Approximating mutual information for multi-label feature selection

J. Lee, H. Lim and D.-W. Kim

Proposed is a new multi-label feature selection method that captures relationships between features and labels without transforming the problem into single-label classification. Using approximated joint mutual information, the proposed incremental feature selection algorithm provides markedly better classification performance than well-known conventional methods.

Introduction: Multi-label classification, in which an object is associated with more than two labels simultaneously, has emerged as a challenging problem in modern applications such as image and video annotation, semantic scene classification, and music emotion recognition [1–4]. Owing to the high dimensionality of multi-label data, the performance of multi-label classification is strongly influenced by the quality of input features. However, few studies have examined the advantages of feature selection methods. Trohidis *et al.* [4] and Docquire and Verleysen [5] proposed preprocessing steps that transform a multi-label problem into one or more single-label problems, and they adopted conventional feature selection for single-label classification. However, such transformation-based methods cannot capture relationships between features and labels owing to the loss of label interaction. Therefore, in this Letter, we propose a multi-label feature selection technique that considers label dependency to evaluate the quality of the given features, without resorting to problem transformation. The computation cost of feature-label dependency increases exponentially with the dimensions of features and labels. Computationally intractable exact calculations are approximated by joint mutual information. To the best our knowledge, this is the first mathematical study to investigate multi-label feature selection on the basis of information theory.

Proposed method: Given input data with N features, $F = \{f_1, \ldots, f_N\}$, and the label set $Y = \{y_1, \ldots, y_m\}$, the objective of multi-label feature selection is to find a feature subset $S \subset F$ with $n < N$ features such that it has the highest dependency on the label set Y. First, let us describe the method for selecting the top-ranked feature f_i from F such that it maximises the mutual information between f_i and Y, $I(f_i; Y) = H(f_i) + H(Y) - H(f_i, Y)$, where $H(x) = \sum_x p(x) \log p(x)$ is the entropy of x. When considering two labels with $Y = \{y_1, y_2\}$, $I(f_i; Y)$ is written as

$$I(f_i; y_1, y_2) = H(f_i) + H(y_1, y_2) - H(f_i, y_1, y_2) \quad (1)$$

Using Shearer's inequality, $H(f_i, y_1, y_2) \leq \frac{1}{2}(H(f_i, y_1) + H(f_i, y_2) + H(y_1, y_2))$ [6], the lower bound of $I(f_i; y_1, y_2)$ can be obtained as

$$I(f_i; y_1, y_2) \geq \frac{1}{2}(I(f_i; y_1) + I(f_i; y_2) - I(y_1; y_2)) \quad (2)$$

From (2), the lower bound of $I(f_i; y_1, y_2)$ increases with the dependency between f_i and each label. Using (2), the dependency between f_i and Y is approximated as

$$\tilde{I}(f_i; Y) = \sum_{y \in Y} I(f_i; y) - \sum_{y, y' \in Y} I(y; y') \quad (3)$$

To deal with the next feature f_{i+1}, (3) can be expanded as

$$\tilde{I}(f_i, f_{i+1}; Y) = \sum_{y \in Y} I(f_i, f_{i+1}; y) - \sum_{y, y' \in Y} I(y; y') \quad (4)$$

After f_i with the highest $\tilde{I}(f_i; Y)$ is selected as the best feature, the feature f_{i+1} selected in the next step should maximise the information gain when it is included, $I(f_{i+1}; y|f_i) = I(f_i, f_{i+1}; Y) - I(f_i; Y)$. Using the approximation in (3), (4) for these terms, we obtain the following estimate denoted by J:

$$J = \sum_{y \in Y} I(f_i, f_{i+1}; y) - \sum_{y \in Y} I(f_i; y) \quad (5)$$

As in the case of (2) and (3), using Shearer's inequality, $I(f_i, f_{i+1}; y)$ in (5) is estimated as

$$I(f_i, f_{i+1}; y) = I(y; f_i, f_{i+1}) \simeq I(f_i; y) + I(f_{i+1}; y) - I(f_i; f_{i+1}) \quad (6)$$

Therefore, using (5) and (6), we can approximate J as

$$\tilde{J} = \sum_{y \in Y} (I(f_{i+1}; y) - I(f_i; f_{i+1})) \quad (7)$$

The feature f_{i+1} that maximises $\tilde{J}$ is selected as subsequent features. The incremental search algorithm is used to find the near-optimal features defined by $\tilde{J}$. Suppose we have already selected a feature subset S_i; then, the incremental algorithm for selecting the feature f_{i+1} in the next step optimises the following equation:

$$\max_{f_{i+1} \in F - S_i} \left[\sum_{y \in Y} I(f_{i+1}; y) - \sum_{f \in S_i} I(f_{i+1}; f) \right] \quad (8)$$

The computational complexity of this incremental search method is $O(|S| \times N)$.

Experiments and results: We compared the performance of the proposed method with that of conventional multi-label feature selection methods (ELA+CHI, LP+CHI, and PPT+CHI [4, 5]); ELA, LP, and PPT are problem transformation methods, and CHI denotes χ^2 statistics. The classification performance of each method was evaluated using a multi-label naive Bayes (MLNB) classifier [7]. Table 1 lists the data sets [8] employed in the experiments; they have been widely used for comparative purposes in multi-label classification. The performance was assessed using three measures: Hamming loss, ranking loss, and multi-label accuracy [7]. Low values of Hamming loss and ranking loss, and high values of multi-label accuracy, indicate good classification performance.

Table 1: Data sets used in experiments

Data sets	Patterns	Features	Labels
Bibtex	7395	1836	159
Enron	1702	1101	53
Scene	2407	294	6
Yeast	2417	103	14

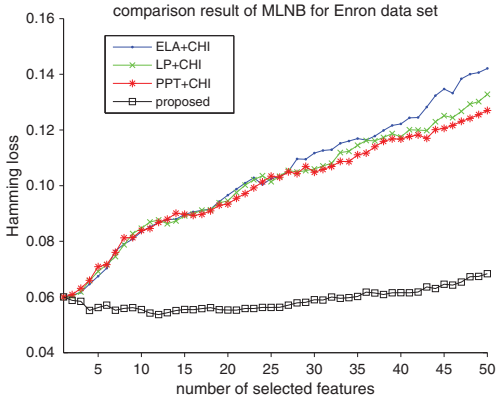

Fig. 1 *Performance comparison of proposed and conventional feature selection methods for Enron data set*

Fig. 1 shows the classification performance of each feature selection method for the Enron data set, which contains 1101 features and 53 labels. The horizontal axis represents the size of the selected feature subset, and the vertical axis represents the Hamming loss. It is evident that the classification performance of the proposed method is superior to that of the conventional methods for any size of the feature subset. The proposed method shows consistently low values of Hamming loss as the size of the selected feature subset increases, whereas the Hamming loss of conventional methods increases linearly as the number of features increases. Fig. 2 shows the classification performance of each feature selection method for the Scene data set. The Hamming loss of the proposed method decreases with increasing size of the feature subset. However, the Hamming loss of the conventional methods rapidly increases as the size of feature subset increases from 1 to 15. Table 2 lists the multi-label accuracy and ranking loss for each method using the four data sets. The best value for each data set is marked in bold text. The proposed method provided a significant improvement in multi-label accuracy and ranking loss. Thus, the proposed method is superior to the conventional methods for all data