A graph-theoretic approach for reducing one-versus-one multi-class classification to ranking

Willem Waegeman WILLEM.WAEGEMAN@UGENT.BE
Department of Electrical Energy, Systems and Automation, Ghent University, Technologiepark 913, B-9052 Ghent, Belgium

Bernard De Baets BERNARD.DEBAETS@UGENT.BE
Department of Applied Mathematics, Biometrics and Process Control, Ghent University, Coupure links 653, B-9000 Ghent, Belgium

Luc Boullart LUC.BOULLART@UGENT.BE
Department of Electrical Energy, Systems and Automation, Ghent University, Technologiepark 913, B-9052 Ghent, Belgium

Abstract

Can a multi-class classification model in some situations be simplified to a ranking model without sacrificing performance? We try to answer that question from a theoretical point of view for one-versus-one multi-class ensembles. To this end, sufficient conditions are derived for which a one-versus-one ensemble becomes ranking representable, i.e. conditions for which the ensemble can be reduced to a ranking or ordinal regression model such that a similar performance on training data is measured. As performance measure, we use the area under the ROC curve (AUC) and its reformulation in terms of graphs. By means of a graph-theoretic analysis of the problem, we are able to formulate necessary and sufficient conditions for ranking representability. For the three class case, this results in a new type of transitivity for pairwise AUCs that can be verified by solving an integer quadratic program.

1. Introduction

Many machine learning algorithms for multi-class classification aggregate several binary classifiers to compose a decision rule (see e.g. (Allwein et al., 2000; Fürnkranz, 2002; Rifkin & Klautau, 2004). In the popular one-versus-one ensemble, a classifier is trained on each pair of categories, but do we really need such a complex model for every multi-class classification task? For example, a one-versus-one ensemble is a very flexible model, since it can represent the trends in the data for each pair of categories separately, but, because of that, it is also a complex model, being difficult to fine tune with a limited amount of data. On the other hand, a one-versus-all ensemble has substantially less free parameters, decreasing the chance of overfitting, but increasing the chance of underfitting the data.

One might agree that different multi-class classification schemes have a different degree of complexity, but no consensus has been reached on which one to prefer. In this work we go one step further and investigate whether a one-versus-one multi-class model can be simplified to a ranking model. We start from the assumption that the optimal complexity of a multi-class model is problem-specific. Reducing a one-versus-one ensemble to a ranking model, can be seen as a quite drastic application of the bias-variance trade-off: a one-versus-one classification scheme is a complex model, resulting in a low bias and a high variance of the performance, while an ordinal regression model is a much simpler model, manifesting a high bias but a low variance. So, we do not claim that a one-versus-one scheme can always be reduced to a ranking model. We rather look for necessary and sufficient conditions for such a reduction.

Appearing in the 6^{th} *International Workshop on Mining and Learning with Graphs*, Helsinki, Finland, 2008.

2. Strict ranking representability

We start with introducing some notations in order to state the problem setting a little bit more formally. Let us assume that examples, training data as well as test data, are identically and independently drawn according to an unknown distribution over $\mathcal{X} \times \mathcal{Y}$ with $\mathcal{X}$ the object space and $\mathcal{Y}$ an unordered set of r categories in an r-class classification task. We use the notation $\mathcal{Y} = \{\overline{y}_1, ..., \overline{y}_r\}$ to denote the respective categories. Furthermore, we formally define a one-versus-one model as a set $\overline{\mathcal{F}}$ of $r(r-1)/2$ ranking functions $f_{kl} : \mathcal{X} \to \mathbb{R}$ with $1 \leq k < l \leq r$. Thus, we consider one-versus-one schemes for which each binary classifier produces a continuous output resulting in a probability estimate or a ranking of the data for each pair of categories. A dataset of size n will be denoted $D = \{(\mathbf{x}_1, y_1), ..., (\mathbf{x}_n, y_n)\}$.

Given a one-versus-one classification model, represented by a set $\overline{\mathcal{F}}$ of pairwise ranking functions, when can we reduce this model to a single ranking $f : \mathcal{X} \to \mathbb{R}$ that gives a better performance on unknown test data? Or, equivalently, when can we simplify the one-versus-one model to a ranking model without decreasing the error on training data? Having in mind the bias-variance trade-off, it would be appropriate to prefer the single ranking model over the one-versus-one scheme if the training error does not increase. In that case, the former model is complex enough to fit the data well in spite of having a lower variance over different training samples. In its most strict form, we can define ranking representability of a one-versus-one classification scheme as follows.

Definition 2.1. *Let $D \subset \mathcal{X} \times \mathcal{Y}$. We call a set $\overline{\mathcal{F}}$ of pairwise ranking functions strictly ranking representable on D if there exists a ranking function $f : \mathcal{X} \to \mathbb{R}$ such that for all $\overline{y}_k, \overline{y}_l \in \mathcal{Y}$ and any $(\mathbf{x}_i, \overline{y}_k), (\mathbf{x}_j, \overline{y}_l) \in D$*

$$f_{kl}(\mathbf{x}_i) \leq f_{kl}(\mathbf{x}_j) \Leftrightarrow f(\mathbf{x}_i) \leq f(\mathbf{x}_j)\,. \tag{1}$$

A graphical reformulation of strict ranking representability is established as follows.

Definition 2.2. *Let $D \subset \mathcal{X} \times \mathcal{Y}$ and let $\overline{\mathcal{F}}$ be a set of pairwise ranking functions. We define the graph $G_{strict}(\overline{\mathcal{F}}, D) = (V, E)$ of $\overline{\mathcal{F}}$ and D such that each node v_i in V is associated with one data object $(\mathbf{x}_i, y_i)$ in D and*

$$f_{kl}(\mathbf{x}_i) \leq f_{kl}(\mathbf{x}_j) \Leftrightarrow (v_i, v_j) \in E\,, \tag{2}$$

with $y_i = \overline{y}_k$, $y_j = \overline{y}_l$, $f_{lk} = -f_{kl}$ for $1 \leq k < l \leq r$.

Proposition 2.3. *Let $D \subset \mathcal{X} \times \mathcal{Y}$. A set $\overline{\mathcal{F}}$ of pairwise ranking functions is strictly ranking representable if and only if $G_{strict}(\overline{\mathcal{F}}, D)$ is a directed acyclic graph (DAG).*

Consequently, strict ranking representability of a set of pairwise ranking functions can be easily checked with a simple algorithm that verifies whether the corresponding graph is a DAG.

3. AUC ranking representability

It goes without saying that strict ranking representability has a very limited applicability to reduce one-versus-one multi-class schemes, since the condition is too strong to be satisfied in practice. When fitting $r(r-1)/2$ functions to the data in a multi-class setting, it is unrealistic to think that all these functions will impose a consistent ranking, i.e. a ranking satisfying Eq. (1). Yet, is it really necessary to require strict ranking representability in order to exchange a one-versus-one model for a single ranking model? The answer is no, since we are interested in a good performance on independent test data. Therefore, demanding that a single ranking gives exactly the same result on training data as a one-versus-one scheme might be a too strong condition. An obvious relaxation could exist in requiring that a single ranking model yields the same *performance* on training data instead of requiring the same results. This makes a subtle difference since it is now allowed that both models make errors on different data objects, as long as the total error of both models is similar. As claimed above, the single ranking model should attain better results on independent test data when the bias-variance trade-off is taken into consideration.

The performance measure that we will consider is the pairwise AUC, which is defined as follows

$$\widehat{A}_{kl}(\overline{\mathcal{F}}, D) = \frac{1}{n_k n_l} \sum_{y_i = \overline{y}_k} \sum_{y_j = \overline{y}_l} I_{f_{kl}(\mathbf{x}_i) < f_{kl}(\mathbf{x}_j)}\,. \tag{3}$$

If we reduce the one-versus-one model $\overline{\mathcal{F}}$ to a single ranking model $f : \mathcal{X} \to \mathbb{R}$, then we are able to compute the pairwise AUC from this simplified model:

$$\widehat{A}_{kl}(f, D) = \frac{1}{n_k n_l} \sum_{y_i = \overline{y}_k} \sum_{y_j = \overline{y}_l} I_{f(\mathbf{x}_i) < f(\mathbf{x}_j)}\,.$$

Given the definitions of $\widehat{A}_{kl}(\overline{\mathcal{F}}, D)$ and $\widehat{A}_{kl}(f, D)$, let us first introduce a more formal definition of AUC ranking representability.

Definition 3.1. *Let $D \subset \mathcal{X} \times \mathcal{Y}$. We call a set $\overline{\mathcal{F}}$ of pairwise ranking functions AUC ranking representable on D if there exists a ranking function $f : \mathcal{X} \to \mathbb{R}$ such*

that

$$\widehat{A}_{kl}(\overline{\mathcal{F}}, D) = \widehat{A}_{kl}(f, D) \quad \forall k, l : 1 \leq k < l \leq r\,. \quad (4)$$

Under which conditions can we represent the set of pairwise AUCs defined on $r(r-1)/2$ ranking functions as a new set of pairwise AUCs but now defined on a single ranking? In other words, when is a one-verus-one model AUC ranking representable. Like strict ranking representability, AUC ranking representability has a graph-theoretic interpretation.

Definition 3.2. *Let $D \subset \mathcal{X} \times \mathcal{Y}$ and let $\overline{\mathcal{F}}$ be a set of pairwise ranking functions. We define $\mathfrak{G}_{AUC}(\overline{\mathcal{F}}, D)$ as a set of complete directed graphs $G = (V, E)$ for which the following three properties hold:*

1. *Each node v_i in V is associated with one data object $(\mathbf{x}_i, y_i)$ in D.*
2. *No cycles occur in the subsets $V_k = \{v_i \in V \mid y_i = \overline{y}_k\}$.*
3. *For $1 \leq k < l \leq r$*

$$\widehat{A}_{kl}(\overline{\mathcal{F}}, D) = \frac{|\{(v_i, v_j) \in E \mid y_i = \overline{y}_k \wedge y_j = \overline{y}_l\}|}{n_k n_l}\,.$$

Remark that in a complete directed graph each pair of nodes is connected by exactly one (directed) edge. So, $(v, v') \in E$ implies $(v', v) \notin E$. Similar graph-theoretic concepts have been introduced in (Waegeman et al., 2008) to construct efficient algorithms for ROC measures in ordinal regression settings.

It follows directly from the definition that $\mathfrak{G}_{AUC}(\overline{\mathcal{F}}, D)$ cannot be an empty set. Usually, its cardinality will be greater than 1 since different graphs satisfying the three conditions in Definition 3.2 will be found for a given $\overline{\mathcal{F}}$ and D. In the following lemma, AUC ranking representability is reformulated in terms of the graph.

Proposition 3.3. *Let $D \subset \mathcal{X} \times \mathcal{Y}$. A set $\overline{\mathcal{F}}$ of pairwise ranking functions is AUC ranking representable if and only if at least one of the graphs in $\mathfrak{G}_{AUC}(\overline{\mathcal{F}}, D)$ is a directed acyclic graph.*

4. Discussion

Unlike strict ranking representability, it is far form trivial to verify whether a set of pairwise rankings $\overline{\mathcal{F}}$ is AUC ranking representable, since examining all graphs in $\mathfrak{G}_{AUC}(\overline{\mathcal{F}}, D)$ will be computationally intractable for large training samples. In the talk we will present a way to tackle the problem by using the graph concepts that we briefly introduced here. It turns out that AUC ranking representability gives evidence of strong similarities with the framework of cycle transitivity (De Baets et al., 2006). Using this framework, we are able to define necessary conditions for AUC ranking representability, since the pairwise AUCs of an AUC ranking representable one-versus-one scheme are reciprocal relations coinciding with dice models (De Schuymer et al., 2003; De Schuymer et al., 2005). These conditions can be easily verified in practice by analyzing the pairwise AUCs.

With the help of the graph formulations that we presented above, sufficient conditions for AUC ranking representability can also be translated into the framework of cycle transitivity. To this end, a new type of cycle transitivity is introduced, leading to a verifiable sufficient condition for the three class case. In this way, AUC ranking representability can be checked by solving an integer quadratic program.

Acknowledgment

Willem Waegeman is supported by a grant of the "Institute for the Promotion of Innovation through Science and Technology in Flanders (IWT-Vlaanderen)".

References

Allwein, E., Schapire, R., & Singer, Y. (2000). Reducing multi-class to binary: A unifying apprach for margin classifiers. *Journal of Machine Learning Research*, *1*, 113–141.

De Baets, B., De Meyer, H., De Schuymer, B., & Jenei, S. (2006). Cyclic evaluation of transitivity of reciprocal relations. *Social Choice and Welfare*, *26*, 217–238.

De Schuymer, B., De Meyer, H., & De Baets, B. (2005). Cycle-transitive comparison of independent random variables. *Jounral of Multivariate Analysis*, *96*, 352–373.

De Schuymer, B., De Meyer, H., De Baets, B., & Jenei, S. (2003). On the cycle-transitivity of the dice model. *Theory and Decision*, *54*, 164–185.

Fürnkranz, J. (2002). Round robin classification. *Journal of Machine Learning Research*, *2*, 723–747.

Rifkin, R., & Klautau, A. (2004). In defense of one-versus-all classification. *Journal of Machine Learning Research*, *5*, 101–143.

Waegeman, W., De Baets, B., & Boullart, L. (2008). On the scalability of ordered multi-class ROC analysis. *Computational Statistics and Data Analysis*, *52*, 3371–3388.