IEEE
Life Sciences Letters

Received 21 March 2015; revised 10 May 2015; accepted 13 June 2015. Date of publication 30 June 2015;
date of current version 18 August 2015.

Digital Object Identifier 10.1109/LLS.2015.2451291

Exploiting Ordinal Class Structure in Multiclass Classification: Application to Ovarian Cancer

BUROOK MISGANAW (Student Member, IEEE), AND MATHUKUMALLI VIDYASAGAR (Life Fellow, IEEE)
Erik Jonsson School of Engineering and Computer Science, The University of Texas at Dallas, Richardson,
TX 75080 USA

CORRESPONDING AUTHOR: M. Vidyasagar (m.vidyasagar@utdallas.edu)

This work was supported in part by the National Science Foundation under Award ECCS 1306630, in part by the Cancer Prevention and Research Institute of Texas under Grant RP140517, and in part the Cecil H. and Ida Green Endowment through The University of Texas at Dallas, Richardson, TX, USA.

ABSTRACT In multiclass machine learning problems, one needs to distinguish between the nominal labels that do not have any natural ordering and the ordinal labels that are ordered. Ordinal labels are pervasive in biology, and some examples are given here. In this note, we point out the importance of making use of the order information when it is inherent to the problem. We demonstrate that algorithms that use this additional information outperform the algorithms that do not, on a case study of assigning one of four labels to the ovarian cancer patients on the basis of their time of progression-free survival. As an aside, it is also pointed out that the algorithms that make use of ordering information require fewer data normalizations. This aspect is important in biological applications, where data are plagued by variations in platforms and protocols, batch effects, and so on.

INDEX TERMS Ordinal classification, ovarian cancer.

I. INTRODUCTION

A STANDARD supervised machine learning problem consists of training a learner on a set of labeled samples $\{(x_i, y_i)\}_{i=1}^m$, where $x_i \in \mathbb{X} \subseteq \mathbb{R}^n$ is an n-dimensional feature vector and $y_i \in \mathbb{Y}$ is a label. In classification, the label space $\mathbb{Y}$ is a finite set. By far, the most common situation is the binary classification, where $|\mathbb{Y}| = 2$, in which case it is customary to take $\mathbb{Y} = \{0, 1\}$ or $\mathbb{Y} = \{-1, 1\}$. Consequently, the problem remains unchanged if the labels are swapped. However, when $|\mathbb{Y}| \geq 3$, this is no longer necessarily true. This situation, namely, multiclass classification with ordered label spaces, is the object of study here.

Two kinds of multiclasses are found in nature: 1) ordinal and 2) nominal classes (or labels). Ordinal labels have a natural ordering among the elements, whereas nominal labels do not. Ordinal classes are ubiquitous in biology. Sometimes ordinal categories arise naturally. In other situations, precise measurement of an underlying continuous latent variable may not be possible; therefore, a coarse discretization may be necessary, which in turn results in ordered multicategorical labels. Here are some examples.

1) Gleason grading system (1, 2, 3, 4, and 5) is a prognostic score for prostate cancer based on the microscopic appearance of the tumor.
2) Stage of the cancer (0, I, II, III, and IV) is a rough score describing the tumor size and the extent of spread of the cancer.
3) Grade of the cancer (I, II, and III) is a score describing the microscopic view of the tumor and how fast the cancer cells are growing.
4) Apgar score [1] (0–10) is the very first test for newborn infants summarizing their health condition.
5) Glasgow Coma Scale (3–15), corresponding to deep coma to fully awake, is a scoring system used to describe the conscious state of a person after a brain injury.
6) Hunt and Hess scale [2] (1–5) is a grading system used to score a diagnostic risk after neurosurgery.

Though ordinal classes are pervasive, not all algorithms take the advantage of this additional information.

In this letter, it is demonstrated that an algorithm that leverages the ordinal structure of the label space not only simplifies the learning process but also achieves a greater classification accuracy and facilitates a functional interpretation of the resulting classifier. This is established by studying the problem of assigning the ovarian cancer patients to one of four linearly ordered classes representing a response to the platinum-based chemotherapy. Two nominal (one-versus-one and one-versus-all) and one ordinal classification (data replication) methods are studied. These three standard approaches considered here decompose the multiclass problem into several binary classification problems. Note that many algorithms have been developed to handle the ordinal labels; see [3]–[5].

The rest of this letter is organized as follows. The three approaches are briefly described in Section II. Section III is devoted to the importance of a sensible normalization method for biological data. Section IV contains the implementation details and results of applying the three methods on a problem in ovarian cancer. Finally, the concluding remarks and the future directions are presented in Section V.

II. OVERVIEW OF THE THREE APPROACHES

Suppose m labeled training examples $\{(x_i, y_i)\}_{i=1}^m$ are given, where $(x_i, y_i) \in \mathbb{X} \times \mathbb{Y}$, $\mathbb{X} \subseteq \mathbb{R}^n$ is an n-dimensional feature/instance space and $\mathbb{Y} = \{1 \prec 2 \prec \cdots \prec k\}$ is a finite set of ordered elements. The objective of the ordinal classification problem is to find a function $\varphi : \mathbb{X} \rightarrow \mathbb{Y}$ that assigns labels to the feature vectors. A common approach is to determine functions $f_i : \mathbb{R}^n \rightarrow \mathbb{R}$, $1 \leq i \leq k$, such that the label assigned to a feature vector $x \in \mathbb{R}^n$ is determined by

$$\varphi(x) = \arg\max_i f_i(x). \tag{1}$$

Typically, a multiclass classification problem is decomposed into several binary classification problems, in each of which a classifier is trained to discriminate between two predetermined subsets of the label space $\mathbb{Y}$. Note that these two subsets must be disjoint, but need not together cover $\mathbb{Y}$. Then, these binary classifiers are systematically aggregated to form the multiclass classifier.

The description of these approaches is simplified using the code-matrix framework of [6], in which the code matrix consists of a $k \times l$ matrix with elements $\pm 1, 0$. The rows correspond to the class labels, while the columns represent binary classifiers. The jth binary classifier discriminates between those classes assigned $+1$ from those assigned -1, while those classes assigned 0 are ignored. Next, we discuss three of the most popular approaches.

The first method, which uses the ordinal information, is the data replication method [7], [8]. Suppose $|\mathbb{Y}| = k$. Then, $k-1$ binary classifiers are constructed, wherein the jth classifier is trained to discriminate between the samples bearing the labels $\{1, \ldots, j\}$ from those bearing the labels $\{j+1, \ldots, k\}$. In the code-matrix framework, a four-class problem can be represented as

$$\begin{array}{l} \\ \text{Class 1} \\ \text{Class 2} \\ \text{Class 3} \\ \text{Class 4} \end{array} \begin{array}{c} \begin{array}{ccc} b_1 & b_2 & b_3 \end{array} \\ \begin{pmatrix} +1 & +1 & +1 \\ -1 & +1 & +1 \\ -1 & -1 & +1 \\ -1 & -1 & -1 \end{pmatrix} \end{array}.$$

Then, class membership scores used in (1) are defined by

$$f_i(x) = \begin{cases} b_i(x), & \text{if } i = 1 \\ b_i(x) - b_{i-1}(x), & \text{if } i \in \{2, \ldots, k-1\} \\ -b_{i-1}(x), & \text{if } i = k. \end{cases}$$

The one-versus-all method, which does not make use of the ordinal information, is a common decomposition scheme, where one binary classifier is built for each class. In this approach, if $|\mathbb{Y}| = k$, then k binary classifiers are built. The jth classifier is trained to discriminate between those samples bearing the label j, and those bearing the labels $\neq j$. To illustrate, a four-class problem is decomposed as

$$\begin{array}{l} \\ \text{Class 1} \\ \text{Class 2} \\ \text{Class 3} \\ \text{Class 4} \end{array} \begin{array}{c} \begin{array}{cccc} b_1 & b_2 & b_3 & b_4 \end{array} \\ \begin{pmatrix} +1 & -1 & -1 & -1 \\ -1 & +1 & -1 & -1 \\ -1 & -1 & +1 & -1 \\ -1 & -1 & -1 & +1 \end{pmatrix} \end{array}.$$

Here, a binary classifier is trained for each class. Therefore, $f_i(x) = b_i(x)$ for all i.

The one-versus-one method, which also does not make use of the ordinal information, is another common decomposition scheme, where the multiclass problem is decomposed into all possible pairwise binary classifiers. In particular, a k-class problem is decomposed into $\binom{k}{2} = k(k-1)/2$ binary classifiers. The code matrix for a four-class problem can be written as

$$\begin{array}{l} \\ \text{Class 1} \\ \text{Class 2} \\ \text{Class 3} \\ \text{Class 4} \end{array} \begin{array}{c} \begin{array}{cccccc} b_{12} & b_{13} & b_{14} & b_{23} & b_{24} & b_{34} \end{array} \\ \begin{pmatrix} +1 & +1 & +1 & 0 & 0 & 0 \\ -1 & 0 & 0 & +1 & +1 & 0 \\ 0 & -1 & 0 & -1 & 0 & +1 \\ 0 & 0 & -1 & 0 & -1 & -1 \end{pmatrix} \end{array}.$$

Assuming that $b_{ij} + b_{ji} = 0$ (additively reciprocal or symmetric binary classifiers), the class membership scores used in (1) are aggregated as

$$f_i(x) = \sum_{j:j \neq i} b_{i,j}(x) \quad \forall i \in \mathbb{Y}.$$

Once the class membership scores are computed using one of these three methods, the classification decision rule is as in (1).

III. NORMALIZATION

One issue in multiclass classification that has not received sufficient attention is the normalization of the data. Recall that the data consist of m labeled samples of the form $\{(x_i, y_i)\}_{i=1}^m$, where each x_i is an n-dimensional vector. Thus, the values of the jth feature across all m samples form an m-dimensional vector. Most of the standard methods for binary classification, such as the support vector machine (SVM), introduced in [9] or the ℓ_1-norm SVM [10] require that each of these m-dimensional feature vectors must be normalized so as to have unit Euclidean norm. The normalization ensures that each feature is given an equal weight during the feature selection phase. In engineering problems, this type of normalization is usually sufficient, because the data are quite reliable. However, in biological problems, there are additional complications such as variations in platform and/or protocol, batch effect, and so on. Therefore, in addition to normalization, the feature vectors are also centered so as to have zero mean, that is, they are converted into a corresponding vector of Z-scores.

Now we come to the point of this section. Both the one-versus-all and the data replication approaches have the advantage that the code matrix does not contain any zero entries. Therefore, every binary classifier uses all the samples. Consequently, once the raw data are normalized or converted into Z-scores, no further transformations are necessary. Each of the binary classification problems uses exactly the same normalized data. However, in the all-pairs approach, each of the binary classifiers works with a different set of samples, as a result of which a further round of normalization and/or centering is needed for each problem. In biological data especially, such multiple normalizations can prove to be problematic. This observation is valid for both the ordinal and the nominal classification methods.

IV. CASE STUDY: OVARIAN CANCER

In this section, we apply the three approaches described in Section II to the problem of assigning ovarian cancer patients into one of four linearly ordered classes, corresponding to the number of days of progression-free survival (PFS). All computations were carried out using MATLAB by the code written by our research group.

A. RAW DATA

Gene expression profiles, measured on Agilent Custom 244K and Affymetrix U133A platform, and clinical data of 566 ovarian cancer patients were downloaded from The Cancer Genome Atlas (TCGA) website [11] on March 14, 2014. Data for two additional cohorts were obtained from the GEO database [12]: the Tothill data set (ID: GSE9891) [13] consists of 285 samples measured on Affymetrix U133 Plus 2.0 platform and the Yoshihara data set (ID: GSE17260) [14] consists of 110 samples measured on an Agilent G4112F platform.

B. TRAINING AND TESTING DATA

PFS times in the TCGA data set range from 0 to 5480 days. The 25, 50, and 75 percentiles of PFS time, corresponding to 233, 404, and 723 days, were used to divide the samples into four risk groups of sizes 142, 141, 142, and 140. Then, 70% of the samples from each of the four groups were selected at random for training. There were four distinct sets of testing data:

1) the remaining 30% of the TCGA Agilent data set that was sequestered from the training process;
2) the TCGA Affymetrix data set that consists of the same tumor samples as training data set, but measured on a different platform;
3) the Tothill data set that is a completely independent data set measured on an Affymetrix platform;
4) the Yoshihara data set that is also a completely independent data set measured on an Agilent platform.

C. PREPROCESSING

The latest probe annotation of the Agilent and Affymetrix probes were obtained from DAVID bioinformatics database [15]. The 54 159 Affymetrix probes and 35 531 Agilent probes were matched with genes on DAVID functional annotation website. First, 8014 Affymetrix and 6674 Agilent probes that are not annotated with any gene or are mapped to multiple genes are discarded.

The remaining probes were then mapped to a unique GeneID as follows. If there was only one GeneID corresponding to one probe, that probe is mapped to that GeneID. If multiple probes correspond to a GeneID, then the median value of all probes was assigned to that gene (note that some authors assign the probe with the highest variance). This resulted in 46 145 Affymetrix probes being mapped to 19 337 unique GeneIDs, and 28 857 Agilent probes being mapped to 18 756 unique GeneIDs.

D. DEVELOPMENT OF BINARY CLASSIFIERS

In order to test all the three methods discussed in the previous section, a total of 11 binary classifiers were determined, as shown in Table 1. Observe that classifiers b1, b5, and b4 are used in the data replication method, classifiers b1–b4 are used in the one-versus-all method, and classifiers b6–b11 are used in the one-versus-one method.

TABLE 1. Binary classifiers.

Binary Classifier	Positive Class	Negative Class	P-Value $\leq$	Fold-Change $\geq$	# Genes Selected
b1	{C1}	{C2,C3,C4}	0.05	1.275	24
b2	{C2}	{C1,C3,C4}	0.05	1.25	27
b3	{C3}	{C1,C2,C4}	0.05	1.325	24
b4	{C4}	{C1,C2,C3}	0.05	1.45	22
b5	{C1,C2}	{C3,C4}	0.05	1.255	25
b6	{C1}	{C2}	0.05	1.275	25
b7	{C1}	{C3}	0.05	1.35	23
b8	{C1}	{C4}	0.05	1.55	22
b9	{C2}	{C3}	0.05	1.425	23
b10	{C2}	{C4}	0.05	1.375	24
b11	{C3}	{C4}	0.05	1.5	25

The following procedure was adopted for each of the 11 binary classifiers. First, the set of genes (features) was prefiltered with two criteria: 1) standard two sample t-test and 2) fold-change. In particular, a feature was retained only if the P-value of the t-test was ≤ 0.05, and the fold-change between the averages across the classes was sufficiently large. The actual threshold was slightly varied to get about 25 genes, as shown in Table 1. It goes without saying that the set of genes used differed from one classifier to the next. Within the training data and each of the four sets of testing data, the gene expression values of each gene were normalized to zero mean ($\mu = 0$) and unit variance ($\sigma = 1$). Then, starting with the selected set of genes, a linear classifier was built using an ℓ_1-norm SVM algorithm [16].

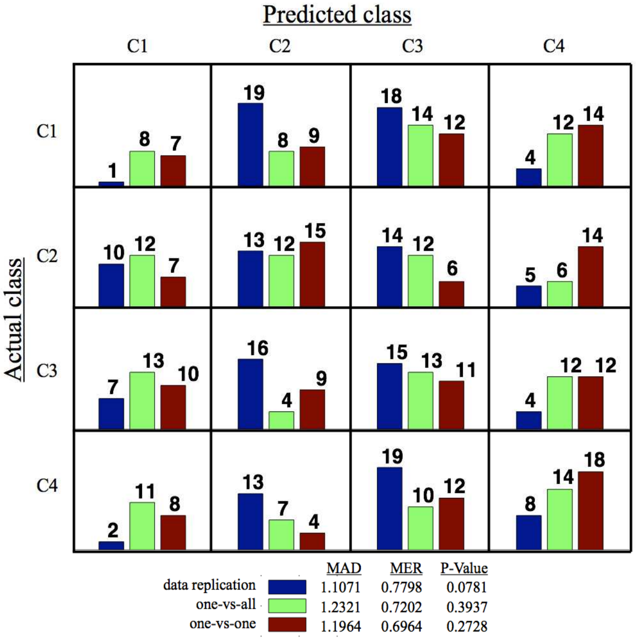
E. TESTING THE CLASSIFIERS

The performances of the three approaches were evaluated in two ways. First, we constructed a 4 × 4 contingency table of the actual class (corresponding to the rows) of each sample and the label assigned by the various methods (corresponding to the columns). These contingency tables are shown here only for the held-out 30% of the TCGA samples. In the interests of compactness, the three different contingency tables, one for each method, are combined into one table. The likelihood of generating the assignments by pure chance was computed using a χ^2-approximation. It can be seen in Table 2 that both the one-versus-all method and the all-pairs method perform very poorly. In contrast, the data replication method that explicitly takes the ordinal structure into account, despite using fewer number of binary classifiers, has better performance.

The second metric is the mean absolute deviation (MAD) rate, a variant of the misclassification error rate (MER). The MER is just the fraction of test set that is misclassified, that is MER $= 1/N \sum_{i=1}^{N} I_{\{\varphi(x_i) \neq y_i\}}$ where N is the total number of samples. However, the MER assumes that any misclassification is equally bad, which is not a valid assumption when the labels are ordered. By assigning integers $1-k$ to the labels in the ordinal set $\mathbb{Y}$, the MAD rate is defined as

$$\text{MAD} = \frac{1}{N} \sum_{i=1}^{N} |\varphi(x_i) - y_i|.$$

TABLE 2. Contingency table on held-out TCGA Agilent samples: data replication method, one-versus-all method, and one-versus-one method.



		MAD	MER	P-Value
data replication	🟦	1.1071	0.7798	0.0781
one-vs-all	🟩	1.2321	0.7202	0.3937
one-vs-one	🟥	1.1964	0.6964	0.2728

TABLE 3. MAD of the three methods on the four validation data sets.

Method \ Dataset	TCGA Agi.	TCGA Affy.	Tothill	Yoshihara
Ordinal	**1.1071**	**0.9413**	**1.0994**	**1.0000**
One-vs-all	1.2321	0.9644	1.1988	1.1636
One-vs-one	1.1964	0.9858	1.2456	1.0636

Table 3 shows the results of applying the ordinal classifier trained on 70% of the TCGA Agilent samples to the four test data sets. It can be seen that the data replication method provides better performance than the other two methods.

A simulation study in [17] demonstrated that ordinal procedures have lower error rate and higher classification power than nominal models when properly applied to ordinal data. This letter debunked a seemingly opposite conclusion from an earlier simulation study [18]. In [19], the value of order information is demonstrated empirically for learning techniques that make use of the ordinal information. To further study the impact of the ordering in the present problem, we carried out three permutations of the ordinal labels, and solved the resulting problem using the data replication method. These permutations and the resulting MAD rates for the 30% held-out TCGA Agilent samples were as follows: 1) natural order: C1, C2, C3, C4, MAD = 1.1071; 2) permutation 1: C1, C3, C2, C4, MAD = 1.1726; 3) permutation 2: C1, C4, C3, C2, MAD = 1.1131; and 4) permutation 3: C3, C2, C1, C4, MAD = 1.0893. Obviously, since order information is not utilized in the nominal methods, the other two methods would produce the same results as in Table 3. Thus, in two out of three cases, the MAD rate for the natural ordering is lower than for the perturbed ordering.

V. CONCLUSION

Despite its frequent occurrence, the ordinal nature of labels is sometimes ignored in machine learning problems, and a nominal classification algorithm is used instead. The primary motivation for using the ordering information is to improve the accuracy of the classifier. Moreover, the additional information about the relationship between classes can be used to simplify the learning process and facilitate the biological interpretation.

In this regard, additional issues that need to be further investigated include the following:

1) the comparative effect of the base binary classifiers (in this letter, we just used a linear SVM);
2) comparison between various ordinal classification algorithms;
3) feature selection for ordinal classification;
4) sensible evaluation metric for ordinal classification (because with ordinal labels, not all misclassification are equally bad);
5) the effect of the number of classes.

For instance, does ordinal information become more important as the number of classes increases?

ACKNOWLEDGMENT

The authors would like to thank their colleagues, M. E. Ahsen and N. Singh, as well as Prof. M. A. White of the UT Southwestern Medical Center, Dallas, for many valuable insights and helpful discussions.

REFERENCES

[1] V. Apgar, "A proposal for a new method of evaluation of the newborn infant," *Current Res. Anesthesia Anal.*, vol. 32, no. 4, pp. 260–267, 1953.

[2] W. E. Hunt and R. M. Hess, "Surgical risk as related to the time of intervention in the repair of intracranial aneurysms," *J. Neurosurgery*, vol. 28, no. 1, pp. 14–20, 1968.

[3] G. Bakir *et al.*, *Predicting Structured Data*. Cambridge, MA, USA: MIT Press, 2007.

[4] S. Nowozin, P. V. Gehler, J. Jancsary, and C. H. Lampert, *Advanced Structured Prediction*. Cambridge, MA, USA: MIT Press, 2014.

[5] I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun, "Large margin methods for structured and interdependent output variables," *J. Mach. Learn. Res.*, vol. 6, pp. 1453–1484, Dec. 2005.

[6] E. L. Allwein, R. E. Schapire, and Y. Singer, "Reducing multiclass to binary: A unifying approach for margin classifiers," *J. Mach. Learn. Res.*, vol. 1, pp. 113–141, Sep. 2001.

[7] E. Frank and M. Hall, *A Simple Approach to Ordinal Classification*. Berlin, Germany: Springer-Verlag, 2001.

[8] J. S. Cardoso and J. F. Pinto da Costa, "Learning to classify ordinal data: The data replication method," *J. Mach. Learn. Res.*, vol. 8, pp. 1393–1429, Dec. 2007.

[9] C. Cortes and V. N. Vapnik, "Support vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1997.

[10] P. S. Bradley and O. L. Mangasarian, "Feature selection via concave minimization and support vector machines," in *Proc. 15th Int. Conf. Mach. Learn. (ICML)*, 1998, pp. 82–90.

[11] *The Cancer Genome Atlas (TCGA)*. [Online]. Available: http://cancergenome.nih.gov, accessed Mar. 14, 2014.

[12] *Gene Expression Omnibus (GEO)*. [Online]. Available: http://www.ncbi.nlm.nih.gov/geo/, accessed Mar. 14, 2014.

[13] R. W. Tothill *et al.*, "Novel molecular subtypes of serous and endometrioid ovarian cancer linked to clinical outcome," *Clin. Cancer Res.*, vol. 14, no. 16, pp. 5198–5208, 2008.

[14] K. Yoshihara *et al.*, "Gene expression profile for predicting survival in advanced-stage serous ovarian cancer across two independent datasets," *PLoS One*, vol. 5, no. 3, p. e9615, 2010.

[15] D. W. Huang, B. T. Sherman, and R. A. Lempicki, "Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources," *Nature Protocols*, vol. 4, no. 1, pp. 44–57, 2008.

[16] M. Vidyasagar, "Machine learning methods in the computational biology of cancer," *Proc. Roy. Soc. A, Math., Phys. Eng. Sci.*, vol. 470, no. 2167, paper 81, 2014.

[17] S. M. Rudolfer, P. C. Watson, and E. Lesaffre, "Are ordinal models useful for classification? A revised analysis," *J. Statist. Comput. Simul.*, vol. 52, no. 2, pp. 105–132, 1995.

[18] M. K. Campbell, A. Donner, and K. M. Webster, "Are ordinal models useful for classification?" *Statist. Med.*, vol. 10, no. 3, pp. 383–394, 1991.

[19] J. C. Huhn and E. Hullermeier, "Is an ordinal class structure useful in classifier learning?" *Int. J. Data Mining, Model. Manage.*, vol. 1, no. 1, pp. 45–67, 2008.