Ordinal Classification and Regression Techniques for Distinguishing Neutrophilic Cell Maturity Stages in Human Bone Marrow

Philipp Gräbel[1(✉)], Martina Crysandt[2], Barbara M. Klinkhammer[3], Peter Boor[3], Tim H. Brümmendorf[2], and Dorit Merhof[1]

[1] Institute of Imaging and Computer Vision, RWTH Aachen University, Aachen, Germany
{graebel,merhof}@lfb.rwth-aachen.de
[2] Department of Hematology, Oncology, Hemostaseology and Stem Cell Transplantation, University Hospital RWTH Aachen University, Aachen, Germany
[3] Institute of Pathology, University Hospital RWTH Aachen University, Aachen, Germany

Abstract. An automated classification of hematopoietic cells in bone marrow whole slide images would be very beneficial to the workflow of diagnosing diseases such as leukemia. However, the large number of cell types and particularly their continuous maturation process makes this task challenging: the boundaries of cell type classes in this process are fuzzy, leading to inter-rater disagreement and noisy annotations. The data qualifies as *ordinal data*, as the order of classes is well defined. However, a sensible "distance" between them is difficult to establish.

In this work, we propose several classification and regression techniques for ordinal data, which alter the encoding of network output and ground-truth. For classification, we propose using the Gray code or decreasing weights. For regression, we propose encodings inspired by biological properties or characteristics of the dataset. We analyze their performance on a challenging dataset with neutrophilic granulocytes from human bone marrow microscopy images. We show that for a sensible evaluation, it is of utmost importance to take into account the relation between cell types as well as the annotation noise. The proposed methods are straight-forward to implement with any neural network and outperform common classification and regression methods.

Keywords: Ordinal classification · Regression · Cell classification

This work was supported by the German Research Foundation (DFG) through the grants SFB/TRR57, SFB/TRR219, BO3755/6-1. The authors would like to thank Reinhild Herwartz and Melanie Baumann for their efforts in sample preparation and annotation.

M. El Yacoubi et al. (Eds.): ICPRAI 2022, LNCS 13364, pp. 186–195, 2022.
https://doi.org/10.1007/978-3-031-09282-4_16

1 Introduction

For the diagnosis of hematopoietic diseases such as leukemia it is necessary to analyze bone marrow samples in addition to peripheral blood. The major advantage of bone marrow analysis is a more detailed insight into hematopoiesis, the cell-forming process. The ratio of immature to mature granulocytes is of major importance, particularly for the detection of chronic myelogenous leukemia. While the process of maturation is a mostly continuous process, hematologists define five classes of granulocytes subsequent to the immature blast stage: promyelocytes, myelocytes, metamyelocytes, band granulocytes and segmented granulocytes. In theory, these maturity stages are well defined. However, manual class assignment by experts is fuzzy at transition stages which results in noisy labels. Inter-rater disagreement between adjacent classes can be observed. This refers not only to the annotations used in training but also to the desired predictions, which hampers a valid, automated evaluation.

Given the described continuous maturation process with transitions between classes and the resulting inter-rater disagreement between adjacent maturity stages, classification may not be the best option. While annotations are typically given as one of five maturity stages – and predictions ought to follow the same format – network optimization could be performed as a regression task. Formulated as regression, the task lies in predicting the maturity stage as a number, instead of a class. As the order of maturity stages is known without an obvious distance metric between those stages, this problem falls into the category of ordinal classification or regression.

The field of regression is a common research area, including some research which takes ordinal data into account. Also ordinal regression and classification have been researched extensively for generic classifiers [6]. Straight-forward approaches in this context, such as assigning regression targets or misclassification costs, can be transferred to deep learning. Other deep learning approaches require extensive changes to the network architecture and/or training process, for example, through pairwise comparisons [9].

In the field of hematopoietic cell classification from bone marrow microscopy images, the relationship between different classes is usually ignored. Song et al. [12, 13] work with bone marrow images but only distinguish between the erythroid and myeloid cell. The maturity grade within individual lineages is not further considered. Choi et al. [3] use a VGG architecture [11] to distinguish between different maturity stages within these two lineages. They treat most classes independently, but refine the prediction of the two most mature neutrophilic cells with a second VGG network. Preliminary experiments on our dataset showed, however, that this network architecture is outperformed by using a DenseNet [5]. Chandradevan et al. [2] perform a simple classification using a VGG network. All aforementioned methods do not perform an analysis on the relationship between adjacent maturity stages and do not take this knowledge into account for training and prediction.

Contribution
In this work, we propose and investigate several strategies to improve common classification and regression techniques in the case of ordinal data and

justify the need for those by providing a multi-rater data analysis of the given dataset. These straight-forward techniques are based on an established classification architecture and require minimal changes to the final linear layer and the class encoding. They achieve improved results in the case of classification and regression of neutrophilic granulocytes in human bone marrow microscopy images in an evaluation on multi-rater data.

2 Materials and Methods

2.1 Image Data

The data is obtained from human bone marrow samples as a purely retrospective, pseudonymized analysis under the Helsinki Declaration of 1975/2000 with written informed consent of all patients. Each sample is stained using the standardized Pappenheim staining procedure [1]. Image acquisition is performed with a whole slide scanner using 63× magnification and automatic immersion oiling. From each sample, relevant regions are selected using a lower magnification overview scan. In each region, cell positions are first proposed by an object detection network and then manually corrected. Each cell is annotated by two medical experts, who have agreed on the cell type. This annotated label is later referred to as an *MTA*-label and it forms the ground truth in all training processes. Examples are shown in Fig. 1.

In this work, the focus lies on predicting maturity stages. To this end, patches that are centered around individual cells of the neutrophilic granulopoiesis are extracted. In total, this results in 4301 cells from six classes: blast, promyelocyte, myelocyte, metamyelocyte, band granulocyte and segmented granulocyte. Of these, 767 cells have been annotated by additional different experts (two to three medical experts, with one annotation as mentioned above). If one expert declares a cell to be one of the considered cell types, while another assigns a completely different cell lineage, this is denoted by the cell type *other*.

Inter-rater Analysis. A comparison between two of the raters highlights the characteristics of inter-rater variability with respect to the maturity progression. Ignoring the *other* class, they agreed on the maturity stage in 65.3% of all cases. In 96.6% of all cases, however, they only differed by, at most, one stage. Apart from a single case, the remainder is, at most, two stages apart.

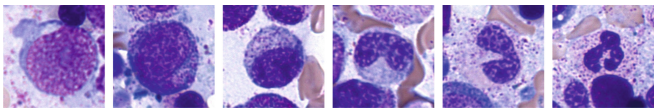

Fig. 1. One example for every cell type, which corresponds to a single maturity stage. From left to right: blast, promyelocyte, myelocyte, metamyelocyte, band granulocyte and segmented granulocyte.

This highlights the importance of a) using methods that not only consider absolute labels but also consider the data as ordinal and b) a multi-rater evaluation. When not taking the ordinality of this data into account, the ground truth provided by a single expert (or even multiple experts working together on a single annotation) could be considered wrong by a second expert in almost a third of all cases. This could lead to undesirable effects in training by putting a focus on "false" predictions that would actually be considered as correct by another rater.

Due to the necessary amount of manual work by medical experts, it is often impossible to obtain a complete multi-rater dataset for training. Instead, we propose strategies to include the ordinality of data into the training on a single-rater dataset. Nevertheless, the evaluation needs to be performed on multi-rater data, in order to get an indication of the actual success of such techniques.

2.2 Classification Techniques

Classification using convolutional neural networks is commonly performed using the Cross-Entropy Loss on the softmax output of the network. However, this loss treats all classes independently – the relationship between adjacent maturity stages can not be modeled using this loss. This can be mitigated by adapting the optimization target, which is usually a one-hot encoded vector representing the ground truth class. We propose two alternative ground truth encodings which are based on (1) declining weights and (2) on the Gray code to incorporate the biological dependencies into the target vector. Both methods are applied by computing the binary cross-entropy loss on a sigmoidal activation of the network output.

According to the taxonomy of Gutiérrez et al. [6], these approaches are similar to *cost-sensitive classification* techniques. They differ, however, in the encoding, which in this work is specifically designed to work with a typical deep learning architecture and the corresponding loss functions.

Declining Weights. Instead of using the one-hot encoding to represent a class (e.g. $[0, 0, 1, 0, 0, 0]$ to represent the myelocyte), we propose an encoding that additionally assigns a smaller number to adjacent classes. Specifically, each class c is represented based on the distance to the ground truth class c_{gt} using $w_{\text{dec}}^{|c-c_{\text{gt}}|}$. The distance here refers to the number of cell types between a class (inclusive) and the ground truth class (exclusive), such that adjacent classes have a distance of 1). For instance, with $w_{\text{dec}} = \frac{1}{2}$ and $c_{\text{gt}} = 2$ (myelocyte), this yields $[\frac{1}{4}, \frac{1}{2}, 1, \frac{1}{2}, \frac{1}{4}, \frac{1}{8}]$ as an encoding. Consequently, mis-classifications between adjacent classes yield lower losses than mis-classifications between distant classes.

Compared to similar ordinal classification techniques [8], we encode the relationship between classes not as a *cost* but as an *acceptability* measure.

Gray Code. As another alternative to one-hot encoding, we propose utilizing the *Reflected Binary Code* (RBC), also known as the Gray code [4]. This code has

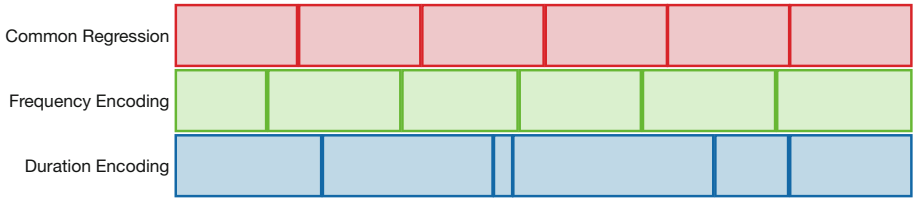

Fig. 2. Target intervals for regression techniques.

the advantage of differing only in one bit in the encoding of adjacent numbers. This encoding allows the representation of n classes in $\lceil \log n \rceil$ bits. Due to the property of RBC, adjacent classes have similar encodings. For the six classes in this dataset, the first six of the eight possible encodings with the required three bits are used. For instance, a myelocyte ($c = 2$) is encoded as $\text{RBC}(c) = [0, 1, 1]$. Again, this code results in lower losses if adjacent classes are mis-classified.

2.3 Regression Techniques

Regression techniques predict a continuous number (in this case correlating to the grade of maturity) instead of a class. The simplest implementation of regression would be to assign an integer number (0–5) to each class. Since the order of the classes is known but the distance is not, we propose using domain knowledge to obtain a more suitable encoding. We propose using either the cell frequencies or biological knowledge about the cell types.

In addition to the established regression technique, we design specific targets for the given use-case. These are not encoded as scalars, as is commonly done, but as intervals. We further investigate different ways of sampling training values from these intervals as well as handling predictions in correctly predicted intervals.

Both proposed intervals are illustrated in Fig. 2.

Frequency-Based Regression. This method utilizes the number of samples per class to obtain a more suitable encoding. More precisely, we compute for each class c a value $f_c = \log \frac{n_c}{\bar{n}}$ based on the number of samples of that class n_c and the average number of samples of all classes $\bar{n}$. In this case, this results in the values $[0.74, 1.09, 0.95, 1.00, 1.09, 1.12]$. Each class c is assigned to a range determined by the cumulative sum to $[\sum_{i<c} f_i, \sum_{i \leq c} f_i]$, which starts at 0 for $c = 0$ and is unlimited for the last class. In order to determine a label in training, we either draw a number at random from this range or use the center.

This encoding ensures that classes with fewer samples are mapped to a smaller range of values.

Stage Duration. During the maturation process, cells at each stage take a specific amount of time to develop. According to the literature, typical times are

$24, 28, 3, 33, 12$ and 20 h for each stage, respectively. These numbers can be used in a similar way as the numbers f_c from the previous section. Ranges can be defined in the same way and training labels can be drawn at random from the range or fixed to the center.

This encoding ensures that classes with shorter duration in the bone marrow are mapped to a smaller range of values.

Loss Computation. In both frequency-based and duration-based encoding, each cell type is represented by an interval of numbers. We analyze four different options on defining a loss for the training process. First, we use the Mean-Squared-Error (MSE) loss between prediction and the mean of the ground truth interval. Second, we use MSE loss between prediction and a randomly drawn number from within the ground truth interval. Furthermore, we use the same two options but set the loss to 0 if the prediction lies within the correct interval.

2.4 Experimental Setup

The base network architecture is a DenseNet-121 [7] pre-trained on ImageNet [10], which showed excellent results in similar tasks [5]. We train with a batch size of 64 with image patches of size $224 \times 224\,\mathrm{px}^2$ normalized to zero mean and unit variance with respect to the ImageNet data. Training is performed to a maximum of 256 epochs and stopped early if the validation score has not improved for 128 epochs. For the final evaluation on the test set, we use the network from the epoch with the highest validation score, which is the macro F1-score. In total, we train the network in five-fold cross-validation (four sets for training, one set for validation). Each trained network is evaluated on the previously excluded part of the dataset with annotations by multiple experts.

For the evaluation, we derive three different measures (*any*, *most* and *MTA*) based on the F1-score which differ in the matching between prediction and the ground truth labels. *Any* results in a true positive if the prediction matches any of the ground truth labels. *Most* results in a true positive only if the prediction matches the most frequent ground truth label (or any of the most frequent ground truth labels in case of a tie). *MTA* results in a true positive only if the prediction matches the label from the team of *MTA*s who also annotated the training and validation datasets. Labels by other experts are ignored in this evaluation mode.

Next to common classification (denoted as *CLF* in Fig. 3) and regression (*REG*), we evaluate the presented methods in different configurations. For classification-based methods, we evaluate the Gray Code method (*RBC*) and the Declining weights method (d_w) with weights $w \in [0.1, 0.2, 0.5]$. For the regression-based methods, we evaluate both frequency-based (f) and duration-based (t) regression. For each of these, we evaluate choosing the labels as mean (denoted by μ as index) or random (r). We further test whether it is beneficial to set the loss to zero if the predicted number falls into the correct interval. The *other* label does not contribute to the loss if it is encountered during training. It

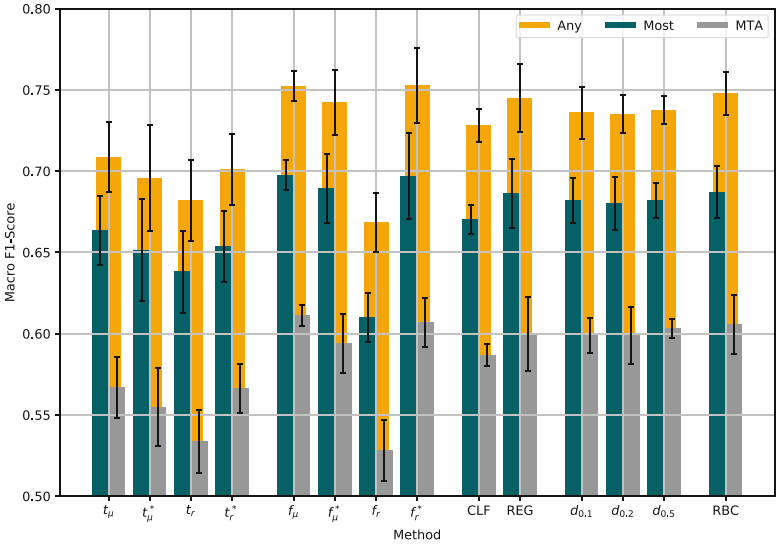

Fig. 3. Resulting F1-Scores (y-axis) for the various methods (x-axis). t refers to duration-based regression, f to frequency-based regression, with μ indicating choosing the mean as a label, and r indicating random drawing. An asterix (*) means that the loss is set to zero if the interval is correctly predicted. *CLF* and *REG* refer to common classification and regression, respectively. d_w denotes the Declining weights and *RBC* the Gray Code method. Colors denote different evaluation metrics as described in Sect. 2.4.

is further ignored in the evaluation (even though false predictions of *other* are still counted as false negatives).

3 Results

Figure 3 shows the evaluation in terms of F1-scores for each method. For all methods, the F1-Score using *most*-matching is higher than *MTA*-matching. In the following, differences in F1-score are given in percentage points (p.p.) rather than percentage.

In terms of common classification and regression methods, regression performs better with increases of 1.7, 1.6 and 1.3 for *any*, *most* and *MTA*-matching, respectively. While the Declining weights method yields better results than common classification (improvements of 1.0, 1.2, 1.6), scores are slightly lower (by 0.7, 0.2, 0.6) compared to common regression. There are only minor differences for different weights. The Gray Code method, however, slightly exceed the common regression scores as well (by 0.3, 0.1, 0.6).

Of the regression methods, frequency-based regression performs generally better than duration-based regression, with the exception of random drawing without setting the loss to zero in the correct interval. For both methods, random drawing performs better with setting the loss to zero if the prediction lies

in the correct interval. Whereas, using the mean performs better if the loss is not set to zero. Depending on the matching, two or three of the frequency-based regression methods outperform common classification and regression. The largest improvement (of 0.7, 1.2, 1.1) to common regression can be observed for frequency-based regression using the mean.

4 Discussion

The results highlight properties of the data as well as the ordinal classification and regression techniques.

The fact that generally higher scores are achieved for *most* rather than for *MTA*-matching indicates that cases exist, in which two raters contradict the *MTA* annotation. The former is used as a ground truth for training and would be used for evaluation if no multi-rater annotations were available. This highlights the importance of taking the inter-rater variability into account both for evaluation and for training. For evaluation, differences of approximately 0.15 in terms of macro F1-Score can be observed in the evaluation of this scenario. A correct and medically relevant interpretation of results and, consequently, of the quality of newly developed methods needs to take this into account. However, trends between evaluation metrics are generally similar such that using only the *MTA*-label as an approximation is a valid choice for comparison between methods. In future work, the inter-rater disagreement ought also to be reflected in the training process.

The importance of taking the ordinal nature of the data into account is further supported by performance differences between common regression and classification. Without further adaption to the data, regression already outperforms common classification by taking the order of classes into account.

Regarding ordinal classification techniques, both approaches presented improve results compared to common classification. The Declining weights method, however, does not reach the performance of common regression. It is furthermore interesting to note that the choice of weights in the evaluated range only has a negligible influence on the results. Even a small weight of $w = 0.1$ already performs consistently better than common classification. A larger weight results in smaller variance between folds and very slight F1-Score improvements. The Gray Code method performs better compared to both common classification and regression. Furthermore, the variance of the results is slightly lower than with regression. Both approaches can easily be transferred to other classification tasks with ordinal data and require no restructuring of the network backbone.

Of the two regression techniques, frequency-based regression generally yields superior results. This can, in part, be explained by the comparatively short duration of cells in the third maturity stage, which may lead to predictive difficulties using duration-based regression. Furthermore, the variation of interval lengths is much larger than in the frequency-based approach. In both approaches, setting the loss to zero within the correct predicted interval is beneficial if random labels are drawn, but not if the mean is used. While the mean "pulls" the prediction towards the interval center, this is not the case for randomly chosen labels,

which even fluctuate for each sample in every epoch. Frequency-based regression outperforms all other approaches in both random sampling and interval mean. The latter achieves this with lower variance, which makes it the most superior method. Frequency-based regression can also be easily transferred to other applications of ordinal classification, as no external domain knowledge but rather a property of the dataset itself is used.

The methods presented are easily applicable to other applications with ordinal data and adapting them to an established network architecture is straightforward. Taking ordinality into account is particularly beneficial for mitigating inter-rater variability: inter-rater disagreement is usually of lower severity (i.e., "off-by-one" disagreements are most likely). By directly incorporating and minimizing the severity of mis-classifications into the training process through ordinal methods, the network becomes capable of learning a similar behavior. While this does not necessarily increase measures such as the F-score (which treats all misclassifications the same when not working with a multi-rater dataset, as used in this work), it improves the clinical soundness of results. It also becomes easily possible to identify "borderline-cases", either by regression values close to class thresholds or through the softmax values for the predicted classes, and validate them in a post-processing step.

Compared to the non-public data reported in related works, the dataset described in this paper is more challenging. Whereas Song et al. [12,13] only differentiate between two lineages, we focus explicitly on the maturity progression within those lineages. The dataset by Choi et al. [3] includes these classes but does not include labels from multiple independent raters (only a label confirmation by a second rater).

Even though they evaluate the most commonly considered cell classes, the image data by Chandradevan et al. [2] is selected to be as simple as possible for classification. This makes the applicability of this dataset in real world clinical applications doubtful.

5 Conclusion

We proposed and evaluated several methods for the handling of ordinal data using various encodings for different regression and classification techniques. All methods are straight-forward to implement without any changes to the network backbone. We show that several of these techniques, particularly Gray Code based classification and frequency-based regression, improve ordinal classification results on a challenging hematopoietic cell dataset. This is supported by an analysis on a dataset annotated by multiple experts. This study highlights the importance of taking into account dependencies between classes and noisy labels between adjacent classes.

References

1. Binder, T., Diem, H., Fuchs, R., Gutensohn, K., Nebe, T.: Pappenheim stain: description of a hematological standard stain - history, chemistry, procedure, artifacts and problem solutions. J. Lab. Med. **36**(5), 293–309 (2012)
2. Chandradevan, R., et al.: Machine-based detection and classification for bone marrow aspirate differential counts: initial development focusing on nonneoplastic cells. Lab. Invest. **100**(1), 98–109 (2019)
3. Choi, J.W., et al.: White blood cell differential count of maturation stages in bone marrow smear using dual-stage convolutional neural networks. PLoS ONE **12**(12), e0189259 (2017)
4. Frank, G.: Pulse code communication. US Patent 2,632,058, 17 Mar 1953
5. Gräbel, P., et al.: Systematic analysis and automated search of hyper-parameters for cell classifier training. In: IEEE International Symposium on Biomedical Imaging (ISBI) (2020)
6. Gutiérrez, P.A., Perez-Ortiz, M., Sanchez-Monedero, J., Fernandez-Navarro, F., Hervas-Martinez, C.: Ordinal regression methods: survey and experimental study. IEEE Trans. Knowl. Data Eng. **28**(1), 127–146 (2015)
7. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4700–4708 (2017)
8. Kotsiantis, S.B., Pintelas, P.E.: A cost sensitive technique for ordinal classification problems. In: Vouros, G.A., Panayiotopoulos, T. (eds.) SETN 2004. LNCS (LNAI), vol. 3025, pp. 220–229. Springer, Heidelberg (2004). https://doi.org/10.1007/978-3-540-24674-9_24
9. Liu, Y., Kong, A.W.K., Goh, C.K.: A constrained deep neural network for ordinal regression. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 831–839 (2018)
10. Russakovsky, O., et al.: ImageNet large scale visual recognition challenge. Int. J. Comput. Vision **115**(3), 211–252 (2015)
11. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
12. Song, T.H., Sanchez, V., ElDaly, H., Rajpoot, N.: Simultaneous cell detection and classification in bone marrow histology images. IEEE J. Biomed. Health Inform. **23**, 1469–1476 (2018)
13. Song, T.H., Sanchez, V., Eldaly, H., Rajpoot, N.M.: Hybrid deep autoencoder with curvature gaussian for detection of various types of cells in bone marrow trephine biopsy images. In: 2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017), pp. 1040–1043. IEEE (2017)