Multi-label Classification Using Rough Sets

Ying Yu[1,2,3,⋆], Duoqian Miao[1,2], Zhifei Zhang[1,2], and Lei Wang[1,2]

[1] Department of Computer Science and Technology, Tongji University, Shanghai 201804, P.R. China
[2] Key Laboratory of Embedded System and Service Computing, Ministry of Education, Tongji University, Shanghai 201804, P.R. China
[3] Software School, Jiangxi Agriculture University, Jiangxi 330045, P.R. China

Abstract. In multi-label classification, each instance may be associated with multiple labels simultaneously which is different from the traditional single-label classification where an instance is only associated with a single label. In this paper, we propose two types of approaches to deal with multi-label classification problem based on rough sets. The first type of approach is to transform the multi-label problem into one or more single-label problems and then use the classical rough set model to make decisions. The second type of approach is to extend the classical rough set model in order to handle multi-label dataset directly, where the new model considers the correlations among labels. The effectiveness of multi-label rough set model is presented by a series of experiments completed for two multi-label datasets.

Keywords: rough sets, multi-label classification, correlation.

1 Introduction

Multi-label classification problems [1] widely exist in various applications where each instance is normally associated with multiple labels and the classes encountered in the problem are not mutually exclusive but may overlap.

There exists uncertainty during the process of multi-label classification due to the finite number of training instances and the ambiguity of concept themselves, which impacts the precision of the prediction. However, there is a lack of study on the uncertainty existing in the multi-label classification. Rough sets form a conceptual vehicle to deal with ambiguous, vague, and uncertain knowledge [2]. In this paper, several methods based on rough sets are proposed for the multi-label decision system.

The rest of this paper is organized as follows. Section 2 briefly reviews the related studies about rough sets and multi-label learning. In Section 3, two types of approaches for multi-label problem are proposed, which are respectively based

⋆ This paper is partially supported by the National Natural Science Foundation of China (Serial No. 61075056, 61273304, 61075056, 61103067, 61202170), and the State Scholarship Fund of China (File No. 201206260047).

D. Ciucci et al. (Eds.): RSFDGrC 2013, LNAI 8170, pp. 119–126, 2013.

on classical rough set model and multi-label rough set model. Section 4 illustrates the effectiveness of multi-label rough set model through some experiments. Finally, Section 5 concludes the studies.

2 Related Works

This section briefly reviews some existing works on rough sets and multi-label learning that are pertinent to our study.

2.1 Rough Sets

Rough set theory, proposed in 1982 by Pawlak [2], is regarded as a tool to process inexact, uncertain or vague knowledge. Indiscernibility relation and Approximations are two important concepts in Pawlak rough set theory.

Rough set theory has attracted worldwide attention of many researchers and practitioners, who have contributed essentially to its development and applications. For example, in order to deal with incomplete information system, some researchers extend the equivalence relations to non-equivalence relations such as tolerance relation [3], similarity relation [4], limited tolerance relation [5], etc.. In order to support numerical attributes, Yao [6] and Hu [7] proposed the neighborhood rough set model based on the neighborhood relations.

2.2 Multi-label Learning

Multi-label classification is different from the traditional task of single-label classification where each instance is only associated with a single class label. An intuitive approach to multi-label learning is to decompose the task into a number of binary classification problems and each for one class. This kind of approaches include binary relevance method (BR) [1], binary pairwise classification approach (PW) [8] and label combination or label power-set method (LC) [9]. Such an approach, however, usually suffers from the deficiency that the correlation among the labels is not taken into account.

There are also numbers of multi-label classification algorithms derived from traditional machine learning methods. For example, Boostexter system [10] provides two boosting algorithms, Adaboost.MH and Adaboost.MR, which are two extensions of Adaboost for multi-label classification. Comit et al. [11] extended the alternating decision tree learning algorithm for multi-label classification. In addition, a number of multi-label methods are based on the popular k Nearest Neighbors (kNN) lazy learning algorithm [12].

3 Rough Sets Based Approaches for Multi-label Classification

In multi-label decision table, an object is associated with a subset of labels and different classes may overlap by definition in the feature space. Fig. 1(a) shows a

multi-label dataset which includes five instances with four labels *grass*, *tree*, *sky* and *water*. If we transfer Fig. 1(a) into Fig. 1(b), we find it looks like a single-label inconsistent decision table, where two objects with the same conditional features belong to different decision classes. In single-label classification system, the classes are mutually exclusive and the inconsistent problem was considered to be caused by noise, such as mistakes in recording process [13], which is in conflict with the definition of multi-label classification. We cannot directly cope with multi-label problem using the existing single-label inconsistent approaches. In this paper, we will present two types of rough sets based approaches for multi-label classification problem.

object	grass	tree	sky	water
1	X		X	
2		X	X	
3	X		X	X
4		X		
5			X	X

(a)

object	label
1	grass
1	sky
2	tree
2	sky
3	grass
3	sky
3	water
4	tree
5	sky
5	water

(b)

Fig. 1. Example of multi-label dataset and its transformation

Before introducing the methods, we present the formal notation in this paper. Let $MDT = \langle U, A \rangle$ be a multi-label decision table, where U is a finite, nonempty set called the universe, and $A = C \cup D$; $C = \{c_1, \ldots, c_n\}$ is the set of conditional attributes and $D = \{l_1, \ldots, l_m\}$ is the set of labels.

The first type of approach is to directly transform the multi-label problem into one or more traditional single-label problems and then use the classical rough set theory to obtain rules. As for the methods of transformation, we can refer to literature [1]. Fig. 1(a) is used as an original example to briefly exemplify these transformations.

For example, we can learn binary classifiers from original dataset, and one for each different label $l_j \in D$. Each dataset contains all instances of original dataset. The instance is labeled as 1, if the original label l_j is included and as 0, otherwise. Fig. 2 shows the result of transformation of Fig. 1(a) using this method. For a new object, its prediction is a set of labels which are output by classifiers. However, the precision of the decision suffers from the imbalance problem existing in the dataset.

In addition, we also can consider each different set of labels that exists in the multi-label dataset as a single-label. Fig. 3 shows the result of transformation of Fig. 1(a) using this method. The new labels come from the power set of D. This method suffers from the sparse problem that the dataset has a large number of classes as well as few examples per class.

object	grass
1	1
2	0
3	1
4	0
5	0

object	tree
1	0
2	1
3	0
4	1
5	0

object	sky
1	1
2	1
3	1
4	0
5	1

object	water
1	0
2	0
3	1
4	0
5	1

Fig. 2. Four datasets with binary labels

object	label
1	grass&sky
2	tree&sky
3	grass&sky&water
4	tree
5	sky&water

Fig. 3. Transformed dataset with power set

The second type of approach is to extend specific rough set model in order to handle multi-label data directly. It can be noticed from Fig. 1(a) that in multi-label dataset, different labels often co-occur in practice. Namely, the labels are not independent with each other. Taking Fig. 1(a) as an example, the probability of an image being annotated with label *sky* would be high if we know it has label *grass*. Thus, effective exploitation of correlation information among labels is crucial for the success of multi-label rough sets.

Generally speaking, the co-occurrence of labels is related with the location of instance. Those instances with multiple labels are usually located in the overlapped region. Fig. 4 gives an example to illustrate the relation between location and co-occurrence. Two labels are respectively marked by '*' and '+' in a 2-D space and examples simultaneously belonging to l_1 and l_2 are denoted by 'X'. For convenience, we assume that the distribution of two classes is circular. There are several instances in example space such as a, b, c,d and we associate a neighborhood with five neighbors to each instance. It can be seen that the instances located in the non-overlapped region only have one label while the instances located in the overlapped region may have two labels simultaneously. Let $\delta(x)$ denote the neighborhood of instance x and $|\delta_j(x)|$ is the number of instance with label $l_j (j = 1, \ldots, m)$ in $\delta(x)$. Let $\Gamma(x)$ denote the sum of all kinds of neighbors in $\delta(x)$ and $|\Gamma(x)| = \sum_{q=1}^{m} |\delta_q(x)|$. The proportion that the neighbors with label l_j accounts for of all kinds of neighbors is represented as $\Upsilon_j(x) = |\delta_j(x)|/|\Gamma(x)|$. Taking instances a and c as examples, $\Upsilon_1(a) = 1$ and $\Upsilon_2(a) = 0$ while $\Upsilon_1(b) = 1/6$ and $\Upsilon_2(b) = 5/6$. The proportion $\Upsilon_j(x)$ varies along the changing of location of instances. A larger value for $\Upsilon_j(x)$ will increase the probability of instance x having label l_j. Here, we first introduce the inclusion degree and then give the definition of upper and lower approximations of multi-label decision table according to the proportion $\Upsilon_j(x)$.

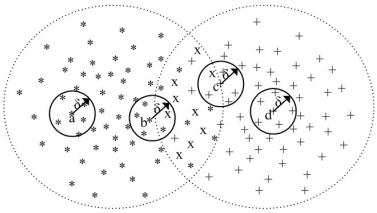

Fig. 4. Illustration of location estimation in multi-label system

Definition 1. *Given two sets A and B in the universe, the inclusion degree of A in B is defined as*

$$I(A,B) = \frac{Card(A \cap B)}{Card(A)} \tag{1}$$

where $Card(\Phi)$ stands for the number of elements in set Φ. The proportion $\Upsilon_j(x)$ can be described using inclusion degree as follows.

$$\Upsilon_j(x) = I(\Gamma(x), Y) = \frac{Card(\Gamma(x) \cap Y)}{Card(\Gamma(x))} \tag{2}$$

where Y represents the set of instances with label l_j in universe. Then the upper and lower approximations of decision class are defined as follows.

Definition 2. *Given a multi-label decision table $MDT = \langle U, A \rangle$, $X_i \in U$ and $A = C \cup D$; Y is the subset of instances with label $l_j (j = 1, \ldots, m)$ and $B \subseteq C$. Then the lower and upper approximations of decision class Y with respect to neighborhood relation R are denoted as $\underline{R_B^\beta}Y$ and $\overline{R_B^\alpha}Y$ respectively, and defined as follows.*

$$\underline{R_B^\beta}Y = \{x_i | I(\Gamma(x), Y) \geqslant \beta, x_i \in U\} \tag{3}$$

$$\overline{R_B^\alpha}Y = \{x_i | I(\Gamma(x), Y) \geqslant \alpha, x_i \in U\} \tag{4}$$

From the definition, we can see that just as decision-theoretic rough set models [14,15], the multi-label rough set model incorporates probabilistic approaches into rough set theory. For each label $l_j \in D$, inclusion degree β and $\alpha (0 \leqslant \alpha < \beta \leqslant 1)$ are different and they are estimated from the training dataset according to maximum posterior probability. Let l_j^1 denote the event of instance x_i having label l_j and l_j^0 denotes the event of instance x_i having no label l_j. $P(l_j^1 | \Upsilon_j(x_i))$ denotes the probability of instance x_i having label l_j, when the proportion is $\Upsilon_j(x_i))$ and $P(l_j^0 | \Upsilon_j(x_i))$ means just the opposite. Then according to Bayesian decision theory, if $P(l_j^1 | \Upsilon_j(x_i)) \geqslant P(l_j^0 | \Upsilon_j(x_i))$ then the instance x_i has label l_j, and otherwise the instance x_i has no relation with label l_j. The threshold β is determined when $P(l_j^1 | \Upsilon_j(x_i)) = P(l_j^0 | \Upsilon_j(x_i))$ and the threshold α is determined when $P(l_j^1 | \Upsilon_j(x_i))$ reaches a satisfied value. Taking Fig. 5 as an example, β is the threshold of lower approximation and α is selected as the threshold of upper approximation when $P(l_j^1 | \Upsilon_j(x_i))$ approaches zero.

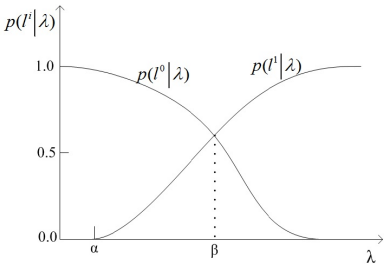

Fig. 5. Illustration of estimation of inclusion degree

For each label l_j, the multi-label rough set model divides the universe into three regions. Decision positive region is denoted by $POS_B(Y) = \underline{R_B^\beta} Y$ where the instances certainly belongs to class l_j. Negative region is denoted by $NEG_B(Y) = U - \overline{R_B^\alpha} Y$ where the instances have no relation with class l_j. The boundary region denoted by $BN_B(Y) = \overline{R_B^\alpha} Y - \underline{R_B^\beta} Y$ is a subset of instances that may have relation with class l_j.

After defining the upper and lower approximations of decision class, we will give the definition of multi-label decision function based on rough sets, which can be used for multi-label classification problem.

Definition 3. *Given a multi-label decision table* $MDT = \langle U, A \rangle$, $x_i \in U$. $\Upsilon_j(x_i)$ $(j = 1, \ldots, m)$ *is the proportion that the neighbors with label* l_j *in* $\delta(x_i)$ *have of all kinds of neighbors in* $\delta(x_i)$. *The multi-label decision function of* x_i *for label* l_j *is defined as* $MD_j(x_i) = l_j^1$, *if* $\Upsilon_j(x_i) \geqslant \beta$ *or* $MD_j(x_i) = l_j^0$, *if* $\Upsilon_j(x_i) \leqslant \alpha$.

$MD_j(x_i)$ is the result assigned to x_i according to the inclusion degree. Obviously, $MD_j(x_i) = l_j^1$ if x_i is located in the positive region of class l_j, or $MD_j(x_i) = l_j^0$ if x_i is located in the negative region of class l_j, or if x_i is located in the boundary region of class l_j, we will assign it a probability of having label l_j.

4 Experiments

To test the effectiveness of the multi-label rough set model(MLRS) presented in this paper, we apply it to two multi-label datasets which come from the the open source Mulan library [1] and Table 1 shows their associated properties. We compare MLRS with various state-of-art multi-label algorithms including the classifier chains algorithm CC, the random k label-set method for multi-label classification RAkEL and back-propagation multi-label learning (BPMLL) learner.

Experimental results of ten-fold cross-validation in terms of *Hamming loss*, *average precision*, *coverage*, *one-error* and *ranking loss* are shown in Table 2 and Table 3. The value following ± gives the standard deviation and the best result on each metric is shown in bold face. The number of the nearest neighbors is set as 10.

It can be seen from Table 2 and Table 3 that MLRS performs well on most evaluation criteria when it applied to the multi-label classification problem. With the enormous increasing of the amount of instances and labels, MLRS still can performs well compared to other multi-label algorithms. It shows that MLRS has some scalability.

Table 1. Multi-label datasets used for experiments

name	instances	attribute	labels	cardinality	density
Scene	2407	294	6	1.074	0.179
Corel5k	5000	499	374	3.522	0.009

Table 2. MLNRS vs. other multi-label algorithms over *Scene*

performance	RAkEL	BPMLL	CC	MLRS
hloss	0.1012±0.0075	0.2667±0.0508	0.1444±0.0164	**0.0912±0.0082**
avgprec	0.8379±0.0156	0.6852±0.0235	0.7176±0.0354	**0.8652±0.0153**
cov	0.5862±0.0593	0.9405±0.0855	1.3504±0.2002	**0.4818±0.0539**
one-error	0.2663±0.0258	0.5450±0.0381	0.3914±0.0453	**0.2255±0.0248**
rloss	0.0999±0.0121	0.1714±0.0165	0.3914±0.0453	**0.0790±0.0116**

Table 3. MLNRS vs. other multi-label algorithms over *Corel5k*

performance	RAkEL	BPMLL	CC	MLRS
hloss	**0.0097±0.0001**	0.5547±0.0213	0.0099±0.0001	0.0105±0.0001
avgprec	0.1075±0.0080	0.0563±0.0097	0.2364±0.0102	**0.2463±0.0092**
cov	336.0374±2.6687	169.0732±4.6338	165.3946±5.8193	**132.1238±5.4093**
one-error	0.7734±0.0201	0.9974±0.0025	**0.7076±0.0172**	0.7398±0.0154
rloss	0.6565±0.0116	0.2273±0.0096	0.1869±0.0083	**0.1513±0.0049**

5 Conclusion

We study the problem of classification under multi-label dataset in this paper. Based on rough set theory, we propose two kinds of approaches to deal with the multi-label problem and present a multi-label rough set model. After applying the model to multi-label datasets, we obtain promising results compared with other well-known multi-label algorithms. Future work will focus on the dimension reduction of multi-label dataset which can improve the accuracy and efficiency of prediction.

References

1. Tsoumakas, G., Katakis, I.: Multi-label classification: An overview. International Journal of Data Warehousing and Mining (IJDWM) 3, 1–13 (2007)
2. Pawlak, Z.: Rough sets. International Journal of Computer and Information Sciences 11, 341–356 (1982)
3. Kryszkiewicz, M.: Rough set approach to incomplete information systems. Information Sciences 112, 39–49 (1998)
4. Stefanowski, J., Tsoukias, A.: Incomplete information tables and rough classification. Computational Intelligence 17, 545–566 (2001)
5. Wang, G.: Extension of rough set under incomplete information systems. Journal of Computer Research and Development 10, 1–9 (2002)
6. Yao, Y.Y.: Relational interpretations of neighborhood operators and rough set approximation operators. Information Sciences 111, 239–259 (1998)
7. Hu, Q., Yu, D., Xie, Z.: Neighborhood classifiers. Expert Systems with Applications 34, 866–876 (2008)
8. Hllermeier, E., Frnkranz, J., Cheng, W., Brinker, K.: Label ranking by learning pairwise preferences. Artificial Intelligence 172, 1897–1916 (2008)
9. 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)
10. Schapire, R.E., Singer, Y.: BoosTexter: A boosting-based system for text categorization. Machine Learning 39, 135–168 (2000)
11. De Comit, F., Gilleron, R., Tommasi, M.: Learning multi-label alternating decision trees from texts and data. In: Perner, P., Rosenfeld, A. (eds.) MLDM 2003. LNCS, vol. 2734, pp. 251–274. Springer, Heidelberg (2003)
12. Zhang, M.L., Zhou, Z.H.: ML-KNN: A lazy learning approach to multi-label learning. Pattern Recognition 40, 2038–2048 (2007)
13. Meng, Z., Shi, Z.: Extended rough set-based attribute reduction in inconsistent incomplete decision systems. Information Sciences 204, 44–69 (2012)
14. Yao, Y.Y., Wong, S.K.M.: A decision theoretic framework for approximating concepts. International Journal of Man-Machine Studies 37(6), 793–809 (1992)
15. Yao, Y.Y., Wong, S.K.M., Lingras, P.: A decision-theoretic rough set model, Methodologies for Intelligent Systems. In: Ras, Z.W., Zemankova, M., Emrichm, M.L. (eds.) Methodologies for Intelligent Systems, vol. 5, pp. 17–25. North-Holland, New York (1990)