of them here and direct the reader to [8].

B. Multi-Label Classification

Multi-label classification is a unique form of classification problem in which a single instance may belong to more than one class [7] (for example, in the classification of satellite image of forest coverage [9]). In the domain explored here, patients may have IDA, β-thal, or both. This complicates the task of binary classification as the data-space can no longer be partitioned into mutually exclusive subspaces.

Existing practices approach the task of multi-label learning by either applying *problem transformation* or *algorithm adaption*. Problem transformation converts the problem into one or more single label tasks, and algorithm adaption extends existing methods of single label classification to multi-label classification. This work applies two popular problem transformation methods; namely, the power set transformation and the binary relevance transformation.

The label power set treats each unique label set as its own class. A sample anemia dataset is depicted in Table I

TABLE I. EXEMPLARY MULTI-LABEL DATASET.

Instance	Attributes	Labels
1	x_1	{ IDA }
2	x_2	{ IDA, Thal }
3	x_3	{ Thal }
4	x_4	{ IDA, Thal }
5	x_5	{ Thal }
6	x_6	{ Thal }

TABLE II. POWER SET TRANSFORMATION OF EXEMPLARY MULTI-LABEL DATASET.

Instance	Attributes	Labels
1	x_1	{ IDA }
2	x_2	{ IDA-β-thal }
3	x_3	{ Thal }
4	x_4	{ IDA-β-thal }
5	x_5	{ Thal }
6	x_6	{ Thal }

and its power set transformation in Table II. In the original form, there are two classes, and instances 2 and 4 are associated with both the IDA class the β-thal class. The transformation adds an additional class that represents the co-occurrence of IDA and β-thal. Thus, the label set becomes $\Omega = \{IDA, \beta{-}thal, (IDA\ \beta{-}thal)\}$. A significant drawback of this transformation is its size complexity of $min(n, 2^k)$ where n is the number of instances and k is the number of classes. Given that there are only two classes, complexity is not an issue here. In addition to the potential complexity, some elements of the power set may be significantly underrepresented; indeed, we see this issue in our data.

The binary relevance transformation creates a new dataset for each class in the original classification task. The resulting datasets are presented in Table III. This transformation creates a set of one-versus-all tasks in which a classifier is induced for each of the k datasets. If, for example, the two classifiers of the BR transformation predict $h_{IDA}(x_i) = 1$, and $h_{Thal}(x_i) = 1$, then the final classification is $\{1, 1\} \Rightarrow \{IDA, Thal\}$.

Unlike PS, the BR transformation method is not susceptible to the explosion of classes in the learning task, nor is it susceptible to the risk of class imbalance. However, by creating a set of one-versus-all learning tasks, there is a significant risk that any dependencies existing within the co-occurrence classes will be lost. Moreover, the co-occurring instances appear as positives in at least two of the derived datasets. Instance two, for example, is listed as a positive in both datasets because it co-occurs with both labels.

In spite of their limitations, these simple methods have proven successful and remain popular as they can be applied with off-the-shelf classifiers. For these reasons we utilize BR and PS in this study. Alternative transformations that exist to facilitate ranking, such as *Ranking by pairwise comparison* [10] and *Calibrated label ranking* [11], may also prove bene-

TABLE III. BINARY RELEVANCE DATASETS PRODUCED FROM THE EXEMPLARY MULTI-LABEL DATASET.

Ex	Label	Ex	Label
1	{ ω_{IDA} }	1	{ $\neg\omega_{Thal}$ }
2	{ ω_{IDA} }	2	{ ω_{Thal} }
3	{ $\neg\omega_{IDA}$ }	3	{ ω_{Thal} }
4	{ ω_{IDA} }	4	{ ω_{Thal} }
5	{ $\neg\omega_{IDA}$ }	5	{ ω_{Thal} }
6	{ $\neg\omega_{IDA}$ }	6	{ ω_{Thal} }

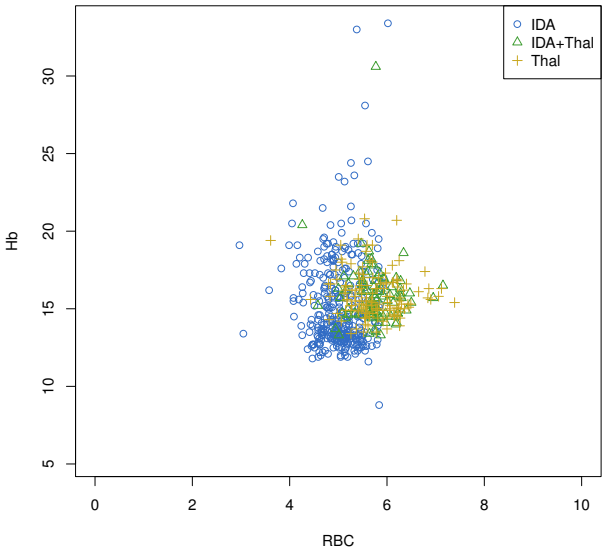

Fig. 2. Plot of data with respect to RBC and Hb with the outlier removed. The outlier had the values {Hb,RBC} = {6,33}.

ficial for medical domains and will be studied in future work.

III. Data

The microcytic anemia domain of classification considered here has two classes: IDA and β-thal. These conditions co-occur in some patients, in which case both classes should be predicted. The dataset utilized in these experiments is relatively small, with 562 instances. Of these, 59.3% are labelled IDA, 23.3% belong to β-thal class and 17.4% are co-occurrence instances (IDA+β-thal). It is clear from this that the class distribution has a significant level of relative imbalance [12]. This is particularly acute between IDA and the co-occurrence class. The data is composed of just 6 features of the form: $\{sex, age, Hb, Hct, RBC, RDW\} = \{(m, f), \mathbb{I}, \mathbb{R}, \mathbb{R}, \mathbb{R}, \mathbb{R}\}$. Specifically, Hb is haemoglobin, Hct is hematocrit, RBC is the red blood cell count and RDW is the red cell distribution width.

A. Preprocessing and Analysis

1) Preprocessing: The size of the dataset rendered preprocessing a simple task. The data was originally stored in an Excel file, and was subsequently converted to CSV format with one column per variable and one row per instance. The final column specifies the class label. The PS transformation resulted in a new dataset with three classes $\Omega = \{IDA, Thal, IDA + Thal\}$. The class priors for this transformation match those in the original form. The BR transformation resulted in two new datasets $\mathbf{D}_{\{\omega_{IDA}, \neg\omega_{IDA}\}}$ and $\mathbf{D}_{\{\omega_{Thal}, \neg\omega_{Thal}\}}$. The class priors are $(76.7\%, 23.3\%)$ in the former dataset and $(40.7\%, 59.3\%)$ in the latter.

TABLE IV. Results of five feature analysis techniques. These show the relative importance of each feature.

Method	Feature ranking order
cfsSubsetEval	RBC, RDW
Attribute Correlation	RBC, age, RDW, Hb, Hct
Gain Ratio	RBC, RDW, Hb, age, Hct, gender
Symmetrical Uncertainty	RBC, RDW, Hb, age, Hct, gender
Information Gain	RBC, RDW, Hb, age, Hct, gender

2) Data Analysis: In this subsection we analyze the data to gain insight into its complexity and its possible impact on classification. Principle component analysis (PCA) along with two- and three-dimensional plots of the data-space were used for this purpose. In addition, we identified an outlier in the data through this process.

In order to rank the impact of the features and determine which features to plot in our analysis, we utilized five feature selection methods; the results are presented in Table IV. In all cases, RBC is found to be the most important feature, with RDW identified as the second most important four out of five times. Gender is removed by the feature selection methods and ranked last by the methods that rank features. Indeed, this is consistent with medical knowledge.

A PCA plot of the data is presented in Fig 1. This plot presents eigenvector and eigenvalues in terms of the direction and length of the red arrows and the data in the PCA-space. This gives an indication of the relative position of the classes and the overall complexity of the data. In order to increase readability, we only plot a subset of the instances. The PCA plot suggests that this is a complex learning task with IDA having many subconcepts and a high degree of spread. In general, the main concept of IDA is visibly separable from β-thal and the co-occurrence class. However, many instances of the β-thal class and the co-occurrence class are clustered at the origin.

For further analysis, the top ranked features are plotted in their two-dimensional data-space in Figure 2 and Figure 3. As is suggested by the PCA plot, the densest concept in IDA appears quite separable from β-thal and IDA+β-thal, whilst the others are almost entirely overlapping. In Figure 3, various combinations of RBC, RWD, age and Hb are plotted on the three axes in order to view the data from different angles. IDA appears most separable in Fig 3 where RBC and Hb are plotted. None of the two-dimensional plots display features that cause β-thal and the co-occurrence to appear overly separable. Adding the third dimension to the plots increases the separability to a certain extent. Nonetheless, β-thal and β-thal+IDA remain extremely challenging.

From our analysis, it is clear that data overlap is a major challenge in this domain. In addition, the distributions are relatively complex and the data is imbalanced. These three facts are considered in our analysis of the final results.

IV. Experimental Methodology

Each algorithm was trained and tested on the BR and PS transformed datasets using 10x10-fold cross-validation, as it is recommended over $10-$fold cross validation and 5x$2-$fold cross validation for choosing between classifiers [13]. On each fold of each iteration, the precision, recall, f-meausure

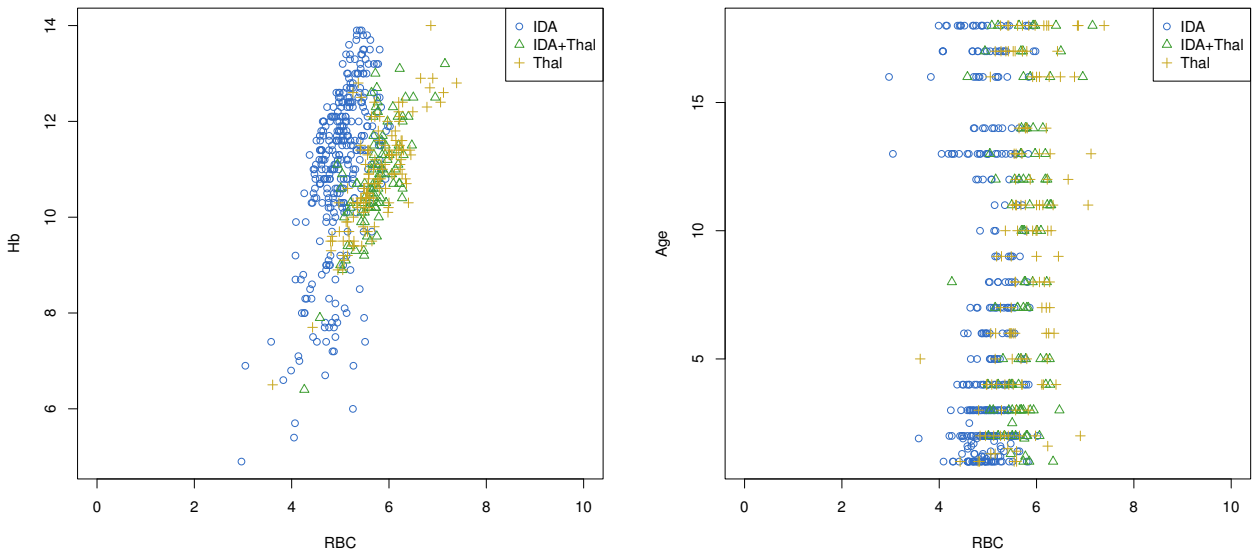

Fig. 3. The RBC, Hb and age features of the microcytic anemia data.

TABLE V. WEIGHTED MEAN CLASSIFICATION RESULTS ON THE PS TRANSFORMED DATA.

PS	Precision	Recall	FM	AUC
MLP	0.7278	0.8308	0.7212	0.9102
SVM	0.7392	0.8160	0.7356	0.8065
DT	0.7136	0.7576	0.7097	0.8001
IBK	0.6219	0.6410	0.6146	0.6919
NB	0.6193	0.7854	0.6486	0.8323

TABLE VI. MEAN CLASSIFICATION RESULTS ON THE BR TRANSFORMED DATA.

BR	Precision	Recall	FM	AUC
MLP	0.7195	0.5773	0.6000	0.7362
SVM	0.7155	0.6007	0.5898	0.7448
DT	0.6756	0.5458	0.5656	0.7113
IBK	0.6015	0.4667	0.4931	0.6572
NB	NA	NA	NA	NA

and area-under-the-ROC curve (AUC) are recorded. For each metric, the mean of the 100 results produced over the 10 iterations of 10-fold cross-validation are reported, and the t-test is used to evaluate the significance of the results for each algorithm across the transformations and between the best two algorithms within each transformation. Each classifier is tested with its default Weka[1] parameter set.

V. RESULTS

A. Weighted Mean Results

On the PS transformed dataset, $AUC(MLP)$ is significantly better than all other methods. These results are presented in Table V. With respect to FM, precision and recall, MLP and SVM achieve similar results. Table VI reports the results produced on the BR transformed datasets. In terms of each metric, the results produced by MLP, SVM, DT and IBK are much closer. Indeed, no statistical significance can be found in the difference between the classifiers.

In addition to the performance within each transformation, we must consider which transformation leads to the best results. For this we can consider the performance of the individual classifiers on each method. MLP, for example, is significantly better in terms of AUC on the PS data, where it produced a score of 0.9102 in comparison to 0.7362 on the BR data. Indeed, when comparing PS to BR, all classification methods are superior in terms of their weighted means on precision, recall, FM and AUC on the former transformation.

[1] http://www.cs.waikato.ac.nz/ml/weka/

B. Results by Class

As a means of further understanding the performance of MLP and SVM, we present their results in the form of pie charts in Figure 4. In each figure, the size of the pies represents the relative prior probability of the class; hence, the IDA pie is the largest and the β-thal+IDA pie is the smallest. In each case, the blue slice indicates the percentage of IDA predictions, yellow indicates the percentage of β-thal predictions and green corresponds to β-thal+IDA predictions. Thus, the larger the blue portion in the IDA pie, the better the classifier did on the IDA class. Likewise, ideally the β-thal pie would be primarily yellow and the β-thal+IDA pie would be mostly covered by the green slice.

This enables us to see the classes that are most accurately classified and least accurately classified. Moreover, we gain insight into the type of prediction errors made on each class. A majority of the co-occurrence instances, for example, are classified as β-thal by both SVM and MLP. This can likely be explained by the degree of overlap between the β-thal and β-thal+IDA classes, along with the class imbalance between

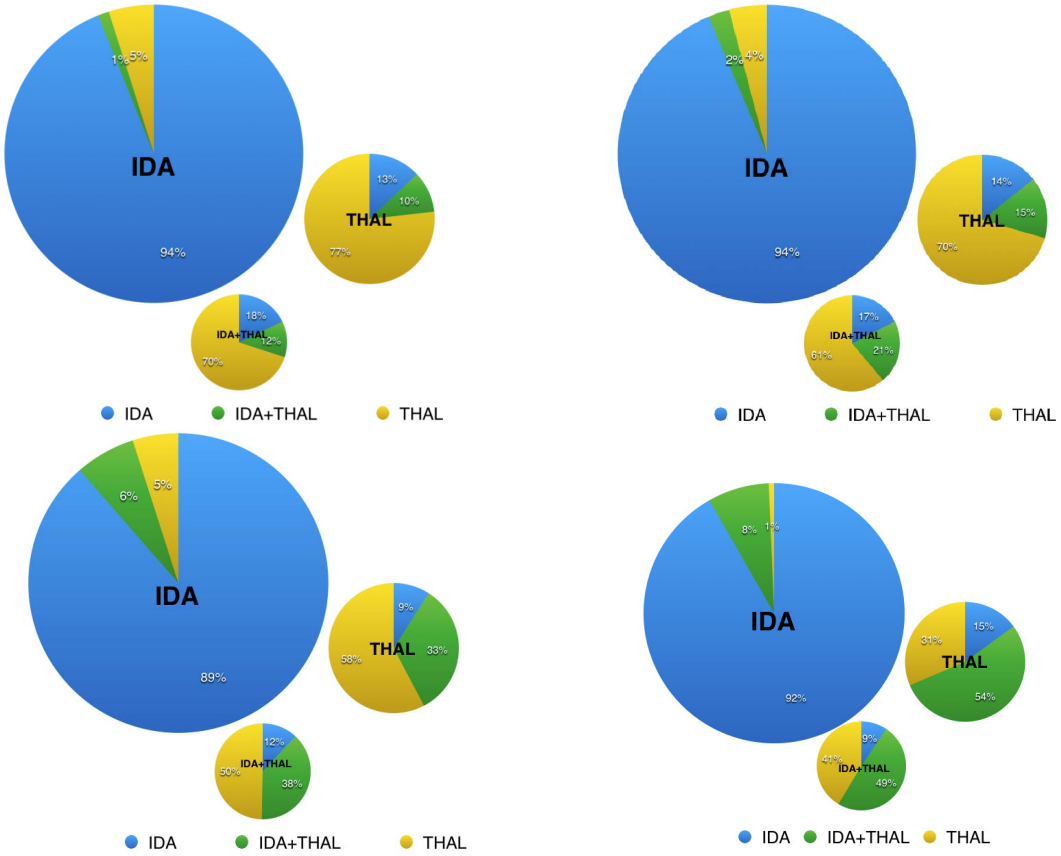

Fig. 4. Pie charts of the predictions by MLP and SVM on both transformations. The MLP predictions and SVM predictions for PS are in the top left and right, respectively. The bottom left and right have the MLP and SVM predictions for BR

them. In addition, it highlights the trade-off in terms of per-class performance between induction on the PS transformation and induction on the BR transformation

In general, both classifiers perform best on the IDA class. With respect to the PS transformation, both classifiers falsely associate IDA instances as one of the alternate classes 6% of the time. When the BR transformation is applied, the number increases to 11% for MLP and 9% for SVM.

MLP correctly identifies more of the β-thal instances and fewer of the co-occurrences instances than SVM on the PS transformed data. MLP identifies 77% and 12% respectively and SVM identifies 70% and 21%.

Performance on the IDA class when the BR transformation is applied is relatively consistent with the performance resulting from the PS transformation. With respect to β-thal and β-thal+IDA, performance changes considerably. Most notably, the total number of β-thal+IDA predictions (true β-thal+IDA and false β-thal+IDA) increase from $\{IDA, Thal, IDA + Thal\} = \{1\%, 10\%, 12\%\}$ by MLP to $\{IDA, Thal, IDA + Thal\} = \{6\%, 33\%, 38\%\}$. Thus, many more of the co-occurrence instances are correctly identified and many more IDA instances and β-thal instances are falsely identified as β-thal+IDA. SVM produces an even greater number of β-thal+IDA predictions. Previously, it produced $\{IDA, Thal, IDA + Thal\} = \{2\%, 15\%, 21\%\}$; with BR, it increases to $\{IDA, Thal, IDA + Thal\} = \{8\%, 54\%, 49\%\}$. Interestingly, the large number of instances incorrectly classified as β-thal+IDA causes a large decrease in the number of correctly identified β-thal instances.

VI. Discussion

A. Data Complexity

Our initial analysis of the data demonstrates that it is both complex in terms of class overlap as well as the presence of subconcepts within the individual classes. In addition, there is a degree of relative imbalance between the three classes, with IDA having significantly more instances than β-thal, and β-thal having more instances than the co-occurrence class.

Due to the overlap, discrimination between β-thal and β-thal+IDA is very challenging. This is confirmed by the results. Additional features, if available, may be very helpful for increasing the performance on these classes. In addition, addressing the imbalance either through machine learning methods or accessing more examples may help to improve the performance on the co-occurrence class. Our preliminary experiments with bagged random undersampling [14] and SMOTE [15] have demonstrated some potential to improve the classification of β-thal+IDA.

B. Multi-Label Classification

As an inaugural work, we only experimented with two multi-class transformation methods. In doing so, we were able to improve upon the existing results; however, other transformation methods exists that may be helpful; the ranking methods that are discussed in the cited work are perhaps appropriate for this domain. They allow more flexibility in the final decision due to the fact that experts can aid in selecting appropriate classification thresholds and/or make case-by-case decisions according to the emitted ranks.

In terms of the overall performance, PS is the clear winner as it leads to significantly better classifiers. PS has the benefit for preserving any dependence relationships that exist within the co-occurrence class. This suggests that it would perform well on the co-occurrence class. When the results are examined in the pie charts, however, we see that this is not the case. We can reasonably suspect that the degree of imbalance between the classes is a cause for the weaker performance. Based on this, we hypothesize that the reduced degree of imbalance that occurs when BR is used leads to better performance on the co-occurrence class. This is in spite to the lost dependency information. Thus, this further suggests the imbalance in the PS data should be directly managed.

VII. Conclusion

This work presents an exploration of the benefits of applying machine learning methods to the domain of microcytic anemia. Existing methods for discriminating between patients with IDA, β-thal are based on linear equations., and fail on the complex co-occurrence class.

Our analysis demonstrates that the data forms a multi-label classification task with imbalance between classes. Moreover, the data is further complicated by subconcepts within the classes and a significant degree of overlap.

Two multi-label transformations are applied to convert the data into a form suitable for multi-class learning. All classifiers induced on the PS transformation produce better weighted scores. With the binary relevance transformation, however, more of the co-occurrence class is correctly classified. MLP and SVM are nearly always the best classifiers, with MLP being significantly better according to the AUC.

The increasing availability of structured and unstructured data in biomedical informatics, along with its potential power to advance healthcare, makes this a field of significant importance [16, 17]. Thus, our future work will explore additional avenues to mine the data and expand upon our study of multi-label classification strategies.

References

[1] W. Mentzer, "Differentiation of iron deficiency from thalassaemia trait," *The Lancet*, vol. 301(7808), no. 882, 1973.

[2] J. England and P. Fraser, "ifferentiation of iron deficiency from thalassaemia trait by routine blood count," *The Lancet*, vol. 1, no. 7801, pp. 449–452, 1973.

[3] I. Shine and S. Lal, "A strategy to detect b thalassaemia minor," *The Lancet*, vol. 1(8013), pp. 692–694, 1977.

[4] R. Green and R. King, "A new red cell discriminant incorporating volume dispersion for differentiating iron deficiency anemia from thalassemia minor," *Blood cells*, vol. 15.3, no. 481-91, 1988.

[5] P. Srivastava and J. Bevington, "Iron deficiency and-or thalassaemia trait," *The Lancet*, vol. 1(7807), p. 832, 1973.

[6] A. Amid, B. Haghi-Ashtiani, M. Kirby-Allen, and M. T. Haghi-Ashtiani, "Screening for Thalassemia Carriers in Populations with a High Rate of Iron Deficiency: Revisiting the Applicability of the Mentzer Index and the Effect of Iron Deficiency on Hb A ¡sub¿2¡/sub¿ Levels," *Hemoglobin*, vol. 39, no. 2, pp. 141–143, 2015.

[7] N. Ghamrawi and A. McCallum, "Collective multi-label classification," *Proceedings of the 14th ACM international conference on Information and knowledge management - CIKM '05*, p. 195, 2005.

[8] T. M. Mitchell, *Machine learning*. McGraw-Hill, 1997.

[9] M. R. Boutell, J. Luo, X. Shen, and C. M. Brown, "Learning multi-label scene classification," *Pattern Recognition*, vol. 37, no. 9, pp. 1757–1771, Sep. 2004.

[10] E. Hüllermeier, J. Fürnkranz, W. Cheng, and K. Brinker, "Label ranking by learning pairwise preferences," *Artificial Intelligence*, vol. 172, no. 16-17, pp. 1897–1916, Nov. 2008.

[11] J. Fürnkranz, E. Hüllermeier, E. LozaMencía, and K. Brinker, "Multilabel classification via calibrated label ranking," *Machine Learning*, vol. 73, no. 2, pp. 133–153, Aug. 2008.

[12] E. Garcia, "Learning from Imbalanced Data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1263–1284, Sep. 2009. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5128907

[13] R. R. Bouckaert, "Choosing between two learning algorithms based on calibrated tests," in *Proceedings of the 20th International Conference on Machine Learning*, 2003, pp. 51–58.

[14] B. C. Wallace, K. Small, C. E. Brodley, and T. a. Trikalinos, "Class Imbalance, Redux," *2011 IEEE 11th International Conference on Data Mining*, pp. 754–763, Dec. 2011. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6137280

[15] N. Chawla, K. Bowyer, L. Hall, and K. W.P., "SMOTE: Synthetic Minority Over-Sampling Technique," *J. Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.

[16] A. Holzinger and I. Jurisica, "Knowledge discovery and data mining in biomedical informatics: The future is in integrative, interactive machine learning solutions," in *Interactive Knowledge Discovery and Data Mining in Biomedical Informatics*, ser. Lecture Notes in Computer Science, A. Holzinger and I. Jurisica, Eds. Springer Berlin Heidelberg, 2014, vol. 8401, pp. 1–18.

[17] A. Holzinger, M. Dehmer, and I. Jurisica, "Knowledge discovery and interactive data mining in bioinformatics-state-of-the-art, future challenges and research directions," *BMC bioinformatics*, vol. 15, no. Suppl 6, p. I1, 2014.