A Ranking-based KNN Approach for Multilabel Classification

Tsung-Hsien Chiang, Hung-Yi Lo and Shou-De Lin
r97041@csie.ntu.edu.tw, hungyi@iis.sinica.edu.tw, sdlin@csie.ntu.edu.tw
Graduate Institute of Computer Science and Information Engineering
National Taiwan University
Taipei, Taiwan

***Abstract*—Multi-label classification has attracted a great deal of attention in recent years. This paper presents an approach exploits a ranking model to learn which neighbor's labels are more trustable candidates for a weighted KNN-based strategy, and then assigns higher weights to those candidates when making weighted-voting decisions. Our experiment results demonstrate that the proposed method outperforms state-of-the-art instance-based learning approaches.**

***Keywords*-multilabel classification; nearest neighbor classification; ranking; optimization; generalized pattern search**

I. Introduction

Multi-label classification problems exist in several domains. In this paper, we propose a KNN-based learning algorithm for multi-label classification. Our objective is to exploit the dependency among labels by incorporating a ranking model into the selection process of trustable neighbors. The experiment results show that the approach outperforms state-of-the-art instance-based methods.

II. Related Work

Existing multi-label classification algorithms can be divided into two categories: Problem Transformation and Algorithm Adaptation [1]. Problem Transformation decomposes a multi-label classification task into one or more single-label classification tasks. Therefore, existing single-label classification algorithms can be applied to problems directly. Binary Relevance (BR) is a popular problem transformation method. It transforms the multi-label classification task into several single-label binary classification tasks, each for one of the labels.

Algorithm Adaptation modifies specific algorithms to handle multi-label data directly. Researchers have tried to extend the KNN concept to propose some algorithms such as Mulit-Label K-Nearest Neighbors algorithm (MLKNN) [2] and Instance-Based Logistic Regression (IBLR) [3]. Both IBLR and MLKNN are considered state-of-the-art multi-label classification algorithms that exploit instance-based learning [2], [3].

III. Methodology

Let X denote the domain of an instance and let $L = \{\lambda_1, ..., \lambda_m\}$ denote the set of labels. Assume we are given a multi-label training data set $T = \{(x_1, Y(x_1)), ..., (x_n, Y(x_n))\}$, whose instances are drawn identically and independently from an unknown distribution D. Each instance $x \in X$ is associated with a label set $Y(x) \subseteq L$. The goal of multi-label classification is to train a classifier $h : X \rightarrow 2^L$ that maps a feature vector to a set of labels, while optimizing some specific evaluation metrics.

Our approach determines the final label set of a test instance x, as shown in Figure 1. We identify the k-nearest neighbors of x. Then the selected neighbors are re-ranked by a ranking model trained on their trustiness (i.e., how close their label sets are to the true label set). After deriving the re-ranked neighbors, we use a weighted voting strategy to produce the final prediction. The ranking model training

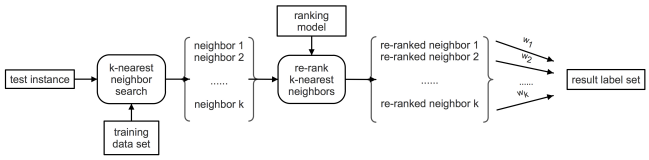

Figure 1. An overview of the testing process

process is shown in Figure 2. First, for each instance x_i, we

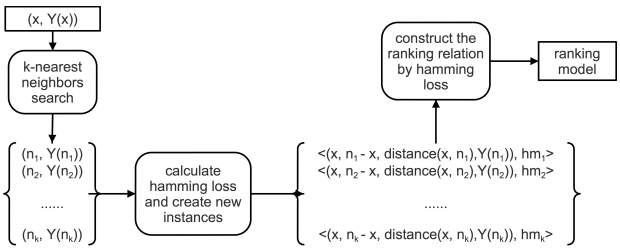

Figure 2. The training process building the ranking model

identify its k-nearest neighbors. Then, for each neighboring instance $\widetilde{x}$ in $\{N_k(x_i, j)|j = 1, ..., k\}$, we create a new instance by using the features related to $\widetilde{x}$ and x_i as the training instances for the ranking model. The new instance q contains the following features:

- The original features of x_i (size = $|x|$)
- The difference between each feature value of $\widetilde{x}$ and x_i (size = $|x|$)
- The Euclidean distance between $\widetilde{x}$ and x_i (size = 1)

- The cosine distance between $\widetilde{x}$ and x_i (size = 1)
- The label set of $\widetilde{x}$ (size = $|L|$)

Since the goal is to train a model that can learn the trustiness of an instance's neighbors, we employ the Hamming loss between the neighbor's label set and x_i's label set to determine the quality of each new instance q. Based on this value, a pair-wise comparison can be made and a ranking-based classifier can be trained. The lower the Hamming loss, the higher will be the rank assigned to a new instance q.

The weight scores for re-ranked neighbors are determined by the solution of an optimization problem, which aims at minimizing Hamming Loss. Let $(w_1, ..., w_k)$ denote weight scores of the re-ranked neighbors $\{N'_k(x,j)|j = 1, ..., k\}$. For each label $\lambda_i \in L$, it is possible to produce an accumulated score as the weighted sum of k scores from each re-ranked neighbor for λ_i

$$f_i(x) = \frac{\sum_{j=1}^{k} w_j \cdot y_i(N'_k(x,j))}{\sum_{j=1}^{k} w_j}. \quad (1)$$

The final prediction of the label set of the test instance x is defined as

$$H(x) = \{\lambda_i | f_i(x) \geq 0.5\}. \quad (2)$$

To determine the optimal weights, the optimization problem is formulated as follows:

$$\begin{aligned} \underset{w_1,...,w_k}{\text{minimize}} \quad & \sum_{x \in T'} HammingLoss(Y(x), H(x)) \\ \text{subject to} \quad & 1 \geq w_1, ..., w_k \geq 0 \\ & w_1 \geq w_2 \geq ... \geq w_k. \end{aligned} \quad (3)$$

Some constraints are added to this objective function. First, we consider all weights are between 0 and 1. Second, the neighbor with a higher rank should be associated with a higher weight. We solve the problem by generalized pattern search [4].

IV. Experiment

We conduct experiments on six commonly used data sets belonging to different domains. The statistics of the data sets are shown in Table I. Ranking SVM [5] is used to train our ranking model. We compare the proposed method's performance with that of two other multi-label instance-based learning algorithms: MLKNN and IBLR. Both algorithms are parameterized by the size of the neighborhood k. Following their experiment setup [2], [3], we set the value of $k = 10$, and use the Euclidean metric as the distance function. For the baseline, we use binary relevance learning with the KNN classifier (k = 10).

We perform 10-fold 5-repeat cross-validation on these data sets. The Hamming Loss results are shown in Table II. The numbers in parentheses represent the rank of the algorithms among the compared algorithms. Overall, the proposed methods significantly outperform the compared methods on each measure.

Table I
STATISTICS OF THE MULTI-LABEL DATA SETS

	instances	features	labels	cardinality	distinct
yeast	2417	103	14	4.237	198
scene	2407	294	6	1.074	15
emotions	593	72	6	1.869	27
audio	2472	177	45	4.119	1553
genbase	662	1186	27	1.252	32
medical	978	1449	45	1.245	94

Table II
HAMMING LOSS RESULT

	MLKNN	IBLR	BR-KNN	Our Method
yeast	0.1944 (3)	0.1935 (2)	0.1983 (4)	0.1910 (1)
scene	0.0857 (3)	0.0839 (2)	0.0931 (4)	0.0817 (1)
emotions	0.2615 (4)	0.1860 (1)	0.1936 (3)	0.1866 (2)
audio	0.0896 (4)	0.0840 (1)	0.0887 (3)	0.0857 (2)
genbase	0.0046 (4)	0.0021 (2)	0.0031 (3)	0.0011 (1)
medical	0.0155 (2)	0.0187 (4)	0.0172 (3)	0.0117 (1)

V. Conclusion

The major contributions of this paper are as follows. First, we observe an interesting phenomenon from the data, namely, it is possible to improve the accuracy of state-of-the-art multi-label classification approaches if the trustable neighbors of instances can be identified. Second, based on the above finding, we present a method that combines weighted KNN and ranked learning methods to solve the multi-label classification problem. The experiment results demonstrate the efficacy of our approach. Note that we have proposed a framework for multi-label classification rather than an algorithm. It is also possible to use another ranked-based learner or search technique based on the characteristics of the dataset.

References

[1] G. Tsoumakas and I. Katakis, "Multi label classification: An overview," *International Journal of Data Warehousing and Mining*, vol. 3, no. 3, pp. 1–13, 2007.

[2] Min-Ling Zhang and Zhi-Hua Zhou, "Ml-knn: A lazy learning approach to multi-label learning," *Pattern Recogn.*, vol. 40, no. 7, pp. 2038–2048, 2007.

[3] Weiwei Cheng and Eyke Hllermeier, "Combining instance-based learning and logistic regression for multilabel classification," *Machine Learning*, vol. 76, no. 2-3, pp. 211–225, September 2009.

[4] Tamara G. Kolda, Robert Michael Lewis, and Virginia Torczon, "Optimization by direct search: New perspectives on some classical and modern methods," *SIAM Review*, vol. 45, pp. 385–482, 2003.

[5] Thorsten Joachims, "Optimizing search engines using clickthrough data," in *KDD '02: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, New York, NY, USA, 2002, pp. 133–142, ACM.