VNU Journal of Science: Comp. Science & Com. Eng., Vol. 35, No. 2 (2019) 1-7

# A new feature reduction algorithm based on fuzzy rough relation for the multi-label classification

Pham Thanh Huyen[1], Ho Thuan[2]

[1]*University of Engineering and Technology, Vietnam National University, Hanoi, Vietnam*

[2]*VietNam Academy of Science and Technology, Hanoi, Vietnam*

**Abstract**

The paper aims at presenting an algorithm for reducing number of features in classification problem. The classification system bases on fuzzy rough relation with multi-label classification. According to the determination of the dependency, more significant features will be retained in the reduction set. The paper focuses on proposing a new method of reducing label-specific features using choosing the most significance features which have the highest dependence in a given fuzzy set.

Received ..... October 2019, revised ....., accepted .....

*Keywords:* Fuzzy rough relation, lable-specific feature, feature reduction set.

## 1. Introduction

Combining fuzzy set theory and rough set theory to apply to data classification has been studied relatively [7, 11], etc. especially the multi-label classification [6] and the reduction of feature space [2]. Fuzzy rough set theory is a tool that allows the implementation of fuzzy approximations of the clear approximation spaces [8]. It is proven effective in diverse data exploitation for classification [5, 7, 10, 11] etc.

Nowadays, the increase in the number of feature dimensions and the excess of received information during the data collection process is one of the most concerned issues. There are many characteristics that are difficult to distinguish and need to be removed. Because they can reduce efficiency in multi-label training. FRS-LIFT and FRS-SS-LIFT [9] are multi-label training algorithms with a distinct label feature reduction that uses approximation to evaluate specific dimension. Based on feature reduction results, classification efficiency has been enhanced. Xu et al. [9] have performed to find a reduction feature set by calculating the dependency of each feature on the decision set at each given label and evaluating the approximate change of that feature set while adding or removing any feature in the original feature space. According to [9], the selection of features for reduction is randomly selected. Although FRS-LIFT improves the performance of multi-label learning via reducing redundant label-specific feature dimensionalities, its computational complexity is high.

Our paper focuses on the fuzzy rough relation to calculate the approximate

[2] Corresponding author. E-mail.: hothuan1812@yahoo.com

dependence between samples on each feature, selecting the purpose-based feature with greatest dependence to include in the reduction set. We propose a new algorithm (called FRR-MLL, Fuzzy Rough Relationship Multi-Label Learning) to improve the LIFT [4] using reducing the feature space. We calculated the degree of the membership function for each element $x$ in universe $\mathcal{X}$ and improved a new methodical reduction via review per feature which has the highest dependence before classification. In fact, we based on the greatest dependency on each feature to select the more dominant feature into the feature reduction set. Thereby, leading to a reduced set with a given threshold.

The article consists of 5 parts. The next section introduces the multi-label training method, LIFT method, the fuzzy rough relationship, FRS-LIFT method and the factors related to feature reduction. Part 3 introduce about the label-specific feature reduction. Part 4 express a proposed algorithm. Finally, part 5 concludes and discusses some plans to develop in the future.

## 2. Related work

### *2.1 Multi-Label trainning*

Multi-label training stated [3]:

Let $\mathcal{X} = \mathbb{R}^d$ be a sample space, $\mathcal{L}$ is a finite set of $q$ labels $\mathcal{L} = \{l_1, l_2, \ldots, l_q\}$.

$\mathcal{T} = \{(x_i, Y_i) | i = 1, 2, \ldots, n\}$ be multi-label training set with $n$ samples with $\forall x_i \in \mathcal{X}$ is $d$-dimensional feature vector,

$x_i = [x_i^1, x_i^2, \ldots, x_i^d]$ and $Y_i \subseteq \mathcal{L}$ be the set of labels associated with $x_i$. The desired purpose is that the training system will create a real-valued function $f: \mathcal{X} \times P(\mathcal{L}) \to \mathbb{R}$; where $P(\mathcal{L})$ is a finite set with $\forall Y_i \subseteq P(\mathcal{L})$. $P(\mathcal{L})$ is the set of the label sets $Y_i$ that connect to $x_i$.

The problem of multi-label classification is also shown in the text semi-supervised multi-label learning model [6]:

Let $D$ be the set of documents in a considered domain. Let $L = \{l_1, \ldots, l_q\}$ be the set of labels. Let $\overline{D}$ and $\overline{D}^U$ be the collections of labeled and unlabeled documents, correspondingly. For each $d$ in $\overline{D}$, $label(d)$ denotes the set of labels assigned to $d$. The task is to derive a multi-label classification function $f: D \to 2^L$, i.e, given a new unlabeled document $d \in D$, the function identifies a set of relevant labels $f(d) \subseteq L$.

### *2.2. Approach to LIFT*

As can be seen in [4], to train multi-label learning successfully, approach to LIFT perform three steps. First, creating label-specific features for each $l_k \in \mathcal{L}$ label which is done by dividing the $\mathcal{T}$ training into two sample sets:

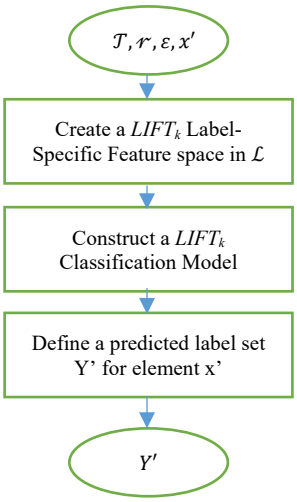
$$P_k = \{x_i | (x_i, Y_i) \in \mathcal{T}, l_k \in Y_i\};$$
$$N_k = \{x_i | (x_i, Y_i) \in \mathcal{T}, l_k \notin Y_i\}; \quad (1)$$

($P_k, N_k$ are called two positive and negative training samples for each $l_k$ label, respectively.)

Then, perform *k-means* clustering, dividing $P_k, N_k$ into discrete clusters with the clustering centers are respectively $\{p_1^k, p_2^k, \ldots, p_{m_k^+}^k\}$ and $\{n_1^k, n_2^k, \ldots, p_{m_k^-}^k\}$, in which:

$$m_k^+ = m_k^- = m_k = \lceil r.min(|P_k|, |N_k|) \rceil \quad (2)$$

($m_k^+, m_k^-$ are respectively the value of cluster numbers divided in $P_k, N_k$; $r$ is the ratio parameter controlling the number of given clusters).




*Flowchart 1: A $LIFT_k$ Classification Model*

Creating the label-specific feature space $LIFT_k$ with $2.m_k$ dimension bases on Euclidean metric to compute distance between samples.

$$\varphi_k: \mathcal{X} \rightarrow LIFT_k \quad (3)$$
$$\varphi_k(x_i) = [d(x_i, p_1^k), \dots, d(x_i, p_{m_k}^k), d(x_i, n_1^k), \dots, d(x_i, n_{m_k}^k)]$$

Second, build a family of $q$ classification models $LIFT_k$ $(1 \le k \le q)$ be $\{f_1, f_2, \dots, f_q\}$ respectively for $l_k \in \mathcal{L}$ labels. In which, a binary training set is created in the form of:

$$\mathcal{B}_k = \{(\varphi_k(x_i), \omega(Y_i, l_k)) | (x_i, Y_i) \in \mathcal{T}\} \quad (4)$$

$(\omega(Y_i, l_k) = 1$ if $l_k \in Y_i$, $\omega(Y_i, l_k) = -1$ if $l_k \notin Y_i)$

Initialize the classification model for each label based on $\mathcal{B}_k$ as follows: $f_k: LIFT_k \rightarrow \mathbb{R}$

Finally, define the predicted label set for $x \in \mathcal{X}$ sample:

$$Y = \{l_k | f(\varphi_k(x), l_k) > 0, 1 \le k \le q\}$$

### *2.3 Approach to fuzzy-rough relation*

Let a nonempty universe $\mathcal{X}$, $R$ is a similarity relation on $\mathcal{X}$ where every $x \in \mathcal{X}$, $[x]_R$ stands for the similarity class of $R$ the represent $x$, i.e. $[x]_R = \{y \in \mathcal{X}: (x, y) \in R\}$.

Given $A$ be the set of condition features, $D$ be the set of decision feature and $F$ be a fuzzy set on $\mathcal{X}$ [5] ($F: \mathcal{X} \rightarrow [0,1]$). A fuzzy rough set is the pair of lower and upper approximations of $F$ in $\mathcal{X}$ on a fuzzy relation $R$.

The fuzzy – rough relation is built such as [5, 8], the fuzzy similarity between two patterns $x$ and $y$ on the feature $a \in A$ is determined:

$$R_a(x, y) = 1 - \frac{|a(x) - a(y)|}{\max\limits_{i=1 \div n} a(z_i) - \min\limits_{i=1 \div n} a(z_i)} \quad (5)$$

Then, the fuzzy similarity relation among all samples in $\mathcal{X}$ on the reductant $B$ in each individual label $l_k$ is determined $\forall x, y \in \mathcal{X}$, $B \subseteq A$:

$$R_B(x, y) = \min_{a \in B}\{R_a(x, y)\} = \min_{a \in B}\left\{1 - \frac{|a(x) - a(y)|}{\max\limits_{i=1 \div n} a(z_i) - \min\limits_{i=1 \div n} a(z_i)}\right\} \quad (6)$$

The relationship $R_B(x, y)$ is the fuzzy similarity relation that satisfies to be reflexive, symmetrical and transitive [8, 11].

Determining the approximations of each fuzzy similarity relation with the corresponding decision set $D_k$ in the label $l_k$, respectively.

$$\underline{R_B D}(x) = \inf_{y \in X} max(1 - R(x, y), F(y));$$
$$\overline{R_B D}(x) = \sup_{y \in X} min\,(R(x, y), F(y)); \quad (7)$$

There may be the method to determine the approximation of $B$ for $D_k$ as follows in Eq. (8):

$$\underline{R_B D}(x) = \inf_{y \in X} max\left(1 - \min_{a \in B}\left\{1 - \frac{|a(x) - a(y)|}{\max\limits_{i=1 \div n} a(z_i) - \min\limits_{i=1 \div n} a(z_i)}\right\}, F(y)\right) \quad (8)$$

The approximate cardinality represents the dependence of the feature set $B$ on $D_k$ in the form [1, 11]:

$$\gamma(B, D) = \frac{\sum_{x \in \mathcal{X}} POS_B(D)}{|\mathcal{X}|} \quad (9)$$

In which, $|\mathcal{X}|$ determine the cardinality of the set. And $POS_B(D) = \bigcup\limits_{x \in \mathcal{X}/D} \underline{R_B D}(x)$, where $POS_B(D)$ is the definite area of the partition $\mathcal{X}/D$ with $B$. In fact, $0 \le \gamma(B, D_k) \le 1$, its meaning is to represent the proportion of all elements of $\mathcal{X}$ which can be uniquely classified $\mathcal{X}/D$ using features $B$. Moreover, the dependency $\gamma(B, D_k)$ is always defined on the

fuzzy equivalence approximation values of all finite samples.

$B$ is the best reducted feature set in $A$ if $B$ be satisfied simultaneously:

$\forall B \subseteq A, \gamma(A, D_k) > \gamma(B, D_k)$ và

$\forall B' \subseteq B, \gamma(B', D_k) < \gamma(B, D_k)$ (10)

Using threshold ε without restrictions [9], $B$ is the reduction of the set $A$ if satisfied:

$(i) \quad \gamma(A, D) - \gamma(B, D) \leq \varepsilon$

$(ii) \; \forall C \subset B, \gamma(A, D) - \gamma(C, D) > \varepsilon$ (11)

The threshold parameter $\varepsilon$ perform a role in controlling the change the approximation quality to loosen the limitations of reduction. The purpose of the using $\varepsilon$ is to reduce redundant information as much as possible [9].

*2.4 A FRS-LIFT multi-label learning approach*

FRS-LIFT is a multi-label learning approach with label-specific feature reduction based on fuzzy rough set [9]. To define the membership functions of the fuzzy lower and upper approximations, Xu et al firstly used a fuzzy set $F$ following in [1]. Then, they carried out calculating the approximation quality to review the significance of specific dimension via perform the forward greedy search strategy. They select the most significant features until no more deterministic rules generating with the increasing of features. There are two determined coefficients to identify the significance of a considered feature in the predictable reduction set $B$ in which: $\forall a_i \in B, B \subseteq A$:

$Sig_{in}(a_i, B, D) = \gamma(B, D) - \gamma(B - \{a_i\}, D)$ (12)

$Sig_{out}(a_i, B, D) = \gamma(B + \{a_i\}, D) - \gamma(B, D)$ (13)

where $Sig_{in}(a_i, B, D)$ means the significance of $a_i$ in $B$ relative to $D$, and $Sig_{out}(a_i, B, D)$ measures the change of approximate quality when $a_i$ is chosen into $B$.

This algorithm improves the performance of multi-label learning using reducing redundant label-specific feature dimensionalities. However, its computational complexity is high. FRS-SS-LIFT is also be limited time and memory consumption.

## 3. The label-specific feature reduction for classification model

*3.1. Problem Formulation*

According to LIFT [4], the label-specific space has an expanded dimension 2 times greater than the number of created clusters. In which, the sample space contains:

$$A = \{a_1, a_2, .., a_{2m_k}\} = \{p_1^k, p_2^k, \dots, p_{m_k}^k, n_1^k, n_2^k, \dots, n_{m_k}^k\}$$

be the feature sets in $\mathcal{X}$.

$\forall x_i \in \mathcal{X}, i = 1 \div n$ be the feature vector,

$x_i = [x_i^1, \dots, x_i^{2m_k}]$, each $x_i^j$ be a distance $d(x_i, p_j^k)$.

$D_k = [d_k^1, d_k^2, \dots, d_k^n]$ be the decided classification,

$d_k^j = 1$ if $x_i \in l_k$; $d_k^j = 0$ if $x_i \notin l_k$;

Thus, when we have the multi-label training set $\mathcal{T}$ and the necessary input parameters, the obtained result is a predicted label set $Y$ for any sample $x$. In order to be able to have an effective set $Y$, it is necessary to solve the label-specific feature reduction [9]. Therefore, our main goal is to build a classification model that represents the mapping form: $\mathcal{F}: \mathcal{X} \rightarrow FRR\text{-}MLL_k$

This proposed task is to build the feature reduction space $FRR\text{-}MLL_k$ based on the properties of the fuzzy rough relation to satisfy:

- Selecting a better fuzzy set for determining the degree of the membership function of approximates.
- The feature $a_i$ which has the highest dependency $\gamma(a_i, D_k)$ is chosen into the reduced feature set $B$ in this space ($B \subseteq A$) on $D_k$. This work is performed if $B$ satisfy Eq. 11 and $\gamma(A, D) - \gamma(B, D)$ obtains the great value with the threshold parameter $\varepsilon \in [0, 0.1]$.

*3.2. Reducing the feature set for multi-label classification*

In this subsection, we propose the reductive feature set $B$ be satisfied simultaneously: The dependency of the feature which is added into reduction set $B$ on the partition $\mathcal{X}/D$, $\gamma(a_i, D)$ is greatest.

The dependency difference between the initial feature in the set $A$ with $D_k$ and the dependency between the reduced feature set $B$ with $D_k$ must be within the given threshold ε ($\varepsilon \in [0,0.1]$), et., $\gamma(A, D_k) - \gamma(B, D_k) \leq \varepsilon$;

We focused on selecting the proposed feature into the reduction set $B$ and conducted experimentally on many datasets:

- The feature that has the greatest dependency and was determined from the fuzzy approximations on the samples, is first selected to be included in the set $B$.

- Next, other features are considered to be included in the reduction set $B$ if guaranteed using threshold ε without restrictions [9] i.e, $B$ is the reduction of the set $A$ if satisfied Eq. (11).

We note that finding a good fuzzy set is more meaningful for identification between elements. It directly effects the result of the membership function of approximates. In fact, searching a great fuzzy set to model concepts can be challenge and subjective, but it is more significance than make an artificial crisp distinction between elements [5]. Here, we temporality based on the cardinality of a fuzzy set $F$ to determine the sum of the membership values of all elements in $\mathcal{X}$ to $F$.

For example: Given the set $\mathcal{X}$ by the under table and the dependency parameter $\varepsilon = 0.1$, respectively determine the fuzzy equivalence relationship $R_A(x, y)$ and the lower approximations of the features with $D_k$. Then, calculate the dependencies $\gamma(A, D_k)$ and $\gamma(a_i, D_k)$:

| $\mathcal{X}$ | $a_1$ | $a_2$ | $a_3$ | $a_4$ | $d_k$ |
|---|---|---|---|---|---|
| $x_1$ | 3.3 | 2.0 | 3.0 | 4.2 | 1 |
| $x_2$ | 1.1 | 3.8 | 1.7 | 2.3 | 1 |
| $x_3$ | 2.0 | 4.7 | 2.1 | 2.5 | 0 |
| $x_4$ | 2.9 | 4.2 | 2.9 | 1.8 | 0 |
| $x_5$ | 1.9 | 2.5 | 1.7 | 2.9 | 0 |
| $x_6$ | 2.4 | 1.7 | 2.3 | 3.1 | 1 |
| $x_7$ | 2.5 | 3.9 | 2.3 | 1.6 | 0 |

$\gamma(A, D_k) = 0.25,$
$\gamma(a_1, D_k) = 0.092$
$\gamma(a_2, D_k) = 0.07$
$\gamma(a_3, D_k) = 0$
$\gamma(a_4, D_k) = 0.094$

First, we choose the feature *a4* and add it to the set $B$. Next, we select the feature *a1* and add it to the set $B$. Calculate $\gamma(B, D) = 0.15$, we obtained: $\gamma(A, D) - \gamma(B, D) = \varepsilon$

So, $B = \{a1, a4\}$ is the obtained reduced feature set in the threshold ε. If this threshold is adjusted $\varepsilon = 0.08$ then $\gamma(B \cup \{a_2\}, D) = 0.19$. We add the feature *a2* to the reductive set $B$ that satisfies the formula (11).

## 4. The proposed algorithms

*4.1. The specific feature reduction algorithm*

Finding the optimal reductive set from the given set $A$ is seen as the significant phase. It is necessary to decide the classification efficiency. So, we propose a new method FRR_RED to search an optimal set.

**Algorithm 1:** *FRR-RED algorithm*

**Inputs**: The finite set of $n$ samples $\mathcal{X}$; The set of condition features $A$; The set of decision $D$; The threshold $\varepsilon$ for controlling the change of approximate quality.

$\mathcal{X} = \{x_1, \ldots, x_n\},$
$A = \{a_1, \ldots a_{2*m}\}, D = \{d_1, \ldots d_n\};$

**Output**: Feature reduction $B$.

**Method:**

1. $\forall x_i \in \mathcal{X}$ compute 2*$m$ the fuzzy equivalent relations between each sample on according to Eq. (5);

2. Compute $\gamma(A,D)$ and $\gamma_i = \gamma(a_i,D)\ \forall a_i \in A$ according to Eq. (9);

3. Create $B$ = {}; $\gamma(B,D) = 0$;

4. **For each** $a_j \in A$

5. **If** $(\gamma(A,D) - \gamma(B,D) > \varepsilon)$ **then**

6. Compute $\gamma_{max}$ thỏa mãn $\forall a_i \in A$ và $\forall a_i \notin B$

7. **If** $(\gamma_{a_j} = \gamma_{max})$ **then** $B = B \cup \{a_j\}$;

8. Compute $\gamma(B,D)$ according Eq. (9);

**9. End if**

10. **End if**

11. **End for**

From step 4 to step 11, selecting the features that have the highest dependency to put into the reductive set $B$ is implemented continuously until satisfy Eq. (11). This proposed method which hopefully finds the optimal reductive set is different to the previous approach because this selecting process is not random.

*4.2. Approach to FRR_MLL for multi-label classification with FRR_RED*

**Algorithm 2:** *FRR-MLL algorithm*

**Inputs**: The multi-label training set $\mathcal{T}$, The ratio parameter $r$ for controlling the number of clusters; The threshold $\varepsilon$ for controlling the change of approximate quality; The unseen sample $x'$.

**Output**: The predicted label set $Y'$.

**Method:**

**1. For $k$ = 1 to $q$ do**

2. Form the set of positive samples $\mathcal{P}_k$ and the set of negative samples $\mathcal{N}_k$ based on $\mathcal{T}$ according to Eq. (1);

3. Perform $k$-means clustering on $\mathcal{P}_k$ and $\mathcal{N}_k$, each with $m_k$ clusters as defined in Eq. (2);

4. $\forall (x_i, Y_i) \in \mathcal{T}$, create the mapping $\varphi_k(x_i)$ according to Eq. (3), form the original label-specific feature space $LIFT_k$ for label $l_k$;

5. **Perform find decision redure B such as FRR-RED**;

6. With $B$, form the dimension-reduced label-specific feature space *FRR-MLL*$_k$ for label $l_k$ (etc., mapping $\varphi'_k(x_i)$);

**7. End for**

**8. For $k$ = 1 to $q$ do**

9. Construct the binary training set $\mathcal{T}_k^*$ in $\varphi'_k(x_i)$ according to Eq. (4);

10. Induce the classification model $f_k$: $FRR - MLL_k \rightarrow \mathbb{R}$ by invoking any binary learner on $\mathcal{T}_k^*$;

**11. End for**

12. The predicted label set:

$$Y = \{l_k | f(\varphi'_k(x_i)) > 0,\ 1 \le k \le q\}$$

The FRR-MLL algorithm is performed to create the $FRR - LIFT_k$ space, then reduce the label-specific feature based on selecting the maximum dependency of the features. The dataset on the reductive feature set is trained in the next step. Finally, build the classification model *FRR_MLL*$_k$ and make the label prediction set $Y$ for the element $x'$.

**Conclusion**

The paper proposed two algorithm for reducing the set of features. Finding the most significance features can determine the new reduction set rapidly, because we have not to

calculate all most features if the reduction set satisfy all conditions. In the future, we continue to conduct experiments on real databases to evaluate the proposed algorithms and improve the fuzzy set $F$ which is the set of the membership functions on $\mathcal{X}$.

## References

[1] D. Dubois, H. Prade, Rough fuzzy sets and fuzzy rough sets, Int. J. Gen. Syst. 17 (1990) 191 – 209.

[2] Daniel Kostrzewa, Robert Brzeski, The data Dimensionality Reduction and Feature Weighting in the Classification Process Using Forest Optimization Algorithm, ACIIDS 2019, pp 97 – 108.

[3] M.L. Zhang, Z.H. Zhou, ML-KNN: A lazy learning approach to multi-label learning, Pattern Recognition. 40 (2007) 2038-2048.

[4] M.L. Zhang, LIFT: Multi-label learning with label-specific features, IEEE Trans, Pattern Anal, Mach, Intell 37 (2015), pp 107-120.

[5] Nele Verbiest (2014). Fuzzy Rough and Evolutionary Approaches to Instance Selection. *PhD Thesis,* Ghent University.

[6] Quang-Thuy Ha, Thi-Ngan Pham, Van-Quang Nguyen, Minh-Chau Nguyen, Thanh-Huyen Pham, Tri-Thanh Nguyen, A New Text Semi-supervised Multi-label Learning Model Based on Using the Label-Feature Relations, International Conference on Computational Collective Intelligence, LNAI 11055, Springer, 2018, pp. 403-413.

[7] Richard Jensen, Chris Cornelis (2011). Fuzzy-Rough Nearest Neighbor Classification and Prediction. Proceedings of the 6th International Conference on Rough Sets and Current Trends in Computing., pp. 310-319.

[8] Richard Jensen, Neil Mac Parthaláin and Qiang Shen (2014). Fuzzy-rough data mining (using the Weka data mining suite). A Tutorial, IEEE WCCI 2014, July 6, 2014, Beijing, China.

[9] Suping Xu, Xibei Yang, Hualong Yu, Dong-Jun Yu, Jingyu Yang, Eric CC Tsang, Multi-label learning with label-specific feature reduction, Knowledge-Based Systems 104 (2016), pp 52-61.

[10] Y. Yu, W. Pedrycz, D. Q. Miao, Multi-label classification by exploiting label correlations, Expert syst. Appl. 41 (2014) 2989 – 3004.

[11] Y.H. Qian, Q. Wang, H.H. Cheng, J. Y. Liang, C.Y. Dang, Fuzzy-Rough feature selection accelerator, Fuzzy Sets Syst. 258 (2014) 61 – 78.