# Three-Way Decisions Based Multi-label Learning Algorithm with Label Dependency

Feng Li[1,2], Duoqian Miao[1,2(✉)], and Wei Zhang[1,2]

[1] Department of Computer Science and Technology, Tongji University, Shanghai 201804, China
tjleefeng@hotmail.com, dqmiao@tongji.edu.cn
[2] Key Laboratory of Embedded Systems and Service Computing, Ministry of Education, Tongji University, Shanghai 201804, China

**Abstract.** A great number of algorithms have been proposed for multi-label learning, and these algorithms usually divide the labels with an optimal threshold according to their relevances to an unseen instance. However, it may easily cause misclassification to directly determine whether an unseen instance has the label with relevance close to the threshold. The label with relevance close to the threshold has a high uncertainty. Three-way decisions theory is an efficient method to solve the uncertainty problem. Therefore, based on three-way decisions theory, a multi-label learning algorithm with label dependency is proposed in this paper. Label dependency is an inherent property in multi-label data. The labels with high uncertainty are further handled with a label dependency model, which is represented by the logistic regression in this paper. The experimental results show that this algorithm performs better.

**Keywords:** Multi-label learning · Label dependency · Three-way decisions · Logistic regression

## 1 Introduction

Multi-label learning is a challenging problem in machine learning field, because multi-label instances have several possible labels simultaneously and labels have correlations with each other in multi-label data. Given a predefined label space $L$, the task of multi-label learning algorithm is to predict a set of relevant class labels $Y$ for an unseen instance through analyzing training instances with known label sets, where $Y \subset L$ and $|Y| \geq 1$ [1–3]. Multi-label objects exist widely in various real-world domains. For example, in the image domain [4], a picture may express multiple semantic classes simultaneously, such as *sea*, *beach* and *sky*. In the text domain [5], a document possibly belongs to several topics, such as *society*, *sport* and *politics*. In the biology domain [6], a gene could have a set of functions, such as *transcription* and *metabolism*. In the video domain [7], a movie may be labeled with several genres, such as *horror*, *cartoon* and *family*.

*Multi-label classification* (MLC) and *multi-label ranking* (MLR) are two major tasks in multi-label learning [1]. MLC predicts binary values for an unseen


V. Flores et al. (Eds.): IJCRS 2016, LNAI 9920, pp. 240–249, 2016.
DOI: 10.1007/978-3-319-47160-0_22

instance instructing relevant or irrelevant to labels, while MLR yields an order of labels according to their relevances to an unseen instance. The outputs of them, especially MLR, greatly depend on the label relevance. There are several ways to measure the label relevance, such as vote, possibility and membership degree. Here, the possibility is used for investigation in this paper, and others have the similar disciplines. Most of the multi-label algorithms firstly predict relevances that an unseen instance has the labels, then find a threshold $t$ to get a bipartition of the labels into relevant or irrelevant. The instance more possibly has the label with greater relevance. On the contrary, those with smaller relevance are more likely to be not associated to the instance. Therefore, it is very certain that the instance has the labels with very great relevance and does not have the labels with very few relevance. However, it is hard to judge whether the instance has the label with a relevance around the threshold, which is full of uncertainty, usually resulting in misclassification.

Three-way decisions theory is an efficient method to solve the uncertainty problem, which is proposed by Yao [8]. The method can improve the algorithm performance, and simplify the complex problem. It divides the problem into three regions, and different decisions are taken for different regions. Normally, the problem in the uncertain region will be further handled to make the right judgement. According to the relevance, the labels can be grouped into three regions in multi-label learning. The region with great relevance is the positive region, the region with few relevance is the negative region, and the region between them is called the boundary region. The labels in the positive region are assigned to the instance, while those in the negative region are not. We are not sure about labels in the boundary region, needing a further learning.

In multi-label data, there usually exists dependency among labels. For example, an action movie is more likely to be an adventure movie than be a romance at the same time. Label dependency is a hot topic, and there are a great number of algorithms about how to explore the label dependency in multi-label learning [10–12]. Hence, the labels in boundary region can be further predicted with the help of labels in the positive and the negative regions by using the label dependency. We propose a multi-label learning algorithm with label dependency based on three-way decisions theory to improve the algorithm performance. A logistic regression model is constructed to represent the label dependency. We experiment the proposed algorithm on multi-label data sets, and the results show that the proposed algorithm can achieve a better performance.

The rest of this paper is organized as follows: Sect. 2 briefly reviews the related work of multi-label learning. In Sect. 3, some basic concepts of three-way decisions theory and multi-label learning are introduced. In Sect. 4, we learn a model to revise the uncertain labels with label dependency. Section 5 displays the experimental results. We conclude the paper in Sect. 6.

## 2 Related Work

In recent years, multi-label learning has attracted significant attentions from various domains, and been a hot topic in machine learning field. A lot of

multi-label learning algorithms have been proposed. These proposed multi-label learning algorithms can be divided into two groups: *problem transportation method* (PTM) and *algorithm adaption method* (AAM) [1]. PTM is independent on algorithm, and transforms the multi-label data into numerous single-label data, such as Binary Relevance (BR) [13], Pairwise Binary (PW) [14], and Label Powerset (LP) [15]. AAM on the other hand extends some specific traditional machine learning algorithms to handle the multi-label data directly, such as decision tree [16], support vector machine [17], neural networks [18] and rough sets [12].

Furthermore, based on rough sets, Yu [12] proposed a multi-label learning with exploiting label correlation, called MLRS-LC. To exploit the label dependency, Zhang [10] proposed a multi-label learning by exploiting label dependency, which uses a Bayesian network structure to efficiently encode the conditional dependencies of the labels as well as the feature set. Kang [11] correlated label propagation with application to multi-label learning, which explicitly models interactions between labels in an efficient manner. In a word, the label dependency should be taken into consideration.

## 3 Preliminaries

### 3.1 Three-Way Decisions

Three-way decisions theory is a proper semantic explanation of probabilistic rough sets and decision-theoretic rough sets [8,9]. The main idea is to divide the whole into three regions, and different regions are treated with different ways. Let $Pr(X|[x])$ denote the conditional probability that $x$ belongs to $X$.

$$Pr(X|[x]) = \frac{|[x] \cap X|}{|[x]|} \tag{1}$$

$[x]$ is the equivalence class of $x$, and $|\cdot|$ stands for the cardinality. Then, the three regions of three-way decisions can be represented by probabilistic rough sets [19] as follow:

$$\begin{aligned} &POS(X) = \{x|Pr(X|[x]) \geq \alpha\}; \\ &NEG(X) = \{x|Pr(X|[x]) \leq 1-\alpha\}; \\ &BND(X) = \{x|1-\alpha < Pr(X|[x]) < \alpha\}. \end{aligned} \tag{2}$$

where $POS(X)$ denotes the positive region of $X$, $NEG(X)$ denotes the negative region, and $BND(X)$ is the boundary region. $\alpha$ is a threshold and $\alpha \in [0.5, 1]$. When $\alpha = 0.5$, the three-way decisions become the two-way decisions.

The three-way decisions theory is a generalized and efficient model for decisions and information processing, not limited for rough sets. There widely exist three-way phenomena in the real-world.

### 3.2 Multi-label Learning

Formally, let $F \subset R^b$ represent the input feature space, and $L = \{l_1, l_2, ..., l_q\}$ denote the label space with $q$ possible labels. Given a multi-label training data

$T = \{(X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n)\}$, $X_i$ is a $b$-dimensional input feature vector, and $Y_i = \{y_i^1, y_i^2, ..., y_i^q\}$ is the binary label vector of $X_i$, where $y_i^j$ equals to 1 if $X_i$ has label $l_j$, and equals to -1, otherwise. The task of multi-label learning is to derive a multi-label classification function $h : F \rightarrow \{0, 1\}^q$, through the training data $T$. For an unseen instance $X$, the multi-label classification function can predict its relevant label vector $Y^{'} = \{y^{1'}, y^{2'}, ..., y^{q'}\}$.

However, most of multi-label learning algorithms do not directly predict whether the instance $X$ has the label $l_j$, but firstly give a relevance $h^j(X)$ between $X$ and $l_j$, which is usually a possibility, then, divide the labels with an optimal threshold $t$ as follows:

$$y^{j'} = \begin{cases} 1, & h^j(X) \geq t \\ -1, & h^j(X) < t \end{cases} \tag{3}$$

The relevance $h^j(X)$ is a certainty degree that the label $l_j$ belongs to $X$. So it is very certain that $X$ belongs to the labels with relevances significantly greater than the threshold $t$. It is almost impossible that the labels with relevances much less than $t$ are assigned to $X$, namely, these labels are very certain to not relate to $X$. It is full of uncertainties that the labels with relevances around $t$. The closer to $t$ the relevance gets, the more uncertain the label is. Therefore, the three-way decisions theory is used to solve the problem in multi-label learning.

## 4 The Proposed Algorithm

Label dependency is important information contained in multi-label data, and a hot topic in multi-label learning. Therefore, it is a practicable way to correct the labels with high uncertainties through label dependency which can be represent by a dependency model of a label on the other $q - 1$ labels. Here, the logistic regression is used to construct the dependency model of label $l_j$ on the others.

$$g^j(X) = \frac{1}{1 + e^{-u_j}} \tag{4}$$

where the equation is a sigmod function, and

$$u_j = \theta_{j1} * y^1 + ... + \theta_{jj-1} * y^{j-1} + \theta_{jj+1} * y^{j+1} + ... + \theta_{jq} * y^q + \theta_{jj} \tag{5}$$

$\theta_{ji(i \neq j)}$ is the weight of $l_i$ to $l_j$ which informs the dependency between $l_i$ to $l_j$, and $\theta_{jj}$ is a constant for $l_j$. Equation (5) can be rewritten as:

$$u_j = \theta_j * y_j^T \tag{6}$$

In the equation, $\theta_j = \{\theta_{j1}, \theta_{j2}, ..., \theta_{jq}\}$ is the weight vector, and $y_j = \{y^1, ..., y^{j-1}, 1, y^{j+1}, ..., y^q\}$ is the input vector where the input for constant is set to 1. Then,

$$g^j(X) = \frac{1}{1 + e^{-\theta_j * y_j^T}} \tag{7}$$

The weight vector $\theta_j$ is trained with the label information in training data set.

Given a test instance $X$, and its label relevance $h(X)$ predicted by a multi-label learning algorithm, the label space $L$ can be grouped into three regions for $X$ according to $h(X)$, namely, the positive region $POS(X)$ in which the labels are assigned to $X$, the negative region $NEG(X)$ in which the labels are not related to $X$, and the boundary region $BND(X)$ where the labels are uncertain and need to be further predicted. The three regions can be defined as:

$$\begin{aligned} &\text{if } h^j(X) \geq t + \beta, \text{ then } l_j \in POS(X); \\ &\text{if } h^j(X) \leq t - \beta, \text{ then } l_j \in NEG(X); \\ &\text{if } t - \beta < h^j(X) < t + \beta, \text{ then } l_j \in BND(X). \end{aligned} \tag{8}$$

where $t$ is the optimal threshold in the original multi-label learning algorithm, and $\beta \in [0, \min(t, 1-t)]$ determines the width of the boundary region, i.e. the uncertainty region. A three-value label vector $Z = \{z^1, z^2, ..., z^q\}$ can be gotten for $X$ as follows:

$$z^j = \begin{cases} 1, & l_j \in POS(X) \\ 0, & l_j \in BND(X) \\ -1, & l_j \in NEG(X) \end{cases} \tag{9}$$

The labels in $POS(X)$ $[NEG(X)]$ have very high certainty degrees belonging [not belonging] to $X$, and do not need to be further processed and changed. Suppose $t \geq (1-t)$, then $\beta \in [0, 1-t]$. $Z$ is used as an input vector of Eq. (7) to obtain a correction term $\varphi_j$ for label $l_j \in BND(X)$

$$\varphi_j = \frac{1}{1 + e^{-(\theta_j * Z^T + \theta_{jj})}} \tag{10}$$

For $l_j \in BND(X)$, $z^j = 0$, the constant $\theta_{jj}$ is added. In the input vector $Z$, the values of the labels in $BND(X)$ are 0, means they have no influence on $\varphi_j$, because of their high certainties. For $l_j \in BND(X)$, $\varphi_j$ is added to the original label relevance $h^j(X)$. Therefore the label relevance after correcting $f^j(X)$ is computed as follows:

$$f^j(X) = \begin{cases} (t + \beta) * h^j(X) + (1 - t - \beta) * \varphi_j, & \text{if } l_j \in BND(X) \\ h^j(X), & \text{otherwise} \end{cases} \tag{11}$$

The formula considers the label relevance predicted from the features by the original multi-label learning algorithm and the label relevance from label dependency simultaneously. $(1 - t - \beta)$ is the weight of the correction term, and determines influence of the correction term. The boundary region becomes lager with the increase of $\beta$, leading to the rising of number of uncertain labels and decreasing of the reliability of the correction term. Therefore, it can be seen that the weight of the correction term decreases as the $\beta$ increases. When $\beta = 0$, there is no uncertain label needing to be corrected, so the label relevance keeps the same. When $\beta = 1 - t$, the certain labels is the least, so the weight of the correction term equals to 0, and no change is on the label relevance.

The label $l_j$ can be predicted whether be associated to $X$ or not by using label relevance $f^j(X)$ after correcting as:

$$y^{j'} = \begin{cases} 1, & f^j(X) \geq t \\ -1, & f^j(X) \leq t \end{cases} \tag{12}$$

**Algorithm 1.** The multi-label learning algorithm with label dependency based on three-way decisions theory

**Input:** Original label relevance $h(X)$; Parameter of the width of boundary region $\beta$; Optimal threshold $t$; Label data $W = \{(Y_1), (Y_2), ..., (Y_n)\}$;
**Output:** Predicted label vector $Y^{'}$
**for** $l_j \in L$ **do**
//Initialize variables;
$z^j \leftarrow 0$;
$\varphi_j \leftarrow 0$;
$f^j(X) \leftarrow 0$;
$y^{j'} \leftarrow 0$;
Compute the value $z^j$ with $h^j(X)$ according to equations (8) and (9);
**end for**
**for** $l_j \in L$ **do**
Construct the logistic regression model with $W$ to get the weight vector $\theta_j$;
Count the correction term $\varphi_j$ according to equation (10);
Calculate $f^j(X)$ according to equation (11);
Determine $y^{j'}$ according to equation (12)
**end for**
Output the predicted label vector $Y^{'} = \{y^{1'}, y^{2'}, ..., y^{q'}\}$;

## 5 Experimental Results

### 5.1 Data Sets

We experiment on three real-world multi-label data sets covering different domains from the Mulan Libary [20]. The statistical information is summarized in Table 1. As shown in Table 1, *Medical* [21] data set has 978 instances, of

**Table 1.** Multi-label data sets in the experiments

| Name | Instance | Feature | Label | Cardinality |
|---|---|---|---|---|
| Medical | 978 | 1449 | 45 | 1.245 |
| Enron | 1702 | 1001 | 53 | 3.378 |
| CAL500 | 502 | 68 | 174 | 26.044 |

which each instance is a radiology text report consisting of the medical history and symptom and is associated with a subset of 45 ICD-9-CM labels. There are 1702 instances in *Enron* [22] data set, and these instances are e-mails of the Enron company and labeled with 53 possible tags. *CAL500* [23] data set contains 502 popular musical tracks, and 174 labels such as style, emotion and instrument.

### 5.2 Evaluation Criteria

Five example based multi-label learning evaluation criteria are considered, *Hamming loss*, *Precision*, *Recall*, *F1-measure*, *Accuracy* [1]. The larger the latter four evaluation criteria are, the better the algorithm performs, while *Hamming loss* in contrast. Given a testing multi-label data set $D = \{(X_1, Y_1), (X_2, Y_2), ..., (X_m, Y_m)\}$, the five evaluation criteria are defined as follows:

*Hamming loss* evaluates how many labels belonging to the instance is not associated, or not belonging to the instance is associated. $\langle\pi\rangle$ equals to 1 if $\pi$ holds and 0 otherwise. $Hamloss = \frac{1}{mq}\sum_{i=1}^{m}\sum_{j=1}^{q}\langle y_{ij}^{'} \neq y_{ij}\rangle$.
*Precision* evaluates how many labels actually belong to the instance in the predicted label set. $Precision = \frac{1}{m}\sum_{i=1}^{m}\frac{|Y_i \cap Y_i^{'}|}{|Y_i^{'}|}$.
*Recall* computes the number of labels that the are correctly predicted in the ground-truth label set. $Recall = \frac{1}{m}\sum_{i=1}^{m}\frac{|Y_i \cap Y_i^{'}|}{|Y_i|}$.
*F1-measure* is the harmonic mean between *precision* and *recall*, common to information retrieval. $F1 = \frac{1}{m}\sum_{i=1}^{m}\frac{|Y_i \cap Y_i^{'}|}{|Y_i|+|Y_i^{'}|}$.
*Accuracy* measures the average degree of similarity between the predicted and the ground-truth label sets of all testing instances. $Accuracy = \frac{1}{m}\sum_{i=1}^{m}\frac{|Y_i \cap Y_i^{'}|}{|Y_i \cup Y_i^{'}|}$.

### 5.3 Results and Discussion

The *ten-fold cross-validations* evaluation is used to evaluate algorithms in the experiment. ML-KNN is a popular multi-label algorithm and chosen to produce the original label relevance. As recommended in [24], the number of neighbors is 10, and the threshold $t$ is set to be 0.5. The $\beta$ arranges from 0 to 0.5 with a step of 0.05. In the following tables, the symbol $'\downarrow'$ represents that the smaller the evaluation criterion value is, the better the performance is, while the symbol $'\uparrow'$ in contrast. Furthermore, the best result is marked in boldface on each evaluation criterion by considering the mean value.

When $\beta$ is set to be 0, there are no labels changed on all data sets. Therefore, the algorithm results with $\beta$ equal to 0 are the same as the original results achieved by the ML-KNN. Tables 2, 3 and 4 show that when $\beta$ is 0.5, the algorithm performance is the same as the original performance, too, the reason of which has been discussed in Sect. 4. In more detail, the performance is improved

**Table 2.** Experimental results (mean ± std. deviation) on the *Medical* data set.

| $\beta$ | *Hamming loss*↓ | *Precision*↑ | *Recall*↑ | *F1*↑ | *Accuracy*↑ |
|---|---|---|---|---|---|
| 0 | 0.0155 ± 0.0070 | 0.6232 ± 0.2132 | 0.5888 ± 0.1727 | 0.2970 ± 0.0929 | 0.0163 ± 0.0049 |
| 0.05 | 0.0154 ± 0.0074 | 0.6428 ± 0.1966 | 0.6012 ± 0.1481 | 0.3048 ± 0.0823 | 0.0166 ± 0.0048 |
| 0.10 | 0.0154 ± 0.0076 | 0.6536 ± 0.2019 | 0.6064 ± 0.1633 | 0.3086 ± 0.0872 | 0.0167 ± 0.0052 |
| 0.15 | **0.0153 ± 0.0081** | **0.6665 ± 0.2049** | **0.6153 ± 0.1807** | **0.3139 ± 0.0930** | **0.0169 ± 0.0056** |
| 0.20 | 0.0157 ± 0.0078 | 0.6624 ± 0.1811 | 0.6134 ± 0.1591 | 0.3125 ± 0.0813 | 0.0168 ± 0.0047 |
| 0.25 | 0.0157 ± 0.0085 | 0.6598 ± 0.1781 | 0.6082 ± 0.1870 | 0.3106 ± 0.0885 | 0.0166 ± 0.0057 |
| 0.30 | 0.0161 ± 0.0087 | 0.6412 ± 0.2011 | 0.5921 ± 0.2113 | 0.3021 ± 0.1011 | 0.0161 ± 0.0063 |
| 0.35 | 0.0174 ± 0.0083 | 0.5649 ± 0.2163 | 0.5210 ± 0.2476 | 0.2656 ± 0.1149 | 0.0142 ± 0.0071 |
| 0.40 | 0.0189 ± 0.0052 | 0.4665 ± 0.1622 | 0.4299 ± 0.1886 | 0.2190 ± 0.0851 | 0.0119 ± 0.0052 |
| 0.45 | 0.0201 ± 0.0060 | 0.3804 ± 0.2210 | 0.3503 ± 0.2039 | 0.1781 ± 0.1033 | 0.0099 ± 0.0060 |
| 0.50 | 0.0155 ± 0.0070 | 0.6232 ± 0.2132 | 0.5888 ± 0.1727 | 0.2970 ± 0.0929 | 0.0163 ± 0.0049 |

**Table 3.** Experimental results (mean ± std. deviation) on the *Enron* data set.

| $\beta$ | *Hamming loss*↓ | *Precision*↑ | *Recall*↑ | *F1*↑ | *Accuracy*↑ |
|---|---|---|---|---|---|
| 0 | 0.0539 ± 0.0074 | 0.5644 ± 0.0915 | 0.3443 ± 0.0731 | 0.1997 ± 0.0354 | 0.0239 ± 0.0038 |
| 0.05 | 0.0540 ± 0.0067 | 0.5715 ± 0.1037 | 0.3549 ± 0.1116 | 0.2047 ± 0.0499 | 0.0243 ± 0.0054 |
| 0.10 | 0.0546 ± 0.0063 | 0.5746 ± 0.1025 | 0.3627 ± 0.1466 | 0.2083 ± 0.0638 | **0.0244 ± 0.0083** |
| 0.15 | 0.0548 ± 0.0058 | 0.5685 ± 0.1056 | **0.3669 ± 0.1825** | **0.2091 ± 0.0774** | 0.0241 ± 0.0106 |
| 0.20 | 0.0553 ± 0.0069 | 0.5477 ± 0.1560 | 0.3411 ± 0.2450 | 0.1967 ± 0.1104 | 0.0232 ± 0.0136 |
| 0.25 | 0.0543 ± 0.0086 | 0.5631 ± 0.1078 | 0.3441 ± 0.1930 | 0.1998 ± 0.0850 | 0.0232 ± 0.0113 |
| 0.30 | 0.0539 ± 0.0080 | 0.5705 ± 0.0996 | 0.3426 ± 0.1699 | 0.2002 ± 0.0759 | 0.0230 ± 0.0096 |
| 0.35 | 0.0539 ± 0.0071 | 0.5765 ± 0.0865 | 0.3518 ± 0.1120 | 0.2046 ± 0.0485 | 0.0238 ± 0.0065 |
| 0.40 | **0.0536 ± 0.0072** | **0.5780 ± 0.0851** | 0.3442 ± 0.1106 | 0.2015 ± 0.0491 | 0.0236 ± 0.0055 |
| 0.45 | 0.0538 ± 0.0069 | 0.5549 ± 0.1034 | 0.3345 ± 0.0917 | 0.1952 ± 0.0430 | 0.0238 ± 0.0046 |
| 0.50 | 0.0539 ± 0.0074 | 0.5644 ± 0.0915 | 0.3443 ± 0.0731 | 0.1997 ± 0.0354 | 0.0239 ± 0.0038 |

to reach the best, then decreases gradually. That is because when $\beta$ is too small, not many labels are corrected, while the certain labels are not enough to produce a reliable correction term, if $\beta$ is too large. Thus, it is a proper $\beta$ that could produce a balance between the number of the certain labels and the number of the uncertain to obtain the best performance.

As shown in Table 2, the proposed algorithm obtains the best performance on all evaluation criteria on *Medical* data set, when $\beta$ is equal to 0.15. The proposed algorithm improves the performance of *Medical*, especially on *precision*, *recall* and *F1*. On the *Enron* data set, the proposed algorithm performs best on *hamming loss* and *precision* when $\beta$ is 0.4, while it achieves the best results on *recall* and *F1* when $\beta$ is 0.15 and *accuracy* when $\beta$ is 0.1. On the *CAL500* data set, the proposed algorithm performs best when $\beta$ is 0.3 on all evaluation criteria except for *precision*, which is the best when $\beta$ is 0.05. All the best results of the proposed algorithm are better than the original ones without correcting, and it promotes the original performance.

**Table 4.** Experimental results (mean ± std. deviation) on the *CAL500* data set.

| $\beta$ | *Hamming loss*↓ | *Precision*↑ | *Recall*↑ | *F1*↑ | *Accuracy*↑ |
|---|---|---|---|---|---|
| 0 | $0.1399 \pm 0.0201$ | $0.5927 \pm 0.0732$ | $0.2247 \pm 0.0370$ | $0.1604 \pm 0.0219$ | $0.0394 \pm 0.0038$ |
| 0.05 | $0.1390 \pm 0.0212$ | $\mathbf{0.6119 \pm 0.0702}$ | $0.2117 \pm 0.0456$ | $0.1547 \pm 0.0277$ | $0.0369 \pm 0.0068$ |
| 0.10 | $0.1394 \pm 0.0240$ | $0.6077 \pm 0.0849$ | $0.2059 \pm 0.0906$ | $0.1511 \pm 0.0562$ | $0.0359 \pm 0.0142$ |
| 0.15 | $0.1400 \pm 0.0215$ | $0.5992 \pm 0.0779$ | $0.2078 \pm 0.0894$ | $0.1513 \pm 0.0559$ | $0.0364 \pm 0.0149$ |
| 0.20 | $0.1411 \pm 0.0211$ | $0.5844 \pm 0.0844$ | $0.2187 \pm 0.0633$ | $0.1561 \pm 0.0384$ | $0.0385 \pm 0.0090$ |
| 0.25 | $0.1409 \pm 0.0202$ | $0.5831 \pm 0.0771$ | $0.2292 \pm 0.0544$ | $0.1615 \pm 0.0319$ | $0.0404 \pm 0.0076$ |
| 0.30 | $\mathbf{0.1389 \pm 0.0211}$ | $0.5962 \pm 0.0696$ | $\mathbf{0.2362 \pm 0.0502}$ | $\mathbf{0.1666 \pm 0.0290}$ | $\mathbf{0.0416 \pm 0.0047}$ |
| 0.35 | $0.1394 \pm 0.0210$ | $0.5990 \pm 0.0781$ | $0.2253 \pm 0.0469$ | $0.1612 \pm 0.0283$ | $0.0397 \pm 0.0051$ |
| 0.40 | $0.1390 \pm 0.0211$ | $0.6077 \pm 0.0901$ | $0.2122 \pm 0.0519$ | $0.1549 \pm 0.0322$ | $0.0372 \pm 0.0070$ |
| 0.45 | $0.1392 \pm 0.0181$ | $0.6038 \pm 0.0722$ | $0.2162 \pm 0.0278$ | $0.1568 \pm 0.0183$ | $0.0380 \pm 0.0045$ |
| 0.50 | $0.1399 \pm 0.0201$ | $0.5927 \pm 0.0723$ | $0.2247 \pm 0.0370$ | $0.1604 \pm 0.0219$ | $0.0394 \pm 0.0038$ |

## 6 Conclusions

By using a logistic regression model of label dependency, this paper proposed a multi-label learning algorithm based on three-way decisions theory to further handle those labels with high uncertainty. The experimental results show that it is helpful to correct the labels near the threshold through the proposed algorithm. How to theoretically choose the best $\beta$ is not researched, hence in the next step, we will propose a theory analysis to choose an optimal width of the boundary region. Furthermore, two variables instead of a single variable $\beta$ are taken into consideration to determine the width of boundary region, which is more generalized and not restricted by the threshold.

**Acknowledgments.** The work is partially supported by the National Natural Science Foundation of China (Nos. 61273304, 61573259), the Specialized Research Fund for the Doctoral Program of Higher Education of China (No. 20130072130004), and the program of Further Accelerating the Development of Chinese Medicine Three Year Action of Shanghai (2014–2016) (No. ZY3-CCCX-3-6002).

## References

1. Tsoumakas, G., Katakis, I., Vlahavas, I.: Mining multi-label data. In: Maimon, O., Rokach, L. (eds.) Data Mining and Knowledge Discovery Handbook, pp. 667–685. Springer, New York (2010)
2. Tsoumakas, G., Katakis, I.: Multi label classification: an overview. Int. J. Data Warehouse. Min. **3**(3), 1–13 (2007)
3. Zhang, M.L., Zhou, Z.H.: A review on multi-label learning algorithms. IEEE Trans. Knowl. Data Eng. **26**(8), 1819–1837 (2014)
4. Yu, Y., Pedrycz, W., Miao, D.Q.: Neighborhood rough sets based multi-label classification for automatic image annotation. Int. J. Approximate Reason. **54**(9), 1373–1387 (2013)
5. Schapire, R.E., Singer, Y.: BoosTexter: a boosting-based system for text categorization. Mach. Learn. **39**(2), 135–168 (2000)

6. Pavlidis, P., Weston, J., Cai, J., Grundy, W.N.: Combining microarray expression data and phylogenetic profiles to learn functional categories using support vector machines. In: Proceedings of the Fifth Annual International Conference on Computational Biology, Montreal, Canada, pp. 242–248 (2001)
7. Snoek, C.G.M., Worring, M., Van Gemert, J.C., et al.: The challenge problem for automated detection of 101 semantic concepts in multimedia. In: Proceedings of the 14th Annual ACM International Conference on Multimedia, pp. 421–430 (2006)
8. Yao, Y.: An outline of a theory of three-way decisions. In: Yao, J.T., Yang, Y., Słowiński, R., Greco, S., Li, H., Mitra, S., Polkowski, L. (eds.) RSCTC 2012. LNCS, vol. 7413, pp. 1–17. Springer, Heidelberg (2012). doi:10.1007/978-3-642-32115-3_1
9. Pawlak, Z.: Rough sets. Int. J. Parallel Prog. **11**(5), 341–356 (1982)
10. Zhang, M.L., Zhang, K.: Multi-label learning by exploiting label dependency. In Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 999–1008. ACM, New York (2010)
11. Kang, F., Jin, R., Sukthankar, R.: Correlated label propagation with application to multi-label learning. In: Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 1719–1726 (2006)
12. Yu, Y., Predrycz, W., Miao, D.Q.: Multi-label classification by exploiting label correlations. Expert Syst. Appl. **41**(6), 2989–3004 (2014)
13. Boutell, M.R., Luo, J., Shen, X., et al.: Learning multi-label scene classification. Pattern Recogn. **37**(9), 1757–1771 (2004)
14. Hllermeier, E., Frnkranz, J., Cheng, W., et al.: Label ranking by learning pairwise preferences. Artif. Intell. **172**(16), 1897–1916 (2008)
15. Tsoumakas, G., Vlahavas, I.P.: Random $k$-labelsets: an ensemble method for multilabel classification. In: Kok, J.N., Koronacki, J., Lopez de Mantaras, R., Matwin, S., Mladenič, D., Skowron, A. (eds.) ECML 2007. LNCS (LNAI), vol. 4701, pp. 406–417. Springer, Heidelberg (2007). doi:10.1007/978-3-540-74958-5_38
16. Clare, A., King, R.D.: Knowledge discovery in multi-label phenotype data. In: Raedt, L., Siebes, A. (eds.) PKDD 2001. LNCS (LNAI), vol. 2168, pp. 42–53. Springer, Heidelberg (2001). doi:10.1007/3-540-44794-6_4
17. Elisseeff, A., Weston, J.: A kernel method for multi-labelled classification. In: Advances in Neural Information Processing Systems, vol. 14, pp. 681–687 (2001)
18. Zhang, M.L., Zhou, Z.H.: Multilabel neural networks with applications to functional genomics and text categorization. IEEE Trans. Knowl. Data Eng. **18**(10), 1338–1351 (2006)
19. Yao, Y.Y.: Three-way decisions with probabilistic rough sets. Inf. Sci. **180**(3), 341–353 (2010)
20. Tsoumakas, G., Spyromitros-Xiousfis, E., Vilcek, I.V.J.: Mulan: a Java library for multi-label learning. J. Mach. Learn. Res. **12**(7), 2411–2414 (2011)
21. Pestian, J., Brew, C., Matykiewicz, P., et al.: A shared task involving multi-label classification of clinical free text. In: Proceedings of the Workshop on BioNLp 2007, pp. 97–104. Association for Computational Linguistics, Stroudsburg (2007)
22. UC Berkeley Enron Email Analysis Project. http://bailando.sims.berkeley.edu/enron_email.html
23. Turnbull, D., Barrington, L., Torres, D., et al.: Semantic annotation and retrieval of music and sound effects. IEEE Trans. Audio Speech Lang. Process. **16**(2), 467–476 (2008)
24. Zhang, M.L., Zhou, Z.H.: ML-kNN: a lazy learning approach to multi-label learning. Pattern Recogn. **40**(7), 2038–2048 (2007)