A Study on How the Training Data Monotonicity Affects the Performance of Ordinal Classifiers

Carlos Milian[1], Rafael Bello[2], Carlos Morell[2], and Bernard de Baets[3]

[1] Universidad de las Ciencias Informáticas, Cuba
[2] Computer Science Department, Universidad Central "Marta Abreu" de Las Villas, Cuba
[3] Department of Applied Mathematics, Biometrics and Process Control, Ghent University, Coupure links 653, B-9000 Gent, Belgium

Abstract. Some classification problems are based on decision systems with ordinal-valued attributes. Sometimes ordinal classification problems arise with monotone datasets. One important characteristic of monotone decision systems is that objects with better condition attribute values cannot be classified in a worse class. Nevertheless, noise is often present in real-life data, and these noises could generate partially non-monotone datasets. Several classifiers have been developed to deal with this problem but its performance is affected when faced with real data that are only partially monotone. In this paper are studied two monotonicity measures for datasets and it is analyzed its correlation with several ordinal classifiers performance. The results allow to a priori estimate the ordinal classifier behavior when faced with partially monotone ordinal decision system.

Keyword:. Ordinal classification, monotonicity, data complexity.

1 Introduction

The problem of the ordinal classification, also called ordinal regression has attracted the interest of the machine learning field due to the fact that many prediction problems or decision taking have present the ordinal values on the decision features [1, 2, 3], and [4]. An ordinal dataset is one with an ordinal variable output. In this case, the classification could be seen as a ranking, with a preference of the type 'higher values are better'.

In the ordinal classification, there is a dataset $D = \{O_1, O_2, \ldots, O_n\}$, where each object is described by a group of features $A = \{a_1, a_2, \ldots, a_n\}$; each feature a_i has a domain with an order relation which establishes an order among the domain's values; this kind of feature is frequently called criterion. Also, each object O_i has a decision feature d, which is also a criterion; therefore the decision values have also an order which defines a preference degree. Two objects O_i and O_j can be compared on the basis of their feature vectors $(O_{i1}, O_{i2}, \ldots, O_{im})$ and $(O_{j1}, O_{j2}, \ldots, O_{jm})$ or their decision values $d(O_i)$ and $d(O_j)$.

One important aspect in this topic is the data's monotony. Because of this concept, it is established the following relation:

$$O_i \leq O_j \Rightarrow d(O_i) \leq d(O_j) \,. \quad (1)$$

which means that if the object O_j is selected over the O_i therefore the class of the object O_j must be as good as the one of the object O_i This means that if only the value of one feature a_i increase (or decrease), while the rest is un-changed; the value of the decision feature d can only increase (decrease) or remain unchanged. An object is called monotone if it does not make up a non-monotone couple with any other object, and a dataset monotone if it contains no non-monotone objects. That is, monotone classification of multi-criteria data simply means that for improving criterion scores, the rank assigned by the ranking algorithm will never get worse. Non-monotonicity is present if there exist O_i and O_j in the dataset for which $O_i < O_j$ and $d(O_i) > d(O_j)$, such instances are said to be 'non-monotone'. Monotone ordinal problems, in which monotonicity constraint is imposed on the relationship between the input variables and the ordinal output is a special, yet common, type of ordinal problem [5].

Nevertheless, noise is often present in real-life data, and these noises could generate partially non-monotone datasets. Data can contain inaccurate rank or criterion values or can be an amalgam of various sources ranked by different experts; this then leads to a training set where a 'better' instance has received a lower rank than a 'worse' instance, when training a ranking algorithm on such a training set, contradictory information will be supplied to the ranking algorithm [6]. Non-monotonicity is in conflict with the domain knowledge where not only two objects with identical feature scores should have identical labels, but additionally, that increasing scores should not lead to a decrease in label; in other words, an object should receive a label at least as good as the best label received by any object that is worse than it [7].

There are many investigations about how monotony affects the way learning algorithms work. Some of the already existent algorithms are unable to be trained by using these partially non-monotone data sets. However, there are other algorithms that select the examples suitable for learning the concepts needed, during the learning process, eliminating those examples with monotonic inconsistencies.

One alternative to face this problem has been to drop the non-monotonicity grade by means of a method for relabeling the datasets, authors in [6] and [8] studied the consequences of a non-monotone training and test a set for a general monotone classification algorithm, to discuss some ways of cleaning up a non-monotone training set and determine whether it is useful in general to do such a thing. In [7], a single-pass optimal ordinal relabeling algorithm is formulated; other algorithm was presented in [9]. Also, other method was formulated in [1].

Algorithms for this kind of classification have been developed, some of which require monotone datasets. Because a monotone instance-based ranking algorithm will classify any new instance in a way monotone w.r.t. the dataset, not all instances will be able to be classified 'correctly' [6]; therefore the need of the methods for relabeling the datasets. Some classifiers guarantee the monotonicity of subsequent predictions, while others do not; ordinal classifiers also differ from each other by the way they handle non-monotone datasets [5]. For a non-monotone classification algorithm, no such 'handicap' exists; therefore, the accuracy might misleadingly be reported higher.

The method of construction of decision trees for ordinal classification in [10] also requires monotone datasets. In other work, a method was formulated to construct monotone ordinal decision trees [11] for the case of partially non-monotone datasets, though it has to relabel some of the noisy objects during the construction of the trees. The TOMASO algorithm [12] does not accept a (partially) non-monotone training set. There are some other algorithms which are mentioned and used in other parts of this work.

In this work, has been studied the relation between the grade of non-monotonicity and the performance of some classifiers. To establish this relation there are considered some measures for measuring the non-monotonicity grade. The rest of the paper is organized as follows. In section 2, we briefly describe a set of data complexity measures which will be used to develop the experimental analysis. Section 3 presents the experiments carried out over several training sets and discusses the relation between the measures and the performance of the some ordinal classifiers. Finally, the conclusions are outlined.

2 Measures of Data Complexity

The monotonicity constraint is so important that some monotone classification algorithms cannot be trained on datasets containing this kind of noise. Research on more flexible algorithms is increasing, but the basic question that if non-monotonicity of a dataset affects the learning process, remains valid. Authors in [6], showed the grade of non-monotonicity present in two datasets, quantified by the number of nonmonotone instances, and the maximum attainable accuracy for a monotone classification algorithm (OSDL); also, the effect of relabeling the instances is presented.

The purpose of this work is showing by means of an experimental study and statistical analysis the relation between monotonicity (non-monotonicity) and the performance of the algorithm. In order to do that, some measures to measure the monotonicity of the datasets are employed, and several classifiers are used.

The problem of characterizing data by means of different measures is present in machine learning, the central idea is that high-quality data characteristics or meta-features provide enough information to differentiate the performance of a set of given learning algorithms. Different studies have been done on data complexity and meta-learning, such as [13, 14, 15 and 16]. To address this problem it is necessary to use data describing the characteristics of the datasets and the performance of the algorithms, which are called meta-data. In this work, are used some measures to characterize the grade of monotonicity (or non-monotonicity) of the datasets, and two measures about the performance of the classifiers.

Two measures are used in this study, *DgrMon* y *OM*. The first one was presented in [4] and the second is proposed by us in this paper. The degree of monotonicity *DgrMon* of a dataset *D* is defined by expression (2):

$$DgrMon = \frac{\#\,Monotone\ pairs\ (D)}{\#\,Comparable\ pairs\ (D)}\,. \tag{2}$$

The pair $(O_1; O_2)$ is called comparable if $O_1 \leq O_2$ or $O_1 \geq O_2$, according to features in A; and if the relationship defined in (1) holds, it is also a monotone pair. If all comparable pairs are monotone then $DgrMon = 1$ and the dataset is called monotone (non-decreasing by assumption).

The other degree of monotonicity OM of a dataset D is defined by expression (3):

$$OM(D) = \frac{\#Monotone\ Objects(D)}{\#\ Objects(D)}\,. \tag{3}$$

As it was stated before, an object is called monotone if it does not make up a non-monotone couple with any other object.

It was studied the performance of the following classifiers: **OLM, OSDL, B-OSDL, 2B-OSDL,** and **OCC.** The ordinal learning model (OLM) is a simple algorithm that learns ordinal concepts by eliminating non-monotonic pairwise inconsistencies [17]; the learning process is based on a rule-based, which is generated during the learning. The Ordinal Stochastic Dominance Learner (OSDL) is an instance-based monotone ranking algorithm based on the concept of ordinal stochastic dominance of which several variants exist [2] and [18]; according to [7], it (really some of its extensions such as B-OSDL, this algorithm reduces to the OSDL when the stochastic training dataset is monotone) is able to use non-monotone training sets to perform a monotone interpolation, without the need of deletion or relabelling of non-monotone samples. The Ordinal class classifier (OCC) is a meta-classifier, because it uses some other classifier, such as C4.5, k-nearest neighbor, Naive bayes, etc., as a base classifier [19], in the experiments developed in this work was used as a base classifier J48, although there were used others obtaining similar results as it will be showed afterwards.; OCC does not guarantee monotonic classifications even when it learns from monotonic data.

3 Experimental Results

In the experimental study were used nine datasets whose dimensions are described in Table1.

In Tables 2-6 is described the measures' value for each dataset and the performance reached by the algorithms using the Weka platform, as measures, the Accuracy and coefficient Kappa. Accuracy (often called Confidence*)* is the number of instances that it predicts correctly, expressed as a proportion of all instances to which it applies.

Table 1. Ordinal datasets.

Dataset	Number of features in A	Number of objects	Number of classes
Car	6	1728	4
Nursery	8	12960	5
Contraceptive	9	1473	3
ERA	4	1000	9
ESL	4	488	9
LEV	4	1000	5
Monks-3	6	432	2
SWD	10	1000	4
Balance	4	625	3

Table 2. Classifier OLM and degree of monotonicity.

Dataset/Classify	OLM			OM	DgrMon
	Correctly	Incorretly	Kappa		
Car	94.85%	05.15%	0.8827	0.9890	0.9996
Nursery	97.88%	02.12%	0.9687	0.9997	0.9999
Contraceptive	44.13%	55.87%	0.0797	0.3985	0.7375
ERA	18.30%	81.70%	0.0647	0.096	0.8237
ESL	56.14%	43.85%	0.4534	0.5819	0.9867
LEV	37.10%	62.90%	0.175	0.318	0.9472
Monks-3	43.51%	56.48%	-0.1002	0.5254	0.5426
SWD	42.20%	57.80%	0.1697	0.418	0.9294
Balance	54.88%	45.12%	0.2021	0.4608	0.3875

Table 3. Classifier OSDL and degree of monotonicity.

Dataset/Classify	OSDL			OM	DgrMon
	Correctly	Incorretly	Kappa		
Car	96.18%	03.82%	0.9181	0.9890	0.9996
Nursery	98.79%	01.21%	0.9823	0.9997	0.9999
Contraceptive	30.96%	69.04%	0.055	0.3985	0.7375
ERA	23.60%	76.40%	0.0975	0.096	0.8237
ESL	68.24%	31.76%	0.6016	0.5819	0.9867
LEV	63.10%	36.90%	0.4665	0.318	0.9472
Monks-3	52.55%	47.45%	-0.0046	0.5254	0.5426
SWD	58.70%	41.30%	0.3636	0.418	0.9294
Balance	12.32%	87.68%	0.0266	0.4608	0.3875

Table 4. Classifier B-OSDL and degree of monotonicity.

Dataset/Classify	**B-OSDL**			**OM**	**DgrMon**
	Correctly	Incorretly	Kappa		
Car	96.35%	03.65%	0.9212	0.9890	0.9996
Nursery	98.69%	01.31%	0.9808	0.9997	0.9999
Contraceptive	42.16%	57.84%	0.0882	0.3985	0.7375
ERA	23.50%	76.50%	0.0967	0.096	0.8237
ESL	69.26%	30.74%	0.6157	0.5819	0.9867
LEV	63.00%	37.00%	0.4658	0.318	0.9472
Monks-3	38.43%	61.57%	-0.2366	0.5254	0.5426
SWD	58.80%	41.20%	0.3677	0.418	0.9294
Balance	57.92%	42.08%	0.2678	0.4608	0.3875

Table 5. Classifier 2B-OSDL and degree of monotonicity.

Dataset/Classify	**2B-OSDL**			**OM**	**DgrMon**
	Correctly	Incorretly	Kappa		
Car	96.35%	03.65%	0.9212	0.9890	0.9996
Nursery	98.68%	01.31%	0.9808	0.9997	0.9999
Contraceptive	42.16%	57.84%	0.0882	0.3985	0.7375
ERA	23.50%	76.50%	0.0967	0.096	0.8237
ESL	69.26%	30.74%	0.6157	0.5819	0.9867
LEV	63.00%	37.00%	0.4658	0.318	0.9472
Monks-3	38.43%	61.57%	-0.2366	0.5254	0.5426
SWD	58.80%	41.20%	0.3677	0.418	0.9294
Balance	57.92%	42.08%	0.2678	0.4608	0.3875

Table 6. Classifier OCC (J48) and degree of monotonicity.

Dataset/Classify	**OCC (J48)**			**OM**	**DgrMon**
	Correctly	Incorretly	Kappa		
Car	92.19%	07.81%	0.8319	0.9890	0.9996
Nursery	97.04%	02.96%	0.9565	0.9997	0.9999
Contraceptive	49.83%	50.17%	0.2561	0.3985	0.7375
ERA	27.30%	72.70%	0.1399	0.096	0.8237
ESL	65.78%	34.22%	0.5685	0.5819	0.9867
LEV	61.40%	38.60%	0.4433	0.318	0.9472
Monks-3	100 %	00.00 %	1	0.5254	0.5426
SWD	58.00%	42.00%	0.3505	0.418	0.9294
Balance	75.84%	24.16%	0.5884	0.4608	0.3875

The Kappa statistic is used to measure the agreement between predicted and observed categorizations of a dataset, while correcting for agreement that occurs by chance [20].

In the following Tables 7 and 8 is shown a review of the statistic correlation between measures OM y $DgrMon$, and the performance of the algorithms, measured, by using the precision and the coefficient Kappa.

This statistical analysis was carried out using bivariate correlations in the Kendall´s tau-b correlation coefficient that is used to analyze data not following a normal distribution. This coefficient establishes that for values smaller than 0.05 the significance between the data is high.

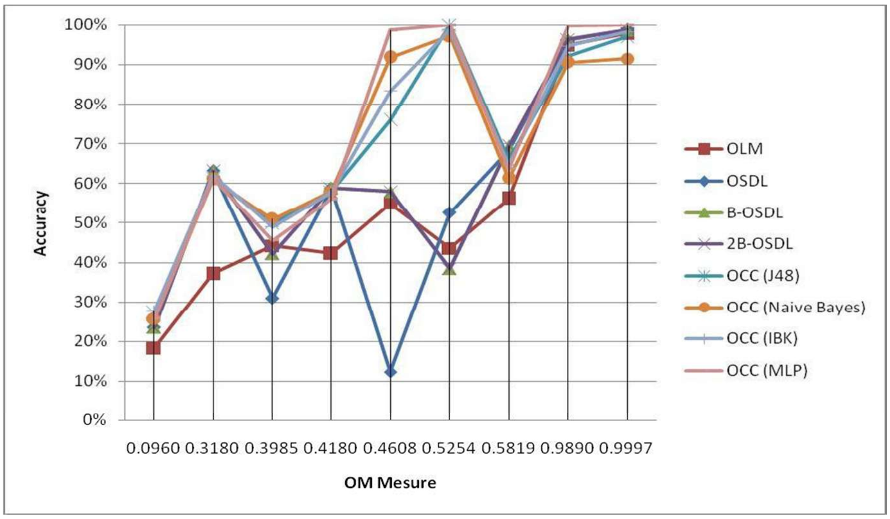

Fig. 1. Algorithms performance vs. Monotonicity measured by OM.

Table 7. Correlation between OM and the algorithms

Ordinal Classifiers	**Correlation OM and precision**	**Correlation OM and Kappa**
OLM	Significant 0.002	Significant 0.022
OSDL	Significant 0.037	Not Significant 0.211
B-OSDL	Significant 0.037	Not Significant 0.95
2B-OSDL	Significant 0.037	Not Significant 0.95
OCC(J48)	Significant 0.012	Significant 0.012

In the case of the OCC other experiments were performed using as a base classifier, classification methods k-NN and Naive Bayes, obtaining a performance similar to that obtained with the base classifier J48; Figures 1 and 2 show the performance of the algorithms.

Table 8. Correlation between DgrMon and the algorithms.

Ordinal Classifiers	Correlation DgrMon and precision	Correlation DgrMon and Kappa
OLM	Not Significant 0.211	Significant 0.012
OSDL	Significant 0.002	Significant 0.000
B-OSDL	Significant 0.007	Significant 0.002
2B-OSDL	Significant 0.007	Significant 0.002
OCC(J48)	Not Significant 0.297	Not Significant 0.297

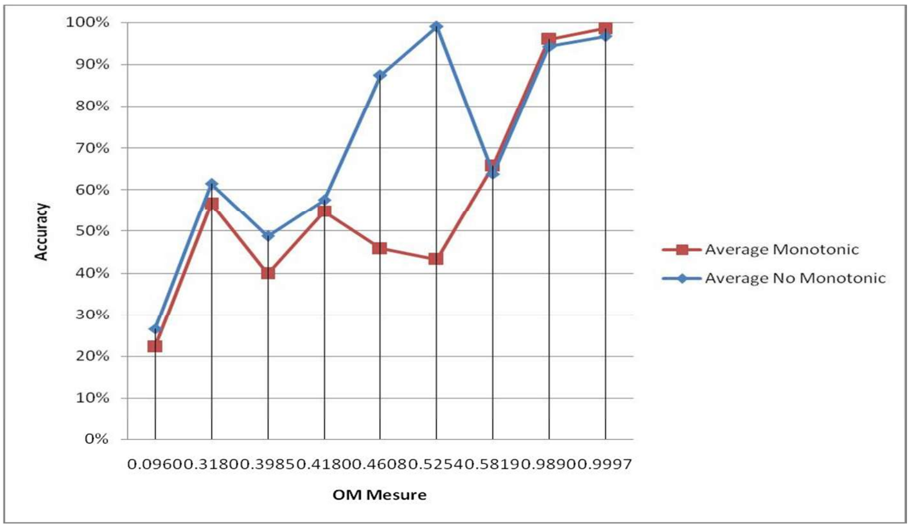

Fig. 2. Monotone and non-monotones algorithms performance vs. Monotonicity

As can be seen from the results shown in Tables 7 and 8 there is a significant statistical correlation between the measures considered to calculate the degree of monotony of the datasets and the performance of the algorithms, especially the monotony extent OM proposed in this work and the performance measure Accuracy, which is also shown in Figures 1 and 2.

The correct correlation existing between the proposed measure OM and the accuracy shown by classifiers is due to the fact that this measure is more sensible to the degree of monotony present in the datasets than the measure DgrMon because the former is more rigorous in selecting the objects to be compared.

4 Conclusions

The study on the relationship between the degree of monotony of ordinal datasets and the performance of some classifiers showed a significant correlation. The degree of monotony of the datasets was calculated by using two measures, one of which is proposed in this work and it showed a stronger relationship. This relationship allows that given a new dataset, can be estimated, by using the measures, its degree of non-monotonicity and to estimate the potential performance of the classifiers.

References

1. H. Daniels and M. Velikova.: Derivation of monotone decision models from noisy data, IEEE Trans. Syst. Man Cybern. Part C Appl. Rev., pp. 705–710 (2006)
2. S. Lievens, S., De Baets, B., and Cao-Van, K.: A probabilistic framework for the design of instance-based supervised ranking algorithms in an ordinal setting, Annals of Operations Research. 163, pp. 115–142 (2008)
3. Ben-David, A., Sterling, L., Tran, T.: Adding monotonicity to learning algorithms may impair their accuracy, Expert Systems with Applications 36, pp 667–66 (2009)
4. Velikova, M. and Daniels, H.: On Testing Monotonicity of Datasets. Ad Feelders and Rob Potharst (eds) Proceedings of MoMo2009, ¨Learning monotone models from data¨, at ECML PKDD 2009, pp. 11–22. September 7, Bled, Slovenia. (2009)
5. Potharst, R., Ben-David, A., and van Wezel, M.: Two algorithms for generating structured and unstructured monotone ordinal datasets. Engineering Applications of Artificial Intelligence 22, pp. 491–496 (2009)
6. Rademaker, M., De Baets, B., and De Meyer, H.: On the Role of Maximal Independent Sets in Cleaning Data Sets for Supervised Ranking. In Proc. of IEEE International Conference on Fuzzy Systems, Vancouver, BC, Canada. July 16-21 (2006)
7. Rademaker, M., De Baets, B., and De Meyer, H.: Optimal monotone relabelling of partially non-monotone ordinal data. Optimization Methods and Software, iFirst, pp. 1–15 (2010)
8. Rademaker, M., De Baets, B., and De Meyer, H.: Loss optimal monotone relabeling of noisy multi-criteria data sets. Information Sciences 179, pp. 4089–4096 (2009)
9. Rademaker, M., De Baets, B.: Optimal restoration of stochastic monotonicity with respect to cumulative label frequency loss functions. Information Sciences 181, 747–757. (2011)
10. Potharst, R. and Bioch, JJ.C.: Decision trees for ordinal classification, Intell. Data Anal. 4 (2), pp. 97–111 (2000)
11. Bioch, J. and Popova, V.: Monotone decision trees and noisy data, Tech. Rep. ERS-2002-53-LIS, Department of Computer Science, Erasmus University Rotterdam (2002)
12. Marichal, J.-L., Meyer, P. and Roubens, M.M.: Sorting multi-attribute alternatives: The TOMASO method, Computational Oper. Research. 32, pp. 861–877 (2005)
13. T. Ho, M. Basu.: Complexity measures of supervised classification problems, IEEE Transactions on Pattern Analysis and Machine Intelligence 24, pp. 289–300 (2002)
14. S. Singh.: Multiresolution estimates of classification complexity, IEEE Transactions on Pattern Analysis and Machine Intelligence 25 (12), pp. 1534–1539 (2003)
15. Brazdil, P. et al.: Metalearning: Applications to Data Mining. Springer, ISSN: 1611–2482, ISBN: 978-3-540-73262-4 (2009)

16. Caballero, Y., Bello, R., Arco, L., and Garcia, M.: Knowledge Discovery using Rough Set Theory, Capitulo en el libro Advances in Machine learning I Dedicated to the memory of Professor Ryszard S. Michalski in Series: Studies in Computational Intelligence Vol 262; Koronacki J., Ras Z.W., Wierzchon S.T.; Kacprzyk J (Eds), ISBN 978-3-642-05176-0 (2010)
17. Ben-David, A.: Automatic Generation of Symbolic Multiattribute Ordinal Knowledge-Based DSSs: methodology and Applications. Decision Sciences. 23:1357–1372 (1992)
18. Lievens, S., De Baets, B.: Supervised ranking in the WEKA environment, Information Sciences 180 (24), 4763–4771 (2010)
19. Frank, E. and Hall, M.: A Simple Approach to Ordinal Classification. In: 12th European Conference on Machine Learning, 145–156 (2001)
20. Witten, I. H. and Frank E.: Data mining: practical machine learning tools and techniques-2nd ed. Elsevier, ISBN: 0-12-088407-0 (2005)