Vers une quantification de l'interprétabilité des espaces latents pour la classification ordinale

Mouad ZINE EL ABIDINE[1], Helin DUTAGACI[1,2], David ROUSSEAU[1]

[1]Université d'Angers, LARIS, UMR INRAe IRHS, 62 Avenue Notre Dame du Lac, 49000 Angers, France

[2]Electrical-Electronics Engineering, Eskisehir Osmangazi University, Eskisehir, Turkey

david.rousseau@univ-angers.fr

Résumé – Dans cette communication, nous proposons une extension d'un travail récent sur les espaces latents dans le cadre de la classification ordinale. Nous rappelons la méthode de réduction de dimension que nous venons de développer pour ce problème et introduisons deux nouvelles métriques permettant de quantifier l'interprétabilité, i.e. l'ordinalité de ces espaces après une étape de réduction de dimension.

Abstract – In this paper, we propose an extension of a recent work on latent spaces in the ordinal classification framework. We mention the dimension reduction method we have just developed for this problem and introduce two new metrics for quantitative interpretation of the ordinality of latent spaces after a dimension reduction step.

1 Introduction

Ordinal classification refers to the classification problems where there is a natural order between categories [1]. The categories are usually represented with one-dimensional discrete values following their inherent order. It is expected that the features used to predict the ordinal categories of the instances also possess an intrinsic order in the high-dimensional space. In order to visualize and assess whether these features follow the ordinality of the categories, dimensionality reduction can be used. Although many techniques can be very useful for dimension reduction none of the classical ones do incorporate the ordinal structure of the categories into their original formulation for ordinal classification problems. Recently, we have introduced a dimension reduction technique, called Best View Point (BVP), especially suited for ordinal classification [2]. In [2], BVP was successfully compared to other classical reduction techniques for visual assessment of the quality of the dimension reduction. In this communication we extend these results and introduce two new metrics to quantify the ordinality after applying dimension reduction. We first shortly review BVP, define our new metrics and then apply them on real ordinal datasets.

2 Best view point dimension reduction

Let us consider an ordinal classification problem, where the features are in 3D space and the categories are ordered. We would like to find a viewpoint on the view sphere such that when viewed from that point the adjacent class centers seem as apart as possible from each other. Our BVP method finds the optimum viewpoint that maximizes the projected square distances between adjacent class centers and projects the data points to the space defined by the optimum viewpoint. To generalize the problem for N-dimensional space, let us first suppose that we have K classes, ordered and identified as $k = 1, 2, ..., K$. A class l is adjacent to class k if $l = k-1$ or $l = k+1$. The instances of class k are represented as N-dimensional column vectors denoted as $x_i^k \in \mathbb{R}^N$, with $i = 1, 2, ..., I_k$, where I_k is the number of instances in class k. The class centers are denoted as c_k corresponding to the arithmetic mean of the instances in class k. For the sake of simplifying the equation of the view sphere, the data is translated beforehand such that the origin of the N-dimensional space corresponds to $\frac{1}{K}\sum_{k=1}^{K} c_k$, i.e. the mean of the class centers. Let us define the n-sphere ($n = N-1$) in the N-dimensional space as $S = \{v \in R^N : \|v\| = 1\}$. Given a viewpoint $v \in S$, we can define an orthogonal projection $P : R^N \to R^N$, whose $N-1$ columns are defined by the vectors orthonormal to v, and whose last column is equal to v. Then, a point $x \in R^N$ can be projected to the $N-1$-dimensional space defined by v by computing $y = Px$ and dropping the last component of y. This point, $x(v) \in R^{N-1}$ can be interpreted as point x as seen from the viewpoint v. Its component parallel to v is invisible to the viewer. Our objective is to find the viewpoint $v*$ on the n-sphere such that the sum of the squared distances between the centers of the adjacent classes is maximized. If we define $\bar{c}_k(v) \in \mathbb{R}^{N-1}$ to be the projected center of class k in the $N-1$-dimensional space defined by viewpoint v, we search for $v*$ maximizing

$$G(v) = \sum_{k=1}^{K-1} \left\|\bar{c}_{k+1}(v) - \bar{c}_k(v)\right\|^2 \tag{1}$$

subject to the constraint $\|v\| = 1$. Maximizing G(v) is equivalent to solving the following minimization problem:

$$\text{Minimize } F(v) = \sum_{k=1}^{K-1} [v^T(c_{k+1} - c_k)]^2 \text{subject to } \|v\| = 1 \ . \quad (2)$$

For more details on the implementation of BVP, the reader can refer to [2].

3 Ordinality metrics

To go beyond the sole visualization of the dimension reduction, we now target to quantify ordinality in the reduced latent space. In the literature, several researches have developed specific metrics dedicated to ordinal classification problems [1, 3–5]. Nevertheless, these references focus on the characterization of the classification performance themselves while we care here about the interpretability of the latent space before classification. Related works [6] proposed a framework, to solve the performance versus interpretability trade-off in the context of ordinal problems. Although this work covers interpretability of ordinality, it is related to In-Model and equation-Model interpretability techniques [5] while we focus on Pre-Model here. As the most related work [7], focused on the intersection between instances of ordinal data in the latent space. In this investigation, the authors proposed a projection method from N-dimensional latent space to 1-dimensional latent space, using insights about the class distribution obtained from pairwise distance calculation between instances of all classes. The idea in [7] is to project an instance in the 1D interval of a given class, according to its distance to instances of other classes. A threshold is set manually, to split the interval on segments of classes. The output projection is then used to perform an ordinal regression. By contrast to the pairwise method in [7], we propose metrics to quantify the intersection between classes by penalizing the ordinal distance of these misclassifications. This can not be deduced from [7], where all instances are mapped in their corresponding class interval. In addition, unlike the thresholds selected manually in [7], our proposed metric is fully automatic. Moreover, we complement our new metric of ordinal intersection between classes with another metric assessing the ordinality at the level of the centroids of the clusters of the classes. This can help to discriminate ordered and unordered latent spaces for noise-free datasets having no class intersection in the latent space. This aspect was not taken into account in [7] which assumes that centroids in latent space are already well aligned and does not quantify their order in this latent space. We detail the expression of these two metrics in the next subsections.

3.1 Deviation from ordinality

The deviation from ordinality (DFO) is a metric that quantifies how much the order of the centroids departs from the expected order after dimension reduction. Technically, it compares the position of centroids in the path connecting them in the expected order $k = \{1, \ldots, K\}$ (reference path), with the position of the same centroids in the shortest path. The shortest path is a path where the nearest centroids are connecting to each other based on the Euclidean distance. This metric can be considered as an edit distance metric. Several edit distance metrics already exist in the literature [8]. Here, we propose a simple binary output: ordered or unordered. The mathematical formulation of deviation from ordinality metric is a simple subtraction between order of centroids in shortest and reference paths

$$DFO^k = \frac{\left|p_k^{ref} - p_k^{short}\right|}{K-2} \ , \quad (3)$$

where p_k^{ref} is the position of centroid k in the reference path and p_k^{short} is the position of centroid k in the shortest path. The subtraction is normalized by the maximum distance $K-2$. Centroid of class 1 is chosen as the starting point for the reference path and the shortest path, hence the normalization by $K-2$ provides a metric between 0 and 1. DFO^k is 0 when the centroid k has the same position in both paths (ordered case) while DFO is 1 if the centroid k is displaced to position K (the extreme case). An average value of the DFO^k over all k can then be computed to provide a global assessment in addition to the local order DFO^k associated to each class.

3.2 Inter-class intersection

We now introduce a second metric to quantify ordinality coined as *Inter-Class intersection*. The metric has two outputs: the first scalar quantifies the severity of intersection between classes and the second is a binary scalar that states if the intersection is only between adjacent classes or also between non-adjacent classes. The definition of intersection between classes depends on the decision boundaries for classes. In this work, we assume elliptic regions for data reduced to two dimensions to account for second order statistics of the data. The inter-class intersection metric we propose can be generalized to higher dimensions (e.g. 3D ellipsoids) and other decision boundaries such as polygonal shapes. Here, the computation of the *Inter-Class intersection* is performed in the following. First, a boundary B_k (ellipse) is computed around each class l_k (algorithm 1). The ratio a_j^k of instances x_i^k inside the boundary B_k is evaluated by counting the number of instances of the class l_k inside B_k, normalized by the dimension I_k. The output is a confusion matrix $K \times K$. The second step is to penalize boundaries in the confusion matrix, containing instances of non-adjacent classes (equation 4). This is achieved by multiplying the ratio a_j^k by the square distance $(j-k)^2$ in the matrix $K \times K$. A normalization is applied on all matrix-elements via dividing them by the sum of square distances $(j-k)^2$, so that the *Inter-Class intersection* value is between 0 and 1, as given in

$$IC^{B_k} = \frac{\sum_{j=1}^{K} a_j^k \times (\left|(j-k)\right|)^2}{\sum_{j=1}^{K} (j-k)^2} \text{for k=\{1,...,K\}} \ . \quad (4)$$

Inter-Class intersection IC^{B_k} equals to 0 when there is no intersection between classes. To be able to separate the case of intersection only between adjacent classes and intersection between non-adjacent classes, we add a complementary information through a binary scalar (BS). If all the non-diagonal and the non-adjacent cell values to the matrix $K \times K$ are equal to 0, the binary output is equal to 0. Otherwise, the binary output is equal to 1.

Algorithm 1: Pseudo-code to compute the *Inter-Class intersection* metric.

```
Data: Coordinates of instances $x_j^k$ of all classes $l_k$.
Result: KxK matrix containing the percentage of
        instances $x_j^k$ of all class $l_k$ in each boundary $B_k$
1 Fit an ellipse $B_k$ over instances $x_j^k$, by computing the
    covariance matrix and eigen vectors and value; find
    instances $x_j^k$ of class $l_k$ inside the boundary $B_k$;
2 Normalize the number of instances $x_j^k$ found by the
    dimension $I_k$ of the class $l_k$;
3 Save all ratios in an KxK matrix, where K is the number
    of classes;
```

4 Data sets

We tested our dimensionality reduction technique on real ordinal classification datasets [9, 10]. The datasets and their properties are given in Table 1.

TAB. 1: Real ordinal datasets used for the experiments [9, 10] (I is the total number of instances, Q is the dimensionality of the original data and K is the number of classes).

Dataset	I	Q	K	**Class Distribution**
pasture	36	25	3	(12,12,12)
bondrate	57	37	5	(6,33,12,5,1)
contact-lenses	24	6	3	(15,5,4)
newthyroid	215	5	3	(30,150,35)
squash-stored	52	51	3	(23,21,8)

5 Results on ordinality metrics

The two ordinal metrics of the previous section have been applied on the ordinal data of Table 1 after dimension reduction by PCA, TSNE, LDA, ISOMAP, MDS, LSDA and our proposed method BVP. The quantitative results are provided in Tables 2-6. The quantitative results are in accordance with the qualitative visualizations provided in our recent paper [2]. It appears that BVP is providing good results with almost no deviation from ordinality and low mean inter-class intersection. By comparison with the other classical dimension reduction methods, BVP provides better results than PCA, TSNE, ISOMAP and MDS. The closest quantitative results of BVP with existing methods is with LDA and LSDA. On some datasets (bondrate) BVP shows a deviation from ordinality not committed by LDA and LSDA. However, in other dataset (contact lenses) BVP outperforms LDA and LSDA. This demonstrates the complementary role of BVP in relation with the existing literature on dimension reduction. It is important to underline the quality of BVP on the Inter-class intersection metrics. Indeed, BVP is designed based on a metric applied on the centroids of the cluster and does not take into account the dispersion around these clusters. The encouraging results found on Inter-class intersection indicates that BVP also has a potential to be used for classification purposes. This is also in agreement with its good performance in comparison with LDA, which is specifically designed for classification. These are interesting perspectives on which we currently work.

6 Conclusion

We proposed the quantitative evaluation of a new and intuitive technique for the visualization of high-dimensional data for ordinal classification (BVP). We provided two new metrics to quantify the ordinality in the latent space after dimension reduction by this technique which confirms its interest and complementarity in comparison with the state of the art.

References

[1] E. Frank and M. Hall, *A Simple Approach to Ordinal Classification*. Berlin, Heidelberg: Springer-Verlag, 2001, p. 145–156.

[2] M. Zine-El-Abidine, H. Dutagaci, and D. Rousseau, “Dimensionality reduction for ordinal classification,” in *2021 29th European Signal Processing Conference (EUSIPCO)*. IEEE, 2021, pp. 1531–1535.

[3] J. S. Cardoso and J. F. Costa, “Learning to classify ordinal data: The data replication method,” *Journal of Machine Learning Research*, vol. 8, no. Jul, pp. 1393–1429, 2007.

[4] J. S. Cardoso and R. Sousa, “Measuring the performance of ordinal classification,” *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 25, no. 08, pp. 1173–1195, 2011.

[5] D. V. Carvalho, E. M. Pereira, and J. S. Cardoso, “Machine learning interpretability: A survey on methods and metrics,” *Electronics*, vol. 8, no. 8, p. 832, 2019.

[6] J. P. Amorim, I. Domingues, P. H. Abreu, and J. A. Santos, “Interpreting deep learning models for ordinal problems,” *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, 2018.

[7] J. Sánchez-Monedero, P. A. Gutiérrez, P. Tiňo, and C. Hervás-Martínez, “Exploitation of pairwise class distances for ordinal classification,” *Neural computation*, vol. 25, no. 9, pp. 2450–2485, 2013.

[8] E. S. Ristad and P. N. Yianilos, “Learning string-edit distance,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 5, pp. 522–532, 1998.

TAB. 2: *Inter-Class intersection* and *Deviation from ordinality* values extracted from latent spaces generated after applying dimension reduction techniques (**P**rincipal **C**omponent **A**nalysis, **T**-distributed **S**tochastic **N**eighbor **E**mbedding, **L**inear **D**iscriminant **A**nalysis, **M**ultidimensional **S**caling, **L**inear **S**equence **D**iscriminate **A**nalysis and **B**est **V**iew **P**oint) on pasture dataset.

DRT	IC^{B1}	IC^{B2}	IC^{B3}	**Mean IC**	**BS**	DFO^2	DFO^3	**Mean DFO**
PCA	0.12	0.3	0.35	**0.25**	1	0	0	0
TSNE	0.3	0.4	0.448	0.38	1	0	0	0
LDA	0	0	0	**0**	0	0	0	0
ISOMAP	0.05	0.384	0.334	0.26	1	0	0	0
MDS	0.12	0.3	0.35	0.26	1	0	0	0
LSDA	0.72	0.35	0.936	0.67	1	0	0	0
BVP	0.05	0.34	0.27	**0.22**	1	0	0	0

TAB. 3: Same as Tab.2 on bondrate dataset.

DRT	IC^{B1}	IC^{B2}	IC^{B3}	IC^{B4}	**Mean IC**	**BS**	DFO^2	DFO^3	DFO^4	**Mean DFO**
PCA	0.81	0.42	0.42	0.97	0.66	1	0.5	0.5	1	0.67
TSNE	0.98	0.42	0.43	0.99	0.70	1	1	0	1	0.67
LDA	0.05	0.10	0.02	0.01	**0.04**	0	0	0	0	0
ISOMAP	0.78	0.42	0.42	0.93	0.64	1	0.5	0.5	1	0.67
MDS	0.81	0.42	0.42	0.97	0.66	1	0.5	0.5	1	0.67
LSDA	0.95	0.42	0.38	0.49	**0.56**	1	0	0.5	0.5	0.33
BVP	0.66	0.42	0.43	0.97	**0.62**	1	0.5	0.5	1	0.67

TAB. 4: Same as Tab.2 on contact lenses dataset.

DRT	IC^{B1}	IC^{B2}	IC^{B3}	**Mean IC**	**BS**	DFO^2	DFO^3	**Mean DFO**
PCA	1	0.4	1	0.8	1	0	0	0
TSNE	1	0.28	1	0.76	1	1	1	1
LDA	1	0.01	0.10	**0.37**	0	0	0	0
ISOMAP	1	0.4	1	0.8	1	1	1	1
MDS	1	0.4	1	0.8	1	1	1	1
LSDA	0.92	0.01	0.10	**0.35**	0	1	1	1
BVP	0.76	0.01	0.10	**0.29**	0	0	0	0

TAB. 5: Same as Tab.2 on newthyroid dataset.

DRT	IC^{B1}	IC^{B2}	IC^{B3}	**Mean IC**	**BS**	DFO^2	DFO^3	**Mean DFO**
PCA	0.12	0.04	0.11	**0.09**	0	0	0	0
TSNE	0.03	0.24	0.02	0.10	0	0	0	0
LDA	0.05	0.04	0.18	**0.09**	0	0	0	0
ISOMAP	0.43	0.04	0.18	0.21	1	0	0	0
MDS	0.50	0.04	0.36	0.30	1	0	0	0
LSDA	0.08	0.04	0.17	0.10	0	0	0	0
BVP	0.09	0.04	0.13	**0.09**	0	0	0	0

TAB. 6: Same as Tab.2 on squash-stored dataset.

DRT	IC^{B1}	IC^{B2}	IC^{B3}	**Mean IC**	**BS**	DFO^2	DFO^3	**Mean DFO**
PCA	0.60	0.40	0.20	**0.40**	1	0	0	0
TSNE	1	0.40	0.42	0.61	1	0	0	0
LDA	0	0	0	**0**	0	0	0	0
ISOMAP	0.60	0.40	0.21	0.40	1	0	0	0
MDS	0.78	0.40	0.24	0.47	1	0	0	0
LSDA	0.97	0.39	0.75	0.70	1	1	1	1
BVP	0.37	0.39	0.04	**0.27**	1	0	0	0

[9] J. Sánchez-Monedero, P. A. Gutiérrez, and M. Pérez-Ortiz, "Orca: A matlab/octave toolbox for ordinal regression," *Journal of Machine Learning Research*, vol. 20, no. 125, pp. 1–5, 2019. [Online]. Available: http://jmlr.org/papers/v20/18-349.html

[10] P. A. Gutiérrez, M. Pérez-Ortiz, J. Sánchez-Monedero, F. Fernández-Navarro, and C. Hervás-Martínez, "Ordinal regression methods: Survey and experimental study," *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 1, pp. 127–146, 2016.