AN ITERATIVE CONSTRAINED OPTIMIZATION APPROACH TO CLASSIFIER DESIGN

Sibel Yaman and Chin-Hui Lee

School of Electrical and Computer Engineering
Georgia Institute of Technology
Atlanta, GA 30332-0250, USA
{syaman, chl}@ece.gatech.edu

ABSTRACT

In this paper, we propose an iterative constrained optimization (ICO) approach to classifier design. When a set of conflicting objectives needs to be simultaneously satisfied, it is often not easy to combine all the utilities in a single overall objective function for optimization. We instead formulate the problem with conflicting objectives as a single-objective optimization scenario while embedding other competing objectives in constraints so that the original problem can be solved by adopting conventional constrained nonlinear optimization techniques. The bounds needed to constrain each objective are determined based on the objective function values obtained in the previous iterate. The so-formed individual constrained optimization problems are solved until a stable solution is obtained. We illustrate the utility of our framework in the context of designing classifiers for text categorization and automatic language identification. The results of our experiments demonstrate that our approach achieves a significant improvement in one objective with only slight degradation of the other conflicting objective.

1. INTRODUCTION

Many problems in a wide variety of engineering disciplines require the need to optimize several nonlinear functions of the design variables simultaneously. Suppose that the system specifications require the minimization of M conflicting objectives, i.e.,

$$\min_{x \in \chi} [f_1(x), ..., f_M(x)]. \quad (1)$$

The common approach is to set a single overall objective function that can combine all the utilities. However, the overall objective function is generally weak in achieving satisfactory levels for all the conflicting objectives. It is very likely that one objective function dominates the optimization process and the others do not meet the goals initially set. In particular, an overall objective function can be constructed with equal importance on each objective, yet the resulting objective function values may turn out to be very different.

There is a recent effort in the machine learning community to develop techniques to approximate empirical error rates are as differentiable functions of the classifier parameters [1]. Such approximations bridge the gap between the classifier's learning objective and its performance since a *single* chosen performance metric can be simulated and then optimized.

A multi-objective approach [2] is better suited for many machine learning applications encountered, which typically require the optimization of several *and* conflicting objectives. This is because multi-objective optimization is concerned with the optimization of all the objectives *simultaneously*.

In this paper, we develop an approach, which we refer to as iterative constrained optimization (ICO), that can simultaneously optimize several conflicting criteria. By iteratively optimizing one objective at a time with constraints on others and adjusting the constraint bounds by using information from previous iterate, ICO framework yields satisfactory achievements for all the objectives simultaneously.

In Section 2, we briefly overview some related concepts in multi-objective optimization that will be used later. In Section 3, we describe our approach. Later, we apply our ICO framework to two areas in machine learning, namely text categorization and automatic language identification. We report the experimental results in Section 4. We finalize the paper by concluding remarks in Section 5.

2. BACKGROUND

In multi-objective optimization problems, there are usually (infinitely) many optimal solutions that constitute the so-called Pareto optimal set.

Definition 2.1 *A decision vector $x^* \in \chi$ is (global) Pareto optimal if there does not exist another decision vector $x \in \chi$ such that $f_i(x) \leq f_i(x^*)$, for all $i = 1, ..., M$ and $f_j(x) < f_j(x^*)$ at least for one index j [2].*

Mathematically, every Pareto optimal solution is an equally acceptable solution to the multi-objective optimization problem. Usually, a decision maker is contained in the problem to select one out of the many Pareto optimal solutions. According to the definition of Pareto optimality, moving from one Pareto optimal solution to the other necessitates trading off. It is said that two feasible solutions are situated on the same indifference curve if a decision maker finds them equally desirable.

In the ϵ-constraint method [2], introduced as a solution to multi-objective optimization problems, one of the objective functions is selected to optimize and all the other objective functions are introduced as constraints by setting up an upper bound for each of them.

3. ITERATIVE CONSTRAINED OPTIMIZATION

We formulate the multi-objective optimization problem as a set of M single-objective optimization problems in the form

$$\begin{aligned} \min_{x\in\chi} \quad & f_\ell(x) \\ \text{subject to} \quad & f_i(x) \le f_i + \delta_i,\ i = 1, ...M, i \neq \ell, \end{aligned} \tag{2}$$

for all $\ell = 1, ..., M$, i.e., the minimization of one objective function with constraints on the other conflicting objectives $f_i(x)$ where $(f_i + \delta_i)$'s are the objective function values to be attained.

3.1. Discriminant Function Approach to Classifier Design

The first step in designing classifiers that optimize an objective function of interest f_ℓ is the approximation of f_ℓ as a function of the classifier parameters ω. The linear discriminant function (LDF) with the category discriminant function $g(x, \omega) = \omega^T x + \omega_0$ is one family of pattern classifiers used to classify the data samples $x \in \Re^n$ to predefined classes. Given the category discriminant function, a category misclassification function $d(x, \omega)$ can be associated to a decision rule: x is classified to the *target* class $\mathcal{C}^+$ if $d(x, \omega) \le 0$; otherwise, it is categorized to the *non-target* class $\mathcal{C}^-$ [1]. A category loss function $\ell(x, \omega) = \frac{1}{1+e^{-(\alpha d(x,\omega)+\beta)}}$ can then emulate the classification operation since true-positives (TP), false-negatives (FN), false-positives (FP) and true-negatives (TN) can be approximated as

$$TP \approx \sum_{x\in\mathcal{C}+} (1-\ell^+), FN \approx \sum_{x\in\mathcal{C}+} \ell^+, FP \approx \sum_{x\in\mathcal{C}-} \ell^-, TN \approx \sum_{x\in\mathcal{C}-} (1-\ell^-) \tag{3}$$

3.2. Single-Objective Optimization

3.2.1. Augmented Lagrangian Function (ALF)

One fundamental approach to the constrained optimization problem in Eq. (2) is to replace the original constrained problem by a sequence of unconstrained subproblems. The approach we take to achieve this is the so-called the augmented Lagrangian method [3, 4]. The problem with inequality constraints in Eq. (2) is first converted to a problem with equality constraints by introducing slack variables s_i by $c_i(x) - s_i = 0, s_i \ge 0$, where $c_i(x) = (f_i + \delta_i) - f_i(x)$ for all $i = 1, ..., M, i \neq \ell$. The slack variables s_i can be eliminated by an explicit minimization with respect to each $s_i \ge 0$ and then can be substituted. This yields the problem of the minimization of the augmented Lagrangian function (ALF)

$$L_A(x, \lambda_i^k, \rho^k) = f_\ell(x) + \sum_{i\in\Im} \psi(x_k, \lambda_i^k, \rho^k)\} \tag{4}$$

$$\psi(x_k, \lambda_i^k, \rho^k) = \begin{cases} -\lambda_i^k c_i(x_k) + \frac{\rho^k}{2} c_i^2(x_k), \text{ if } c_i(x_k) - \frac{\lambda_i^k}{\rho_k} \le 0 \\ -\frac{(\lambda_i^k)^2}{2\rho^k}, \text{ otherwise} \end{cases} \tag{5}$$

The sequences of the penalty parameters $\{\rho_k\}$ and the Lagrangian multipliers $\{\lambda_k\}$ are generated by a general algorithm by Powell (1969) [5] to ensure global convergence.

3.2.2. Unconstrained Optimization of ALF

Replacing the original constrained optimization problem with the unconstrained optimization of the augmented Lagrangian function in Eqs. (4) and (5), we need to decide how to move from one iterate x_k to the next x_{k+1} so that the objective L_A is minimized. The first step is to find a direction d_k satisfying the descent condition $\nabla_x L_A^T d_k < 0$. Then, a step-size α_k in the direction d_k should be chosen so that L_A is lowered as much as possible.

In the quasi-Newton methods [3, 4], the descent direction is found as $d_k = -H_k \nabla_x L_A(x_k)$ where H_k is an approximation to the inverse Hessian $\nabla_{xx}^2 L_A$. For the update of H_k, we use the BFGS (Broyden-Fletcher-Goldfarb-Shanno) algorithm for which there is growing evidence that it is the best quasi-Newton method [3]. The Armijo rule [3, 4] is used so that we find the largest step-size α_k in the direction of d_k such that $L_A(x_k + \alpha d_k, \lambda_k, \rho_k) \le L_A(x_k, \lambda_k, \rho_k) + \alpha_k \nabla_w L_A^T(x_k, \lambda_k, \rho_k) d_k$.

The BFGS algorithm and the Armijo rule yield a sequence of iterates x_k that converges to the optimizer of the problem at a super-linear rate [3, 4].

3.3. Iterative Refinement

Starting with an iterate $x^{(1)}$ and the corresponding objective vector $f^{(1)} = [f_1^{(1)}(x), ..., f_M^{(1)}(x)]$, it is desired to move into another iterate $x^{(2)}$ yielding an objective function vector $f^{(2)}$ where at least one objective function $f_k, 1 \le k \le M$ achieves a considerably improved value, while others are only slightly degraded.

In Figure 2, a feasible objective space Z for two-objective case is illustrated, where z_1 and z_2 are properly scaled. Starting with an initial feasible objective vector z, it is desired to reach a point on the Pareto optimal set. In the simplest case of two conflicting objectives, one iteration of moving from $x^{(1)}$ to $x^{(2)}$ requires solving two subproblems of minimizing f_1 and f_2 with constraints on the other.

3.3.1. Generation of the Constraint Bounds

Suppose that the solution to the single-objective optimization problem in Eq. (2) yielded an objective vector $f^{(j)} = [f_1^{(j)}, ..., f_M^{(j)}]$. We can perturb $f^{(j)}$ only slightly by a $\delta \in \Re^M$ vector and set as the constraint bounds: *Improving one objective at the expense of slight degradation on others, ICO technique refines the solution $x^{(j)}$ at every stage until it becomes impossible to reach a more favorable indifference curve.*

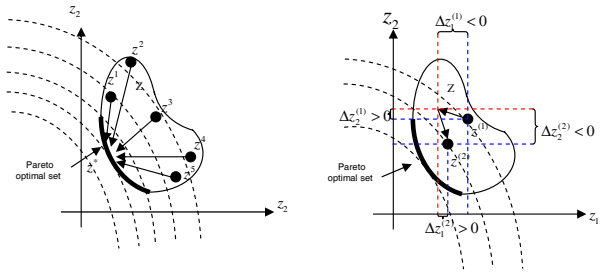

Fig. 1. A point on the Pareto optimal set is sought starting at a feasible objective vector z^1, z^2, z^3, z^4, z^5. Each intermediate step yields an iterate on a more favorable indifference curve. A slight change in the first objective $\Delta z_2^{(1)} > 0$ makes it possible to gain considerable improvement in f_1 by $\Delta z_1^{(1)} < 0$ in the first subproblem.

In our framework, the constraint bounds $\delta = [\delta_1, ..., \delta_M]$ for the next subproblem are set based on the objective vector from the previous subproblem as

$$\delta^{(j+1)} = \Psi^{(j+1)} f^{(j)} \tag{6}$$

where $\Psi = diag(\psi_1, ..., \psi_M)$ is a diagonal matrix such that $0 \leq \psi_i \ll 1, i = 1, ..., M$. Each $\psi_i^{(j+1)}$ is referred to as the *relaxation factor* for the i^{th} objective in the $(j+1)^{st}$ iteration. The higher the ψ_i is, the higher the degradation allowed in the i^{th} objective is. Note that $\psi_i = 0$ means that the i^{th} objective is not allowed at all to be degraded. Although this situation is desirable, remember that one objective can improve at the expense of the conflicting objectives.

We consider two schemes for setting the relaxation factors ψ_i. In the first scheme, we choose the relaxation factors as $(1)\ 0 < \psi_i \ll 1, \forall i = 1, ..., M,\ \ (2)\ \psi_i^{(p)}$ constant, $\forall i = 1, ..., M$.The first condition implies that each objective is *only slightly* degraded while the second condition implies a small constant degradation is allowed in all iterations. In the second scheme, we choose the M relaxation factors by $(1)\ 0 < \psi_i \ll 1, \forall i = 1, ..., M,\ \ (2) \psi_i^{(j)} \to 0$ as $n \to \infty, \forall i = 1, ..., M$. This second condition implies that the degradation allowed is reduced in each iteration. The latter can be satisfied by reducing the relaxation factor as a geometric time series, i.e., $c^j, 0 < c \ll 1$.

3.3.2. ICO Approach to Classifier Design

Through the use of Eq. (3), one can simulate classifier evaluation metrics as differentiable objective functions of the classifier parameters. After the proper selection of a set of M conflicting objectives, one can translate the ICO approach to a classifier learning algorithm. The major advantage is that any objective that is not satisfactory can be improved at the expense of a minor degradation of the conflicting objectives. In case one of the objectives is of greater importance, the constraints can be adjusted accordingly; at the same time, the remaining objectives still attain satisfactory levels since they are subject to some worst-case upper bounds.

	Unconstrained			ICO		
EARN	P	R	F_1	P	R	F_1
MAX PRECISION	0.998	0.543	0.703	0.998	0.788	0.881
MAX RECALL	0.414	1.000	0.586	0.751	0.999	0.857
CORN	P	R	F_1	P	R	F_1
MAX PRECISION	1.000	0.125	0.222	1.000	0.393	0.564
MAX RECALL	0.140	0.982	0.245	0.319	0.964	0.480

Table 1. For the maximization of precision, we let $\psi = 0.005$ for precision and $\psi_2 = 0.200$ where each relaxation factor is reduced in each iteration j as 0.5^j. For the maximization of recall, we let $\psi_1 = 0.200$ and$\psi_2 = 0.01$ and both are reduced as 0.1^j in the j^{th} iteration.

4. EXPERIMENTAL STUDY

4.1. Text Categorization

Text categorization [6] is the process of assigning text documents to a set of predefined categories based on document contents. It is an important component for many applications, including document routing/filtering, web organization and word sense disambiguation. In text categorization, two standard performance evaluation metrics are precision and recall, $\frac{TP}{TP+FP}$ and $\frac{TP}{TP+FN}$, respectively. The common approach is to have a compromise between these two conflicting objectives.

In text categorization, each document is converted to a document vector by counting the term occurrences in the document. This results in a very large term-document matrix of size $10,118$x7770 for the Reuters-21578 text corpus [1]. The LSI-based feature extraction technique [7] can be used so that the dimension of the document vectors is reduced from 10,118 to 1618 while still retaining effective indexing information.

We designed two classifiers for each objective following our classifier design outline in Section 3.3.2. The advantage of our ICO framework to the unconstrained approaches becomes clear when one objective has higher importance than the other. In Table 1, the performance of the unconstrained and constrained approaches are compared for two categories where the objectives have asymmetric importance: `EARN` (a large set) and `CORN` (a smaller set). For the category `EARN`, when precision is maximized, our ICO framework achieves a precision of 0.9942 whereas unconstrained approach achieves 0.9983, which means a loss of about 0.004. However, our ICO framework results in recall 0.788 whereas unconstrained approach results in 0.5428. Hence, *with little or no degradation in one objective, our ICO approach can considerably enhance the other*. Similar arguments are valid for the maximization of recall as well as for the category `CORN` with a lot fewer category training documents.

4.2. Automatic Language Identification (LID)

The technique illustrated for the text categorization problem can be applied to the problem of the classification of multi-

media data. In [8], an acoustic segment modeling technique is described for tokenizing the audio data for the LID task. Upon tokenization, a speech segment or an utterance can be represented as a vector of detected patterns and treated as a *spoken document*. Documents from the same language are then grouped into the same topic category. Similarly in [9], a framework where image data is represented as a vector of features and treated as text is developed. Therefore, by proper tokenization of continuous signals such as audio or image, we can construct corresponding multimedia feature vectors and treat them as document vectors. Doing so, any text feature extraction and classification method can be directly applied to the LID, image annotation or other multimedia classification problems.

LID is the process of determining the language identity corresponding to a given set of spoken queries. It is used in applications such as multilingual speech recognition, spoken language translation and spoken document retrieval. The common evaluation metrics in the LID task are false rejection (FR) and false alarm rates (FA), $\frac{FN}{|C^+|}$ and $\frac{FP}{|C^-|}$, respectively. The usual approach is to have a tradeoff between these two conflicting metrics to reach a balance.

In our experiments, we used the data provided by Bin Ma [8]. The training data set consists of more than 170,000 training samples and the evaluation set consists of 1492 unknown samples. We designed two classifiers for each objective following our design outline in Section 3.3.2.

In Figure 2, the changes of FR, FA and their arithmetic average (C_{det}) for both the training and test data are illustrated separately as a function of iteration number. With an inspection of the change of errors for "Mandarin" (MA), we see that FA is reduced from 31.90% to 7.69% (i.e. by more than 24%) at the expense of an increase in the FR from 1.11% to 3.09% (i.e. by less than 2%) in the training data. Similarly in the test data, FA is reduced from 31.29% to 9.21% (i.e. by more than 22%) at the expense of an increase in FR from 0.00% to 1.28%. It is remarkable that the two conflicting objectives move towards a balance: To improve one objective, we deliberately degrade the other objective a little. Similar results were obtained for the other languages, and further experiments support that the improvement is obtained only with a slight degradation.

5. CONCLUSION

In this paper, we laid the formal treatment of an augmented Lagrangian-based multi-objective optimization approach to classifier design, which we refer to as iterative constrained optimization (ICO). The original optimization problem with conflicting objectives is formulated as an iterative process of single-objective optimization problems in which the remaining objectives are embedded as constraints. We refine the bounds needed to constrain the conflicting objectives based on the *slight perturbation* of the previous subproblem in an iterative manner. We illustrated the utility of our ICO framework developed in the context of classifier design in text categorization and automatic language identification. We observed that ICO approach can achieve a remarkably better performance in one objective with a slight degradation on the other conflicting objective.

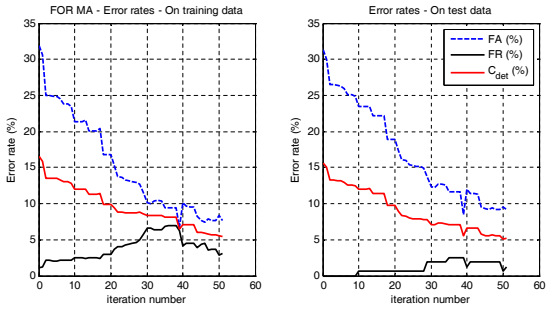

Fig. 2. The change of false alarm (FA), false rejection (FR) and their average C_{det} for the language "Mandarin" (MA). We let the relaxation factors be $\psi_1 = 0.003$ for the false alarm rate and $\psi_2 = 0.006$ for the false rejection rate.

6. ACKNOWLEDGEMENTS

The authors are grateful to Bin Ma for providing the LID data for use in the experiments, and Jinyu Li for his very helpful discussions on pattern recognition problems, and Wen Wu for his help with the text categorization problem and with the Reuters-21578 database.

7. REFERENCES

[1] S. Gao, W. Wu, C.-H. Lee, and T.-S. Chua, "A mfom learning approach to robust multiclass multi-label text categorization," in *21st International Conference on Machine Learning*, 2004.

[2] K. Miettinen, *Nonlinear Multiobjective Optimization*, Springer, 1999.

[3] J. Nocedal and S.J. Wright, *Numerical Optimization*, Springer, 1999.

[4] D.P. Bertsekas, *Constrained Optimization and Lagrange Multiplier Methods*, Athena Scientific, 1996.

[5] R. Fletcher, *Practical Methods of Optimization*, John Wiley & Sons, 2000.

[6] T. Joachims, "A probabilistic analysis of the Rocchio algorithm with TFIDF for text categorization," in *14th International Conference on Machine Learning*, 1997.

[7] J. R. Bellegarda, "Exploiting latent semantic information in statistical language modeling," *In Proc. IEEE*, vol. Vol. 88, pp. pp. 1279–1296, 2000.

[8] B. Ma, H. Li, and C.-H. Lee, "An acoustic segment modeling approach to automatic language identification," in *9th European Conference on Speech Communication and Technology*, 2005, pp. 2837–2840.

[9] S. Gao, D.-H. Wang, and C.-H. Lee, "Automatic image annotation through multi-topic text categorization," *submitted to ICASSP2006*, 2006.