Design and Evaluation of Large-Scale Cost-Sensitive Classification Algorithms

Professor Hsuan-Tien Lin and the Computational Learning Laboratory
Department of CSIE, National Taiwan University

March 01, 2009

The *classification* problem in machine learning aims at designing a computational system that learns from some given training examples in order to separate input instances to pre-defined categories. The problem fits the needs of a variety of applications, such as classifying emails as spam and non-spam ones automatically. Traditionally, the regular classification setup intends to minimize the number of future mis-prediction errors. Nevertheless, in some applications, it is needed to treat different types of mis-prediction errors differently. For instance, in terms of public health, if there is some infectious diseases like SARS (Severe Acute Respiratory Syndrome), the cost of mis-predicting an infected patient as a healthy one may be higher than the other way around. In an animal recognition system, the silliness of mis-predicting a person as a fish may be higher than the silliness of mis-predicting her/him as a monkey. Such a need can be formalized as the *cost-sensitive classification* setup, which is drawing much research attention throughout the years because of its many applications, including targeted marketing, fraud detection, medical decision, and web analysis (Abe, Zadrozny and Langford 2004). As shown in Table 1, there is a gap between the theoretical guarantee and the empirical performance of most of the existing cost-sensitive classification algorithms. The major topic of this research project is to fill the gap.

Our past research results (Lin 2008) were targeted towards the *ordinal ranking* setup. Instead of asking the computational system to separate input instances to

theoretical guarantee / empirical performance	none/weak	strong
bad/unclear	not useful	some algorithms (e.g. Beygelzimer, Langford and Ravikumar 2007)
okay/good	many algorithms (e.g. Margineantu 2001)	***only a few algorithms*** (e.g. Abe, Zadrozny and Langford 2004)

Table 1: current status of research on designing cost-sensitive classification algorithms

categories, ordinal ranking asks the computational system to distinguish the *ranks* of input instances. It is an important setup in machine learning for modeling our preferences. For instance, we rank hotels by stars to represent their quality; we give feed-backs to products on Amazon using a scale from one to five; we say that an infant is younger than a child, who is younger than a teenager, who is younger than an adult, without referring to the actual age. Ordinal ranking enjoys a wide range of applications from social science to behavioral science to information retrieval, and hence attracts lots of research attention in recent years.

Note that we can view ordinal ranking as a special case of cost-sensitive classification. In particular, because there is a natural order among the ranks (e.g., infants, children, teenagers, adults—ordered by "age"), the penalty of a mis-prediction depends on its "closeness." For example, the penalty of mis-predicting a child as an adult should be higher than the penalty of mis-predicting the child as a teenager. Thus, ordinal ranking can be casted as a cost-sensitive classification problem with V-shaped costs, as illustrated in Figure 1 (where costs are denoted as $C_{y,k}$).

Many machine learning algorithms are designed in recent years to understand ordinal ranking better, but the design process can be time-consuming. Our work presents a novel alternative—a reduction framework that systematically transforms ordinal ranking to simpler yes/no question answering, which is called *binary clas-*

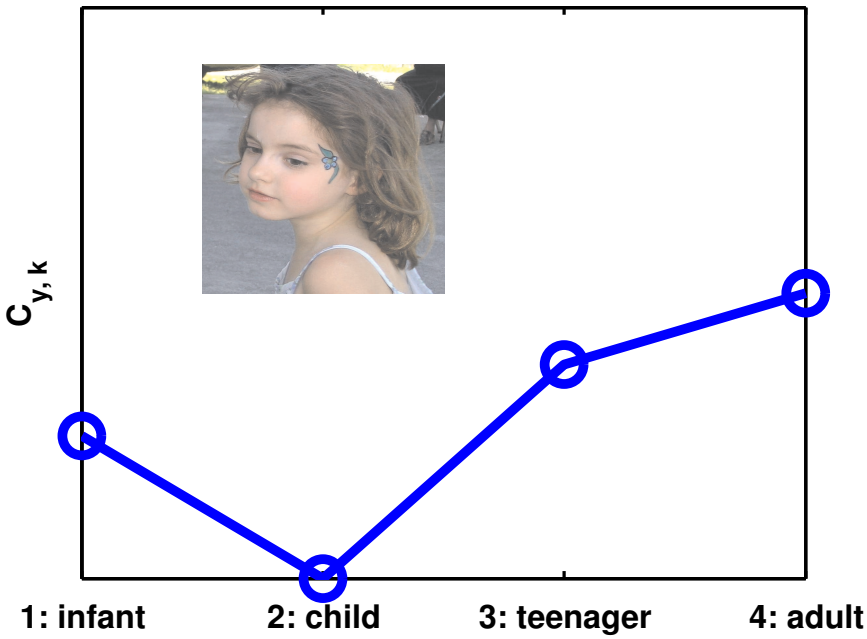

Figure 1: a V-shaped cost vector

sification (Li and Lin 2007; Lin 2008). At first glance, ordinal ranking seems more difficult than binary classification. Nevertheless, our framework reveals a surprising theoretical consequence: ordinal ranking is, in general, *as easy as* (or as hard as) binary classification (Lin 2008). Most importantly, our framework immediately brings research in ordinal ranking *up-to-date* with decades of study in binary classification. In particular, well-tuned binary classification algorithms can be effortlessly casted as new ordinal ranking ones, and well-known theoretical results for binary classification can be easily extended to new ones for ordinal ranking. Along with the reduction results, we proposed several new ordinal ranking algorithms, all of which inherited strong theoretical guarantees and empirical benefits from binary classification (Lin and Li 2006; Li and Lin 2007; Lin 2008).

Given the success stories in the special ordinal ranking setup, we are interested in extending our results to the more general cost-sensitive classification setup. One specific research question and some preliminary results are as follows.

How do we design better large-scale cost-sensitive classification algo-

rithms?

By "better", we mean better-suited for specific purposes. There is one current focus point: more efficient cost-sensitive classification algorithms when the number of categories or the number of examples is large. There is a strong need of such algorithms in real-world applications like computer vision. In computer vision, there are usually hundreds of categories in a typical object recognition problem, and there can be many training examples in total. Then, existing cost-sensitive classification algorithms either become too slow or do not perform well. Since one of the major applications of cost-sensitive classification is object recognition (e.g. human is closer to monkey than to fish), we hope to design some concrete algorithms for those applications. We have designed two novel algorithms, the "cost-sensitive one-versus-one" (CSOVO) and "cost-sensitive one-versus-all" (CSOVA). The latter is especially suited when the number of categories is large (Lin 2008).

In our previous work (Lin 2008), we have obtained the following experimental results when comparing the proposed CSOVA and CSOVO algorithms with their original versions. All these algorithms obtains a decision function by calling a binary classification algorithm several times. We take the support vector machine (SVM) with the perceptron kernel (Lin and Li 2008) as the binary classification algorithm in all the experiments and use LIBSVM (Chang and Lin 2001) as our SVM solver.

We use six benchmark classification data sets: VEHICLE, VOWEL, SEGMENT, DNA, SATIMAGE, USPS (Table 2).[1] The first five comes from the UCI machine learning repository (Hettich, Blake and Merz 1998) and the last one comes from Hull (1994).

The six data sets in Table 2 were originally gathered as regular classification problems. We follow the procedure used by Abe, Zadrozny and Langford (2004) to test the algorithms. In particular, we generate the cost vectors from a cost function $C(y, k)$ that does not depends on the input. $C(y, y)$ is set as 0 and $C(y, k)$

[1]They are downloaded from `http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets`

Table 2: Classification data sets

data set	# examples	# categories (K)	# features (D)
VEHICLE	846	4	18
VOWEL	990	11	10
SEGMENT	2310	7	19
DNA	3186	3	180
SATIMAGE	6435	6	36
USPS	9298	10	256

is a random variable sampled uniformly from $\left[0, 2000\frac{|\{n:\, y_n=k\}|}{|\{n:\, y_n=y\}|}\right]$.

We randomly choose 75% of the examples in each data set for training and leave the other 25% of the examples as the test set. Then, each feature in the training set is linearly scaled to $[-1, 1]$, and the feature in the test set is scaled accordingly. The results reported are all averaged over 20 trials of different training/test splits, along with the standard error.

SVM with the perceptron kernel takes a regularization parameter (Lin and Li 2008), which is chosen within $\{2^{-17}, 2^{-15}, \ldots, 2^3\}$ with a 5-fold cross-validation (CV) procedure on the training set (Hsu, Chang and Lin 2003). For the original OVA and OVO, the CV procedure selects the parameter that results in the smallest cross-validation regular classification cost. For the other algorithms, the CV procedure selects the parameter that results in the smallest cross-validation cost-sensitive classification cost based on the given setup. We then rerun each algorithm on the whole training set with the chosen parameter to get the decision function Finally, we evaluate the average performance of the decision function on the test set.

We compare CSOVA and CSOVO with their original versions in Table 3. We see that CSOVA and CSOVO are often significantly better than their original version respectively, which justifies the validity of the cost-transformation technique and our proposed algorithms. We intend to use the computing power of the NTU CC clusters for more large-scale experiments.

Table 3: Test cost of cost-sensitive classification algorithms

data set	one-versus-all		one-versus-one	
	OVA	CSOVA	OVO	CSOVO
VEHICLE	189.064±17.866	**158.215±19.833**	185.378±17.235	**145.745±18.404**
VOWEL	14.654±1.766	14.386±1.717	**11.896±1.955**	19.277±1.899
SEGMENT	**25.263±2.015**	**25.434±2.208**	**25.153±2.109**	**25.618±2.664**
DNA	44.480±2.771	**39.424±2.521**	48.152±3.333	51.961±4.543
SATIMAGE	93.381±5.712	77.101±4.762	94.075±5.488	**65.812±4.463**
USPS	23.087±0.709	**22.793±0.710**	23.622±0.660	**22.103±0.721**

(those within one standard error of the lowest one are marked in bold)

References

Abe, N., B. Zadrozny, and J. Langford (2004). An iterative method for multi-class cost-sensitive learning. In W. Kim, R. Kohavi, J. Gehrke, and W. DuMouchel (Eds.), *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 3–11. ACM.

Beygelzimer, A., V. Daniand, T. Hayes, J. Langford, and B. Zadrozny (2005). Error limiting reductions between classification tasks. In L. D. Raedt and S. Wrobel (Eds.), *Machine Learning: Proceedings of the 22rd International Conference*, pp. 49–56. ACM.

Beygelzimer, A., J. Langford, and P. Ravikumar (2007). Multiclass classification with filter trees. Downloaded from `http://hunch.net/~jl`.

Chang, C.-C. and C.-J. Lin (2001). *LIBSVM: A Library for Support Vector Machines*. National Taiwan University. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

Domingos, P. (1999). MetaCost: A general method for making classifiers cost-sensitive. In *Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 155–164. ACM SIGKDD: ACM.

Hettich, S., C. L. Blake, and C. J. Merz (1998). UCI repository of ma-

chine learning databases. Downloadable at `http://www.ics.uci.edu/~mlearn/MLRepository.html`.

Hsu, C.-W., C.-C. Chang, and C.-J. Lin (2003). A practical guide to support vector classification. Technical report, National Taiwan University.

Hsu, C.-W. and C.-J. Lin (2002). A comparison of methods for multiclass support vector machines. *IEEE Transactions on Neural Networks 13*(2), 415–425.

Hull, J. J. (1994). A database for handwritten text recognition research. *IEEE Transactions on Pattern Analysis and Machine Intelligence 16*(5), 550–554.

Langford, J. and A. Beygelzimer (2005). Sensitive error correcting output codes. In P. Auer and R. Meir (Eds.), *Learning Theory: 18th Annual Conference on Learning Theory*, Volume 3559 of *Lecture Notes in Artificial Intelligence*, pp. 158–172. Springer-Verlag.

Li, L. and H.-T. Lin (2007). Optimizing 0/1 loss for perceptrons by random coordinate descent. In *Proceedings of the 2007 International Joint Conference on Neural Networks (IJCNN 2007)*, pp. 749–754. IEEE.

Lin, H.-T. (2008). *From Ordinal Ranking to Binary Classification*. Ph. D. thesis, California Institute of Technology.

Lin, H.-T. and L. Li (2006). Large-margin thresholded ensembles for ordinal regression: Theory and practice. In J. L. Balcazár, P. M. Long, and F. Stephan (Eds.), *Algorithmic Learning Theory*, Volume 4264 of *Lecture Notes in Artificial Intelligence*, pp. 319–333. Springer-Verlag.

Lin, H.-T. and L. Li (2008). Support vector machinery for infinite ensemble learning. *Journal of Machine Learning Research 9*, 285–312.

Margineantu, D. D. (2001). *Methods for Cost-Sensitive Learning*. Ph. D. thesis, Oregon State University.

Xia, F., L. Zhou, Y. Yang, and W. Zhang (2007). Ordinal regression as multiclass classification. *International Journal of Intelligent Control and Systems 12*(3), 230–236.

Zadrozny, B., J. Langford, and N. Abe (2003). Cost sensitive learning by cost-proportionate example weighting. In *Proceedings of the 3rd IEEE International Conference on Data Mining (ICDM 2003)*. IEEE Computer Society.