한국경영과학회지
제39권 제4호
2014년 11월
http://dx.doi.org/10.7737/JKORMS.2014.39.4.071

Short Note on Optimizing Feature Selection to Improve Medical Diagnosis

Cui Guo[1] · Hong Seo Ryoo[2]
[1]Business School, Shantou University
[2]Industrial Management Engineering, Korea University

■ Abstract ■

A new classification framework called 'support feature machine' was introduced in [2] for analyzing medical data. Contrary to authors' claim, however, the proposed method is not designed to guarantee minimizing the use of the spatial feature variables. This paper mathematically remedies this drawback and provides comments on models from [2].

Keywords : Support Features, Feature Selection, Integer Programming, Multiobjective Programming

1. Introduction

A new classification framework called support feature machine (SFM) was introduced in [2]. The authors claim that SFM is specialized for analyzing medical data in that it "matches temporal patterns (in time series analysis) using the nearest neighborhood rule, while optimizing the selection of good (spatial) features" and presented two integer programming (IP) classification models and their mixed integer and linear programming (MILP) extensions. Specifically, for classifying two types of n (training) data, let $x_j \in \{0, 1\}$, $j \in \{1, \cdots, m\}$, be a decision variable indicating if feature j is selected and $y_i \in \{0, 1\}$, $i \in \{1, \cdots, n\}$, be a decision variable indicating if data i is correctly classified by a SFM rule to be discovered. Let

논문접수일 : 2014년 09월 01일 논문게재확정일 : 2014년 10월 19일
논문수정일(1차 : 2014년 10월 16일)
[1] This author acknowledges the support of a research grant by Social Science Program of Shantou University (Grant Number : 07404861).
[2] This author acknowledges the support of a research grant by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (Grant Number : NRF-2013R1A1A2011784).
† 교신저자 hsryoo@korea.ac.kr

$$\Theta := \left\{ \begin{array}{c} \boldsymbol{A}\mathbf{x} - \left(\frac{1}{2}\boldsymbol{e}^T\mathbf{x}\right)\boldsymbol{e} \le M\mathbf{y} \\ \left(\frac{1}{2}\boldsymbol{e}^T\mathbf{x}\right)\boldsymbol{e} - \boldsymbol{A}\mathbf{x} + \epsilon\boldsymbol{e} \le M(\boldsymbol{e}-\mathbf{y}) \\ \mathbf{x}\in\{0,1\}^m,\ \mathbf{y}\in\{0,1\}^n \end{array} \right\}$$

where $\boldsymbol{e}$ is a vector of 1's of appropriate dimension, $M = m/2$, $0 < \epsilon < 1/2$, and $\boldsymbol{A} = [a_{ij}]$ for $i \in \{1, \cdots, n\}$ and $j \in \{1, \cdots, m\}$ is an accuracy matrix, where $a_{ij} = 1$ indicates that the nearest neighbor rule correctly classified training sample i at feature dimension j or 0 otherwise. Let

$$\Delta := \left\{ \begin{array}{c} \bar{\boldsymbol{D}}\mathbf{x} - \boldsymbol{D}\mathbf{x} \le M_1\mathbf{y} \\ \boldsymbol{D}\mathbf{x} - \bar{\boldsymbol{D}}\mathbf{x} \le M_2(\boldsymbol{e}-\mathbf{y}) \\ \mathbf{x}\in\{0,1\}^m,\ \mathbf{y}\in\{0,1\}^n \end{array} \right\}$$

where $d_{ij}(\bar{d}_{ij})$ is the average distance between training sample i and all training samples from the same (different) class at feature dimension j and $\boldsymbol{D} = [d_{ij}]$ ($\bar{\boldsymbol{D}} = [\bar{d}_{ij}]$) for $i = 1, \cdots, n$ and $j = 1, \cdots, m$. (We note that, unlike in Θ, M_1 and M_2 values are not specified for Δ in [2].) The two integer programming classification models from [2] are

$$\text{V-SFM}: Z_V = \max_{(\mathbf{x},\mathbf{y})\in\Theta} \{c(\mathbf{x},\mathbf{y}) = \boldsymbol{e}^T\mathbf{y} : \boldsymbol{e}^T\mathbf{x} \ge \mathbf{1}\}$$

and

$$\text{A-SFM}: Z_A = \max_{(\mathbf{x},\mathbf{y})\in\Delta} \{c(\mathbf{x},\mathbf{y}) = \boldsymbol{e}^T\mathbf{y} : \boldsymbol{e}^T\mathbf{x} \ge \mathbf{1}\},$$

and their four MILP extensions in [2] are the ones that are obtained from the two IP models by simply relaxing the 0-1 integrality requirement on feature variables from $\mathbf{x}\in\{0,1\}^m$ to $\mathbf{x}\in[0,1]^m$.

Support features are a minimal subset of all features that can describe the data under analysis without contradiction. Note in the formulations of V-SFM and A-SFM that they are designed for maximizing the number of correctly classified data only by the nearest neighborhood rule, and there is no 'mechanism' in these models that deals with the selection of support features other than the 0-1 integrality on feature variables $\mathbf{x}$. To be more specific, since an optimal solution to any IP is not likely to have all its 0-1 variables to take value 1, simply requiring integrality on $\mathbf{x} \in \{0, 1\}^m$, one obtains a naive and passive way of identifying a subset of the spatial features, and this is exactly what V-SFM and A-SFM are designed to do.

In summary, we note that the classification framework as proposed in [2] cannot minimize the number of (spatial) features in maximizing the number of correctly classified data, and, hence it is difficult to categorize it as a 'support feature machine' that is specialized for effectively analyzing medical data with both temporal and spatial characteristics.

In this paper, we present a way to mathematically remedy the drawback of the two IP classification models above to transform them into valid support feature machines that minimize the number of features in maximizing the number of correctly classified data. We also provide three comments in regard to classification models from [2].

2. Remedying V-SFM and A-SFM for Selecting Support Spatial Features

In order for the two SFM models of the previous section to select a minimal number of spatial features, the objective functions in V-SFM

and A-SFM need to be modified to include a function (or extra terms) involving feature variables $\mathbf{x}$.

Taking V-SFM for example, let us consider the following IP model:

$$\text{V}'\text{-SFM}: \ z^{V'} = \max_{(\mathbf{x},\mathbf{y})\in\Theta} \{c'(\mathbf{x},\mathbf{y}) = \boldsymbol{e}^{\boldsymbol{T}}\mathbf{y} - \omega\boldsymbol{e}^{\boldsymbol{T}}\mathbf{x} : \boldsymbol{e}^{\boldsymbol{T}}\mathbf{x} \geq \mathbf{1}\}$$

As seen, the multiobjective function of V′-SFM maximizes the number of correctly classified data via the maximization of $\boldsymbol{e}^{\boldsymbol{T}}\mathbf{y}$ and, at the same time, minimizes the number of features to use via the minimization of $\boldsymbol{e}^{\boldsymbol{T}}\mathbf{x}$. As for the coefficient ω, one can choose any real value from the interval

$$\omega \in \left(0, \frac{1}{m}\right).$$

For example, one can simply choose $\omega = \frac{1}{m+1}$. Then, an optimal solution of V′-SFM can be shown to correctly classify the same number of data as an optimal solution of V-SFM while using a minimal subset of the spatial features. This is shown by the following lemma.

Lemma 1. *Consider V-SFM and V′-SFM formulated on the same set of (training) data. Let* $(\boldsymbol{x}^*, \boldsymbol{y}^*)$ *and* $(\boldsymbol{x}', \boldsymbol{y}')$ *denote an optimal solution of V-SFM and V′-SFM, respectively. Then, we have*

$$\boldsymbol{e}^T\boldsymbol{y}^* = \boldsymbol{e}^T\boldsymbol{y}'.$$

Proof: First, note that the two IP models have the same set of constraints, hence, by the nature of optimization principles, we have

$$z_V = \boldsymbol{e}^T\boldsymbol{y}^* \geq \boldsymbol{e}^T\boldsymbol{y}' \qquad (1)$$

Next, we note that $(\mathbf{x}^*, \mathbf{y}^*)$ is a feasible solution to V′-SFM, and this naturally yields

$$\boldsymbol{e}^T\mathbf{y}' - \omega\boldsymbol{e}^T\mathbf{x}' \geq \boldsymbol{e}^T\boldsymbol{y}^* - \omega\boldsymbol{e}^T\mathbf{x}^*.$$

Since $\omega \in \left(0, \frac{1}{m}\right)$, we have

$$0 < \omega\boldsymbol{e}^T\mathbf{x}' < 1 \text{ and } 0 < \omega\boldsymbol{e}^T\mathbf{x}^* < 1,$$

which, along with $\boldsymbol{y} \in \{0, 1\}^n$, yields

$$\boldsymbol{e}^T\mathbf{y}^* \leq \boldsymbol{e}^T\mathbf{y}'.$$

Combining the above with (1), we obtain the desired result. □

In brief, this lemma shows that V′-SFM matches temporal patterns by using the nearest neighborhood rule while using a minimal subset of all spatial features. Likewise, one can let

$$\text{A}'\text{-SFM}: \ Z^{A'} = \max_{(\mathbf{x},\mathbf{y})\in\Delta} \{c'(\mathbf{x},\mathbf{y}) = \boldsymbol{e}^{\boldsymbol{T}}\mathbf{y} - \omega\boldsymbol{e}^{\boldsymbol{T}}\mathbf{x} : \boldsymbol{e}^{\boldsymbol{T}}\mathbf{x} \geq \mathbf{1}\}$$

and use $\omega \in \left(0, \frac{1}{m}\right)$ to obtain a rectified model of A-SFM for selecting support spatial features.

3. Additional Remarks

We close this short note with three remarks on classification models from [2].

Remark 1: No specific values are given in [2] for M_1 and M_2 in the definition of Δ for A-SFM. We determine these values in this remark.

One immediately notes that $y_i = 1$ when $\Sigma_{j=1}^m \bar{d}_{ij}x_j - \Sigma_{j=1}^m d_{ij}x_j > 0$ only if M_1 takes a value that is greater than or equal to the maximum value of $\Sigma_{j=1}^m \bar{d}_{ij}x_j - \Sigma_{j=1}^m d_{ij}x_j$ for $i \in \{1, \cdots, n\}$.

Thus, one can use $M_1 \coloneqq \max_{i \in \{1, \cdots, n\}} \Sigma_{j=1}^{m} \bar{d}_{ij}$ and $M_2 \coloneqq \max_{i \in \{1, \cdots, n\}} \Sigma_{j=1}^{m} d_{ij}$ in A-SFM.

Remark 2: A-SFM model counts all unclassified data i with $\Sigma_{j=1}^{m} \bar{d}_{ij} x_j = \Sigma_{j=1}^{m} d_{ij} x_j$ toward correct classifications as the two constraints for data i in Δ reduce to $y_i \geq 0$ and $1 - y_i \geq 0$, hence $0 \leq y_i \leq 1$. Now, since the objective function $\Sigma_{i=1}^{n} y_i$ is to be maximized, A-SFM sets all such $y_i = 1$ in an optimal solution.

This is a mathematical 'deficiency' of A-SFM but can prove beneficial from a classification perspective. As seen, this can inflate the training performance of A-SFM classifiers (refer <Tables 4> and <Table 5> in [2]). Furthermore, this can allow A-SFM classifiers to be less affected by support vectors (the harder-to-classify data) and, thus, can help them perform better in testing (refer <Table 6> in [2]).

Likewise, $\epsilon \boldsymbol{e}$ terms in the second set of constraints in Θ, which were introduced for a strict separation of different types of data, can be dropped from the formulation of V-SFM for the same effect.

Remark 3: The four MILP models in [2] without the integrality on $\mathbf{x}$ are regular nearest neighborhood-based classification models that are hardly seen as support feature machines. In fact, with all due respect to the authors, we suspect it might have been the case that these MILP classification models were developed before V-SFM and A-SFM.

In summary, one can take a random classification model (for example, a popular linear programming-based model from [1, 3]) and require 0-1 integrality on the feature variables. This will force the modified classification model to implement a decision rule on a subset of the features, as it is not likely that $x_j = 1$ for all $j \in \{1, \cdots, m\}$ in its optimal solution. However, such a model would hardly qualify as a support feature machine designed specifically for medical data.

References

[1] Bennett, K. and O. Mangasarian, "Robust linear programming discrimination of two linearly inseparable sets," *Optimization Methods and Software*, Vol.1(1992), pp.23-34.

[2] Fan, Y.J. and A. Chaovalitwongse, "Optimizing feature selection to improve medical diagnosis," *Annals of Operations Research*, Vol.174 (2010), pp.169-183.

[3] Mangasarian, O., "Multisurface method of pattern separation," *IEEE Transactions on Information Theory*, Vol.14, No.6(1968), pp. 801-807.