IEICE TRANS. INF. & SYST., VOL.E99–D, NO.5 MAY 2016

LETTER

Accelerating Multi-Label Feature Selection Based on Low-Rank Approximation

Hyunki LIM[†], **Jaesung LEE**[†], ***Nonmembers***, ***and*** **Dae-Won KIM**[†a)], ***Member***

SUMMARY We propose a multi-label feature selection method that considers feature dependencies. The proposed method circumvents the prohibitive computations by using a low-rank approximation method. The empirical results acquired by applying the proposed method to several multi-label datasets demonstrate that its performance is comparable to those of recent multi-label feature selection methods and that it reduces the computation time.

key words: *multi-label feature selection, multivariate feature selection, feature dependency, Nyström method*

1. Introudction

Recently, with the advancement of multi-label data analysis related to modern applications that involve multiple concepts [1], knowledge-mining research has provided information that is vital to achieve the distinct objectives of these applications. Such applications include conventional text categorization [2], image annotation, sentiment analysis for brand and social network service such as Twitter [3].

Large numbers of features degrades the speeds of machine learning algorithms, the generality of knowledge, and the interpretability of the explored models [4]. Multi-label feature selection is considered a solution that can effectively avoid the aforementioned problems [5], [6]. Conventional multi-label feature selection methods evaluate the importance of each feature independently; therefore, the dependencies among features are ignored [2]. As a result, a compact multi-label feature subset cannot be obtained because a selected feature subset will necessarily contain redundant features, that is, features that are similar to one another [6]. To resolve this practical problem, a multi-label feature selection method must consider the feature dependencies during its feature selection process. However, these methods typically require additional computation to evaluate the feature dependencies.

Recently, multi-label quadratic programming feature selection (MLQPFS) was introduced by Lim et al. [7]. It has the advantage that it concurrently considers the dependencies between the features and labels and among the features by using a quadratic function without a special search algorithm. However, although this method has this advantage, it still requires additional computational time $O(N^2)$ (N is the number of features) to determine the feature dependencies.

In this paper, we propose a fast multi-label feature selection method that considers the feature dependencies. To develop this method, we extended the MLQPFS method and endeavored to reduce the computational requirements involved in determining the feature dependencies by using a low-rank approximation. We decreased the time required for feature dependency determination from $O(N^2)$ to $O(Nk)$ (k is the selected number from N features and is much smaller than N) by using the MLQPFS method.

2. Proposed Method

Let $W \subset \mathbb{R}^N$ denote an input space constructed from a set of features F, where $|F| = N$, and let the patterns obtained from W be assigned to a certain label subset $\lambda \subseteq Y$, where $Y = \{y_1, \cdots, y_M\}$ is a finite set of labels with $|Y| = M$. The feature selection problem involves obtaining the subset S composed of n features from $F (n \ll N)$ that is the most dependent upon multiple labels Y.

We formulated an objective function that simultaneously considers the dependencies among features and between the features and labels [7]. In this section, the MLQPFS objective function is introduced. The n features with the highest weight values can be determined by minimizing the objective function of an N-dimensional vector. Similar features should not be included in S because the number of features selected is limited to n. Thus, the dependencies among the selected features in S should be minimized, whereas the dependency between S and Y should be maximized. This concept can be naturally represented by a quadratic objective function. The objective function of $x \in \mathbb{R}^N$ can be written as

$$f(x) = \frac{1}{2} x^T Q x - c^T x, \tag{1}$$

subject to $x_1, \ldots, x_N \geq 0$. In this study, $Q \in \mathbb{R}^{N \times N}$ was computed as

$$Q_{ij} = I(f_i; f_j), \tag{2}$$

where Q_{ij} represents the dependency between f_i and f_j and $I(f_i; f_j)$ means mutual information. Mutual information can be calculated as

$$I(f_i; f_j) = H(f_i) + H(f_j) - H(f_i, f_j) \tag{3}$$

where $H(T) = -\sum_{t \in T} P(t) \log P(t)$ is the joint entropy of T.

Manuscript received November 27, 2015.
Manuscript revised January 18, 2016.
Manuscript publicized February 12, 2016.

[†]The authors are with School of Computer Science and Engineering, Chung-Ang University, Seoul 156–756, Korea.
a) E-mail: dwkim@cau.ac.kr

DOI: 10.1587/transinf.2015EDL8243

Element c_i of a non-negative vector $c \in \mathbb{R}^N$ in Eq. (1) represents the dependency between feature f_i and the multiple labels in the set Y and can be computed as

$$c_i = \sum_{y_j \in Y} I(f_i, y_j). \tag{4}$$

Calculating all of the feature dependencies for the matrix Q using Eq. (2) may require significant time because the number of features is large. To avoid excessive computation time, our strategy involves calculating some of the elements of matrix Q by using Eq. (2) and approximating the rest of elements by employing low-rank approximation using the previously calculated elements.

The matrix Q can be represented as the block matrix shown in Eq. (5) to separate the elements being calculated and approximated.

$$Q = \begin{pmatrix} A & B \\ B^T & E \end{pmatrix}, \tag{5}$$

where $A \in \mathbb{R}^{k \times k}$, $B \in \mathbb{R}^{k \times (N-k)}$, and $E \in \mathbb{R}^{(N-k) \times (N-k)}$ (Note that the matrix Q is symmetric). $A = Q_{1:k,1:k}$ and $B = Q_{1:k,k+1:N}$. Suppose we only know $[AB]$ of matrix Q. In other words, the elements of $[AB]$ are exactly calculated by Eq. (2) and the elements of E (unknown part) will be approximated by the low-rank approximation.

We used the Nyström method to approximate E of Q. The Nyström method is one of the most widely used low-rank approximation methods [8], and it was employed to solve single-label feature selection problem [9]. By applying the Nyström method, the approximated block matrices $\hat{E}$ and $\hat{Q}$ can be represented as

$$E \approx \hat{E} = B^T A^+ B \tag{6}$$

and

$$\hat{Q} = \begin{pmatrix} A & B \\ B^T & B^T A^+ B \end{pmatrix}, \tag{7}$$

where A^+ indicates the pseudo-inverse of A. An approximated $\hat{Q}$ can be obtained only through $O(Nk)$ in Eq. (7) instead of $O(N^2)$. We defined $\frac{k}{N}$ (p) as the selection ratio. If p is 1, the approximation of Q is not calculated in this method.

Because the Nyström method approximates the elements of E by using the elements of $[AB]$, the selection of $[AB]$ from Q is important. Features high selection probabilities should be calculated exactly, while features low selection probabilities need not. Because features with large c_i and small Q_{ij} are considered to be selected features with high selection probabilities, the Q_{ij} of features with large c_i should be calculated preferentially. Therefore, we selected the features of $[AB]$ by ordering the values obtained Eq. (4), that is, the dependencies of the features on labels. $[AB]$ features with larger label dependencies selected. We defined this selection method as the label selection method. In the proposed method, Q is approximated according to Eq. (7) and then the feature weight vector x can be obtained after solving QP. The proposed method is summarized in Algorithm 1.

Algorithm 1 Procedures of the proposed method

procedure PROPOSED METHOD
input $\{f_i\}_{i=1}^N, \{y_i\}_{i=1}^M, p, r$ ▷ n is number of selected features
$k :=$ round$(N \times p)$ ▷ p is selection ratio
initialize $Q \in \mathbb{R}^{N \times N}$, $c \in \mathbb{R}^N$
  for $i = 1 \rightarrow N$ **do**
    compute c_i by $\sum_{y_j \in Y} I(f_i, y_j)$
  end for
sort c in descending order of c
sort f in descending order of c
  for $i = 1 \rightarrow k$ **do**
    for $j = i \rightarrow N$ **do**
      compute Q_{ij} by $I(f_i, f_j)$ ▷ $O(Nk)$
    end for
  end for
$Q := Q + Q^T$
$A := Q_{1:k,1:k}$
$B := Q_{1:k,k+1:N}$
$Q_{k+1:N,k+1:N} := B^T A^+ B$ ▷ Nyström method
solve min $\{\frac{1}{2} x^T Q x - c^T x\}$
sort f according to x in descending order
output top n features in f
end procedure

Table 1 Data sets used in the experiments.

Datasets	Patterns	Features	Labels	Domain
Corel5k	5,000	499	374	Images
Delicious	16,105	500	983	Text
Medical	978	1,449	45	Text
Scene	2,407	294	6	Images
Yeast	2,417	103	14	Biology

3. Experimental Results

To analyze the applicability of the proposed method, we present a comparison of the results obtained by applying the feautre selection methods to real-world data sets. Table 1 lists the data sets used in our experiments; these sets have been widely used for comparative purposes in multi-label classification [10]. To evaluate the performances of the feature selection methods, we compared their execution times and classification accuracies. The feature subsets selected by each multi-label feature selection method were evaluated by using a multi-label naive Bayes (MLNB) classifier [5]. The performance was assessed by using multi-label accuracy and Hamming loss [1]. High multi-label accuracy and low Hamming loss indicate good multi-label classification performance respectively. We evaluated the performance of the method using a 20% holdout set. The experiments were repeated 30 times, and the average value was used to represent the classification performance.

We compared the execution time of the ELA+CHI [2], MLQPFS [7] and pairwise multi-label utility (PMU) [6] method with that of newly proposed method. The selection ratio for the feature dependency approximation of the proposed method is set by 5%. Table 2 shows the execution times (in seconds) and classification performance ($n = 10$)

Table 2 Comparison of multi-label feature selection methods.

Datasets	Methods	Time	M. Acc.	H. Loss
Corel5k	ELA+CHI	54.0	0.006	0.010
	PMU	65,423.5	0.000	0.009
	MLQPFS	314.5	0.006	0.010
	Proposed	202.1	0.006	0.010
Delicious	ELA+CHI	20,566.0	0.025	0.020
	PMU	$\geq$ 3 days	-	-
	MLQPFS	1,828.1	0.065	0.021
	Proposed	1,478.6	0.060	0.021
Medical	ELA+CHI	3.9	0.371	0.020
	PMU	2,877.2	0.487	0.018
	MLQPFS	383.6	0.463	0.022
	Proposed	87.1	0.511	0.018
Scene	ELA+CHI	1.2	0.264	0.271
	PMU	71.0	0.411	0.148
	MLQPFS	19.3	0.335	0.201
	Proposed	2.8	0.341	0.221
Yeast	ELA+CHI	3.2	0.469	0.227
	PMU	49.6	0.457	0.226
	MLQPFS	2.6	0.469	0.226
	Proposed	0.8	0.459	0.232

for the proposed and conventional methods. The time consumed by the proposed method is reasonable in comparison with conventional methods. In particular the proposed method is always faster than the PMU and MLQPFS that consider feature dependency. For the Delicious data set, we could not obtain the execution time of the PMU because it consumed more than 3 days.

Figure 1 shows the classification accuracies of the proposed and conventional methods. The vertical axis represents the multi-label accuracy, and the horizontal axis represents the number of selected features. In all data sets, the performance of the proposed method and MLQPFS is similar. The results of Corel5k and Delicious data sets show that the proposed method and MLQPFS are the best performance regardless of the number of selected features. In the Medical data set, when the number of features is smaller than 15, ELA+CHI shows the worst performance. When the number of features is larger than 15, almost all methods shows similar performance. In the Scene data set, the PMU showed the better performance than other methods. The PMU, which calculates higher-order mutual information, is appropriate for the Scene data set because it contains dense features and labels. However, we see that the PMU is the slowest in the Scene data set, while the proposed method and MLQPFS show similar performance. In all experiments, the proposed method is statistically the same as the MLQPFS method according to the paired t-test. From Table 2 and Fig. 1, it can be concluded that consumed time of the proposed method is reasonable in comparison with the conventional methods and its accuracy is similar to that of the MLQPFS method.

In previous Nyström method studies, various selection methods (so-called sampling methods) were introduced for the Nyström method. Different selection methods have different sensitivities to approximation error. The most widely used are the random and diagonal techniques [8]. To analyze the effects of the selection method used for the Nyström method and to demonstrate the superiority of the proposed

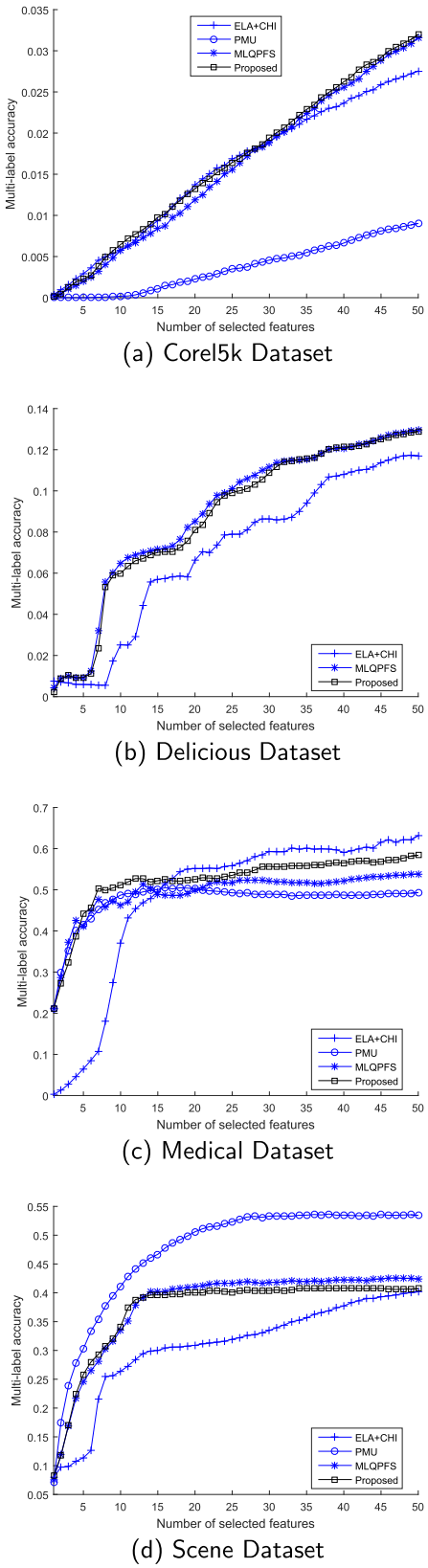

Fig. 1 Comparison of multi-label accuracies of proposed and conventional methods.

Table 3 Comparison of classification accuracies when $n = 10$.

Datasets	Sampling method	Selection ratio (%)			
		5	10	15	20
Corel5k	Random	0.0057	0.0057	0.0058	0.0056
	Diagonal	0.0057	0.0057	0.0056	0.0056
	Label	**0.0068**	**0.0066**	**0.0067**	**0.0067**
Medical	Random	0.4419	0.4400	0.4441	0.4470
	Diagonal	0.4470	0.4470	0.4484	0.4477
	Label	**0.4967**	**0.4868**	**0.4888**	**0.4870**

Table 4 Comparison of classification accuracies when $n = 20$.

Datasets	Sampling method	Selection ratio (%)			
		5	10	15	20
Corel5k	Random	0.0122	0.0118	0.0119	0.0119
	Diagonal	0.0122	0.0120	0.0122	0.0121
	Label	**0.0138**	**0.0143**	**0.0145**	**0.0144**
Medical	Random	0.4837	0.4853	0.4819	0.4921
	Diagonal	0.4962	0.4982	0.5011	0.5008
	Label	**0.5289**	**0.5270**	**0.5280**	**0.5361**

method, we compared the feature dependency approximation errors of the selection methods. Tables 3 and 4 present the more detailed results of the approximation methods when the number of selected feature was 10 and 20, respectively. The bold text indicates the best performance. The classification accuracies for each of the selection methods are shown for selection ratios of 5%, 10%, 15%, and 20%. For the Corel5k and Medical data sets, the proposed method (denoted as "Label" in Tables 3 and 4) shows performance better than those of the conventional methods in all of the experimental settings. From Tables 3 and 4 we can conclude that the approximation error affects the feature selection results and that when the error is low, better feature selection performance is expected.

4. Conclusion

In this paper, we presented a low-rank approximation-based multi-label feature selection method. To efficiently assess the dependencies of the input features in multivariate situations, our proposed method uses the Nyström method to calculate the feature dependencies. The results of the comparisons conducted on three real-world data sets from different domains indicate that our proposed method consumes less time than conventional methods and that it maintains accuracy. Because the proposed method accelerates the multi-label feature selection process, it can be applied to several modern applications that incur explosive features.

However, our proposed method still requires further investigation, as the c time is still excessive and no theoretical analysis has been performed. Future studies should also include label dependency approximations. If we can approximate the label dependency for multi-label applications, then the proposed method may be more useful for multi-label feature selection. In future work, we will study these issues further.

Acknowledgements

This research is supported by Ministry of Culture, Sports and Tourism (MCST) and Korea Creative Content Agency (KOCCA) in the Culture Technology (CT) Research & Development Program 2015.

References

[1] M.-L. Zhang and Z.-H. Zhou, "A review on multi-label learning algorithm," IEEE Trans. Knowl. Data Eng., vol.26, no.8, pp.1819–1837, 2014.

[2] W. Chen, J. Yan, B. Zhang, Z. Chen, and Q. Yang, "Document transformation for multi-label feature selection in text categorization," Proc. 7th IEEE Int. Conf. Data Mining, pp.451–456, Omaha, USA, Oct. 2007.

[3] Y. Rao, Q. Li, X. Mao, and L. Wenyin, "Sentiment topic models for social emotion mining," Information Sciences, vol.266, pp.90–100, 2014.

[4] Y. Zhang and Z.-H. Zhou, "Multilabel dimensionality reduction via dependence maximization," ACM Trans. Knowledge Discovery from Data (TKDD), vol.4, no.3, Article No.14, 2010.

[5] M.-L. Zhang, J.M. Peña, and V. Robles, "Feature selection for multi-label naive Bayes classification," Inf. Sci., vol.179, no.19, pp.3218–3229, 2009.

[6] J. Lee and D.-W. Kim, "Feature selection for multi-label classification using multivariate mutual information," Pattern Recognit. Lett., vol.34, no.3, pp.349–357, 2013.

[7] H. Lim, J. Lee, and D.-W. Kim, "Multi-label learning using mathematical programming," IEICE Trans. Inf. & Syst., vol.E98-D, no.1, pp.197–200, Jan. 2015.

[8] S. Kumar, M. Mohri, and A. Talwalkar, "Sampling techniques for the Nyström method," Int. Conf. Artificial Intelligence and Statistics, pp.304–311, 2009.

[9] I. Rodriguez-Lujan, R. Huerta, C. Elkan, and C.S. Cruz, "Quadratic programming feature selection," J. Machine Learning Research, vol.11, pp.1491–1516, 2010

[10] J. Read, B. Pfahringer, G. Holmes, and E. Frank, "Classifier chains for multi-label classification," Machine Learning, vol.85, no.3, pp.333–359, 2011.