Multi-label Online Streaming Feature Selection Based on Spectral Granulation and Mutual Information

Huaming Wang, Dongming Yu, Yuan Li, Zhixing Li(✉), and Guoyin Wang(✉)

Chongqing Key Laboratory of Computational Intelligence,
Chongqing University of Posts and Telecommunications,
Chongqing 400065, People's Republic of China
{lizx,wanggy}@cqupt.edu.cn

Abstract. Instances in multi-label data sets are generally described as a high-dimensional feature vector, as brings the "curse of dimensionality" problem. To ease this problem, some multi-label feature selection algorithms have been proposed. However, they all handle feature selection problems with the assumption that all candidate features are available beforehand. While in some real applications, feature selection must be conducted in the online manner with dynamic features, for example, novel topics arise constantly with a set of features in social networks. Online streaming feature selection (OSFS), dealing with dynamic features, has attracted intensive interest in recent years. Some online feature selection methods are designed for single-label applications, They can not be directly applied in multi-label scenarios. In this paper, we propose a multi-label online streaming feature selection algorithm based on spectral granulation and mutual information (ML-OSMI), which takes high-order label correlations into consideration. Moreover, comprehensive experiments are conducted to verify the effectiveness of the proposed algorithm on twelve multi-label high-dimensional benchmark data sets.

Keywords: Multi-label feature selection · Streaming features
Mutual information · Granular computing

1 Introduction

Multi-label data emerge on various real-world domains, such as image processing, text classification, bioinformatics and information retrieval [1–5]. In these applications, each instance is associated with multiple labels simultaneously. For example, a document may belong to many topics and a gene could have several functions [5]. Moreover, multi-label data are generally represented by very high dimensional vectors, as brings a large number of features and most of them are irrelevant or redundant [6]. Unnecessary features may not only reduce the performance of classifiers but result in the increment of memory storage and computation time. To ease these problems, feature selection techniques have been wildly

H. S. Nguyen et al. (Eds.): IJCRS 2018, LNAI 11103, pp. 215–228, 2018.
https://doi.org/10.1007/978-3-319-99368-3_17

studied, which select a relative small subset of features from the original feature space to remove irrelevant and redundant features without losing discriminative information for later processing.

A number of feature selection methods dealing with multi-label data have been proposed [7–9]. However, they handle feature selection problems with the assumption that all candidate features are available before the learning starts and have to wait for the calculation of all the features, which is very deficient in practice. Online streaming feature selection [10], evaluating features dynamically with the arrival of new features, is a more time efficiency and intuitive way to solve such problems. Existing online feature selection methods [11–14]. But they are designed for single-label learning tasks and cannot be directly applied to multi-label tasks. One commonly encountered way is transforming the multi-label problems into single-label problems. Then single-label online feature selection methods can be adopted. Nevertheless, it ignores the correlation among labels which may carry useful information for learning task, or leads to extremely high and unbalanced label space [9,15].

In this paper, we analyze multi-label online streaming feature selection problem and design an online streaming feature selection algorithm based on spectral granulation and mutual information. The proposed algorithm first granulates labels using spectral clustering. Then it transforms label granules into new multi-class labels and performs feature selection on the new label space. The main contributions of this study are summarized as follows: (1) Although there are multi-label feature selection methods for constant features and single-label feature selection algorithms for dynamic features, we introduce dynamic feature selection into multi-label scenarios. (2) We designed a novel multi-label online streaming feature selection algorithm. (3) Comprehensive experiments are conducted to compare our proposed methods with traditional multi-label methods and single-label online streaming feature selection algorithms on various benchmark multi-label data sets.

2 Related Works

2.1 Multi-label Feature Selection

In multi-label learning tasks, each instance is associated with multiple labels and these labels are generally correlated, as makes multi-label feature selection tasks more complicated than single-label ones. Moreover, there are evidences showing that taking label correlations into consideration can benefit the learning model [7]. Hence, exploring label dependence is an important issue. Multi-label feature selection algorithms can be divided into three categorizes by the type of correlations they considered, first-order, second-order and high-order methods.

First-order ones, such as BR [15], consider each label independently and transform the multi-label feature selection task into several binary single-label sub-problems. LCFS [16] is a second-order algorithm. It builds new labels based on relations among the original labels to capture pair-wise label correlations

and then conducts BR approach on the expanded label space to select a subset of informative features. First-order and second-order algorithms assume that labels are independent to each other or pair-wise correlated. However, correlations among labels in real applications are more complicated. LP transforms multi-label data set to a new single-label multi-class data set, then any single-label feature selection could be adopted [15]. However, when the number of labels is extreme large, LP based methods could suffer from terribly class-imbalance problems [6].

MDMR [17] defines mutual information based evaluations to guide feature selection procedure, considering multi-label feature selection problems in two aspects, namely feature dependency and feature redundancy. [18] implements a multi-label feature selection method similar to MDMR named MLMRMR based on the single-label feature selection algorithm mRMR [19]. [9] partitions labels into clusters according to their similarity using a balanced k-means methods and then undertakes feature selection based on mRMR viewing each cluster of labels as a new multi-label subtask. RFS [20] introduces $\ell_{2,1}$-norm on both loss function and regularization to eliminate unnecessary features. [21] solves multi-label feature selection with streaming labels by ranking features iteratively, where the labels arrive one at a time. [7] proposes a multi-label feature selection method called MIFS. The labels are first mapped to a low-dimensional space with less noisy. Then it conducts feature selection on the reduced label space.

2.2 Online Streaming Feature Selection

Online streaming feature selection focuses on the feature selection problems with dynamic features. Grafting [13], Alpha-investing [14], fast-OSFS [11] and SAOLA [22] are several state-of-the-art algorithms proposed to solve online streaming feature selection problems. Grafting treats the feature selection task as a stream-wise regularized risk minimization problem. New features are selected if the improvement of accuracy made by them is greater than a predefined threshold. However, it has no mechanism to remove redundant features selected previously, rendering it suffering from the nesting effect. Alpha-investing [14] uses a step-wise linear regression model and a p-value to determine new features which are selected or not. Furthermore, alpha-investing and Grafting used prior information about the structure of feature space, which is impossible to obtain on the original streaming tasks. Hence, they might not produce good performance in real applications. Wu [11] proposed the fast-OSFS algorithm, needing no prior knowledge about the feature space, which contains two major steps: online relevance analysis and online redundancy analysis. The first step discards irrelevant features and the second eliminates redundant features. SAOLA [22] is another online feature selection method dealing with dynamic features using mutual information based criterions to guide feature selection heuristically.

Though there are several online feature selection methods proposed, they are designed for single-label tasks and can not apply directly in multi-label scenarios. In this paper, we study the multi-label feature selection problems with dynamic

(or streaming) features and propose a multi-label online streaming feature selection algorithm.

3 The Proposed Method

In this section, we first describe the multi-label online streaming feature selection problem. Then, we design a multi-label online streaming feature selection method. The proposed method applies spectral clustering which granulates labels into clusters and captures high-order label correlations. Moreover, the relevance and redundancy of features are redefined using mutual information to guide multi-label feature selection procedure.

3.1 Problem Statement

Definition 1 (Traditional Multi-label Feature Selection). Let $\boldsymbol{X}$ be the sample space and $\boldsymbol{x_i} \in \boldsymbol{X}$ is a feature vector. $\boldsymbol{Y} = \{\boldsymbol{l_1}, \boldsymbol{l_2}, ..., \boldsymbol{l_m}\}$ is a set of labels. Multi-label learning is objective to produce a function $H = \{X \rightarrow 2^L\}$ which assigns each instance with a set of relevant labels. Traditional multi-label feature selection holds the assumption that instances are represented with a fixed dimensional feature space $\boldsymbol{F} = \{\boldsymbol{f_1}, \boldsymbol{f_2}, ..., \boldsymbol{f_d}\}$. They aim to select an optimal subset of features $\boldsymbol{SF} \subseteq \boldsymbol{F}$ without harming the predictive performance.

Definition 2 (Streaming Features). Streaming features denote a feature space where features flow in one by one over time with fixed number of instances. With a dynamic feature space, the dimensionality may tend to very high or even infinite. Besides, each feature is required to be processed when its arrival. Hence, feature selection procedure should be conducted in the online manner.

Definition 3 (Multi-label Online Streaming Feature Selection). Multi-label online streaming feature selection copes with a streaming feature vector $\boldsymbol{F_s^t}$, where $\boldsymbol{F_s^t} = \{\boldsymbol{f_1}, \boldsymbol{f_2}, ..., \boldsymbol{f_t}\}$ and $\boldsymbol{f_t}$ denotes the feature arrives at time $\boldsymbol{t}$. As the features flow in continuously, multi-label streaming feature selection task is objective to remove irrelevant and redundant features from the available feature set $\boldsymbol{F_s^t}$ while holds discriminative information with more than one targets $\boldsymbol{Y} = \{\boldsymbol{l_1}, \boldsymbol{l_2}, ..., \boldsymbol{l_m}\}$.

There are three major challenges in the multi-label streaming feature selection scenario:

- **The dynamic and uncertain nature of the feature space.** The dimensionality of the feature space grows over time and may even tend to infinite.
- **The streaming nature of the feature space.** The subset of selected features should be updated timely with new features flow in one at a time.
- **The complex correlations among labels.** There are complex correlations among labels and evidences show that taking label correlations into consideration will benefit learning model.

3.2 The Framework of ML-OSMI

The framework of the proposed multi-label online streaming feature selection algorithm is shown in Fig. 1. To capture label correlations, the original label space is first transformed into a multi-class multi-target one with much lower dimensionality. Then, the new labels are used to select features. To conduct feature selection procedure with many labels and streaming features, we adopt relevance test and redundancy test to guide the online feature selection, motivated by single-label online streaming feature selection methods [11]. Section 3.3 gives the details of label space transformation and Sect. 3.4 redefines the relevance and redundancy of features.

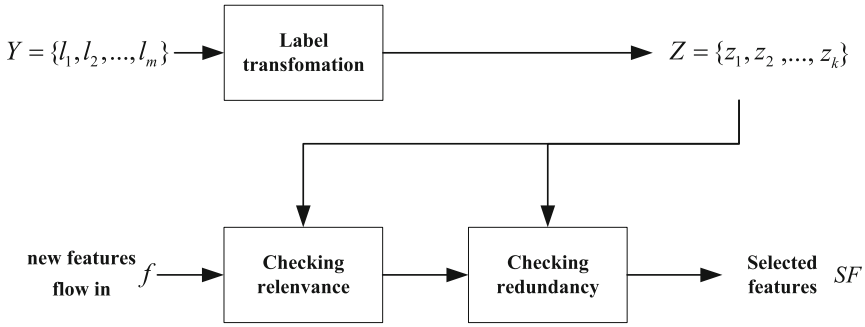

Fig. 1. Framework of the proposed algorithm

3.3 Capturing Label Correlations by Spectral Granulation

In multi-label data, a label is generally related to a small set of labels from the entire label space [9,23]. Hence, the label correlations can be explored as much as possible by dividing labels into partitions, where the labels in one partition are relevant to each other and the labels in different partitions are irrelevant. The partitions of labels are considered as granulas in this paper. The labels in the same granula are high correlated while the labels in different granula are mutually independent or weakly related. To generate the granulas, labels are clustered using spectral clustering with cosine similarity. Then, each label clusters is transformed into a multi-class label applying LP framework [6]. Finally, we get a new label space consists of multi-class labels with much lower dimensionality than the original label space. The new multi-class labels are used to steer feature selection processing taking label correlations into account.

3.4 Evaluations Based on Mutual Information

To perform multi-label feature selection, an algorithm must be able to measure the dependency between features and labels. Mutual information is often

employed to characterize this dependency. Given two random variables $\boldsymbol{x}$ and $\boldsymbol{y}$, their mutual information is defined in terms of probability density functions $p(x)$, $p(y)$ and $p(x, y)$:

$$mi(x, y) = \int \int p(x, y) log \frac{p(x, y)}{p(x)p(y)} dxdy \ . \tag{1}$$

the normalized version of mutual information is:

$$nmi(x, y) = \frac{2 \times mi(x, y)}{h(x) + h(y)} \ . \tag{2}$$

where $h(x) = \int p(x) \log p(x) dx$. Given conditional variable $\boldsymbol{z}$, the conditional mutual information between $\boldsymbol{x}$ and $\boldsymbol{y}$ is

$$cmi(x, y|z) = \int \int p(x, y|z) log \frac{p(x, y)}{p(x)p(y)} dxdy \ . \tag{3}$$

Given a finite set of features $\boldsymbol{F}$ and a finite set of labels $\boldsymbol{L}$, mutual information based feature selection methods is objective to find the optimal subset of features $\boldsymbol{SF^* \subseteq F}$ without reducing the information shared by features and labels, as can be written as:

$$SF^* = \underset{SF \subseteq F}{\arg\min} \{|SF| : mi(SF, L) = mi(F, L)\} \ . \tag{4}$$

It can also be considered as removing every unnecessary feature from $\boldsymbol{F}$. Using conditional mutual information, this formulation can be expressed as:

$$SF^* = \underset{SF \subseteq F}{\arg\min} \{|SF| : \forall f \epsilon F - SF, cmi(f, L|SF) = 0\} \ . \tag{5}$$

The Eq. (5) indicates that an optimal reduction of the original feature set $\boldsymbol{F}$ should contain no irrelevant or redundant features. However, either Eqs. (4) or (5) is difficult to calculate. In the following, we redefine the relevance and redundancy of features based on mutual information to guide the feature selection procedure to achieve this target.

Definition 4 (Relevance) given a finite label set $\boldsymbol{L = \{l_1, l_2, ..., l_n\}}$, the relevance of the feature $\boldsymbol{f}$ and the label set $\boldsymbol{L}$ is defined as:

$$rel(f, L) = \max\{nmi(f, l_i), l_i \epsilon L\} \ . \tag{6}$$

The $rel(f, L)$ measures the relevance between feature $\boldsymbol{f}$ and the label set $\boldsymbol{L}$. Moreover,it delivers in pairwise manner, as can be calculated with efficiency. Obviously, if $rel(f, L) = 0$, f shares little information with any label $l_i \in L$. In other words, f can be discarded without harming the predictive performance. However, 0 is a threshold which is too strict to use in real applications. A compromise choice is using a small positive relevance threshold $\boldsymbol{\alpha}$. If $rel(f, L) \leq \alpha$, f is considered to be an irrelevant feature.

Definition 5 (Redundancy). let $\boldsymbol{F}$ is a finite feature set, for any feature $\boldsymbol{g}\epsilon\boldsymbol{F}$, the significance of $\boldsymbol{g}$ on $\boldsymbol{L}$ given another $\boldsymbol{h}\epsilon\boldsymbol{F}$ is defined as

$$sig(g, L|h) = \max\{cmi(g, l_i|h), l_i \epsilon L\} \ . \tag{7}$$

which means that a feature $\boldsymbol{g}$ is redundant and can be removed from $\boldsymbol{F}$ if there exists a feature $\boldsymbol{h}\epsilon\boldsymbol{F}-\boldsymbol{g}$ satisfying $sig(g, L|h) = 0$. This is a loosed and approximate version of the formulation $cmi(g; l_i|F - g) = 0$ described in Eq. 5. It only considers second-order conditional dependency but is much easier and efficient to calculate.

3.5 The Proposed Method

We propose a multi-label online feature selection algorithm named multi-label online streaming feature selection based on spectral granulation and mutual information (ML-OSMI) on the basis of Sects. 3.2, 3.3 and 3.4. The pseudo-code

Algorithm 1. ML-OSMI

Input: Feature stream $\boldsymbol{F}$, label space $\boldsymbol{L}$ and the relevance threshold $\boldsymbol{\alpha}$
Output: selected features $\boldsymbol{SF}$

granulating labels into $\boldsymbol{Z} = \{\boldsymbol{z_1}, \boldsymbol{z_2}, ..., \boldsymbol{z_k}\}$ using spectral clustering;
$\boldsymbol{SF} = \emptyset$;
repeat
  get $\boldsymbol{f}$ from the stream $\boldsymbol{F}$;
  /*checking relevance */
  if $rel(f, L) \leq \alpha$ **then**
    continue;
  end
  /*checking redundancy */
  $added = 1$;
  for $\boldsymbol{a_j}$ *in* $\boldsymbol{SF}$ **do**
    /*checking whether f is redundant*/
    if $sig(f, z_i, a_j) == 0$ **then**
      $added = 0$;
      break;
    end
    /*checking whether a_j is redundant*/
    if $sig(a_j, z_i, f) == 0$ **then**
      $\boldsymbol{SF} = \boldsymbol{SF} \backslash \boldsymbol{a_j}$;
    end
  end
  if $added == 1$ **then**
    $\boldsymbol{SF} = \boldsymbol{SF} \cup \boldsymbol{f}$;
  end
until *no new features or stopping criteria met*;
return $\boldsymbol{SF}$

of ML-OSMI is shown in Algorithm 1. ML-OSMI delivers as follows. As a new feature f flows in, if $rel(f, L) \leq \alpha$ is satisfied, f is considered to be a irrelevant feature and discarded. The online feature selection waits for the next feature. If f passes the relevance checking at Step 6, the algorithm assesses two kinds of redundancy, the redundancy of f and the redundancy of selected features before time t. Suppose $\boldsymbol{SF}$ is the set of selected features before f arrives. Firstly, the algorithm checks the redundancy of f to determine whether there exists a feature $a_j \in SF$ making f conditionally independent to the label set. If there has no such a feature in $\boldsymbol{SF}$, f is selected. Then, the algorithm removes all features made to be redundant by f from $\boldsymbol{SF}$. If there has no new features, the algorithm terminates.

3.6 Analysis of Time Efficiency

The time complexity of the proposed algorithm consists of two parts: the complexity of conducting relevance analysis and the complexity of removing redundant features. In the analysis, the number of samples is omitted for simplicity. Let $\boldsymbol{F_t}$ be the features arrived before time $\boldsymbol{t}$. $\boldsymbol{F_t^r}$ is a subset of $\boldsymbol{F_t}$ containing all features which are relevant to the label set. Suppose $\boldsymbol{SF_t}$ is the selected feature subset at time $\boldsymbol{t}$ and $\boldsymbol{r} = |\boldsymbol{SF_t}|$. Let $\boldsymbol{m} = |\boldsymbol{F_t}|$ be the number of features in $\boldsymbol{F_t}$ and $\boldsymbol{p} = |\boldsymbol{F_t^r}|$. When the number of feature is extremely high, it has $\boldsymbol{m \gg p \gg r}$. Hence, the average time complexity of the proposed algorithm is $\boldsymbol{O(km + kpr)}$, where $\boldsymbol{k}$ is the number of label granulas and $\boldsymbol{k \ll r}$. If all features are discarded on the relevance test, the best time complexity is $\boldsymbol{O(km)}$. While all features pass the independence test, the worst-case complexity is $\boldsymbol{O(kmr)}$. Noticing that $\boldsymbol{k \ll n}$ and $\boldsymbol{r \ll m}$, where $\boldsymbol{n}$ is the cardinality of the original label set, one can concludes that $\boldsymbol{O(kmr) \ll O(nm^2)}$.

4 Experiment Results

4.1 Experiment Settings

We use twelve multi-label high-dimensional benchmark data sets from various domains as our test beds. The details of data sets are shown in Table 1. The *scene* is from the image processing application. *emotions* and *CAL500* involve emotions classification of music. *genbase* and *yeast* are obtained in biology domain. The rest seven data sets are from text and natural language processing topics. All data sets are available at the MEKA website[1]. The experiments are conducted on a personal computer with Windows Server 2016, Inter(R) Core (TM) i7-6850K CPU and 64 GB memory employing MATLAB R2016a platform.

To illustrate the effectiveness of the proposed algorithm, we compare our algorithm with four state-of-the-art multi-label feature selection algorithms and two state-of-the-art single-label online feature selection algorithms. The comparisons contain the number of selected of features, running time and prediction

[1] http://meka.sourceforge.net/#datasets.

performances. The predictions are delivered by the multi-label k-nearest neighbors algorithm(ML-KNN) [24] trained with the selected features. ML-KNN is a well known multi-label classification method for its efficiency. In our experiments, the number of nearest neighbors is set to the recommended value 10 and the smoothing factor is 1. Five widely used evaluations are used to measure the predictive performances, namely Hamming Loss, Coverage, One Error, Ranking Loss and Average Precision [6]. The greater the value of Average Precision, the better the performance of the model. For the other four evaluations, the less their value are, the better the model is.

Table 1. Details of the benchmark data sets

Ind	Dataset	Instance	Feature	Label	Domain
1	emotions	593	72	6	music
2	bibtex	7395	1836	159	text
3	CAL500	502	68	174	music
4	delicious	16105	500	983	text
5	enron	1702	1001	53	text
6	genbase	662	1186	27	biology
7	languagelog	1460	1004	75	text
8	medical	978	1449	45	ext
9	scene	2407	294	6	images
10	tmc2007	28596	49060	22	text
11	20NG	19299	1006	20	text
12	yeast	2417	103	14	biology

4.2 Comparisons with Traditional Multi-label Feature Selection Methods

The comparative multi-label feature selection algorithms are F-Score [25], MLM-RMR [18,19], RFS [20] and MIFS [7]. Comparisons on running time and predictive performances are given. The implements of these algorithms can be found on Github[2] and the parameters such as the size of selected features are set as their default value. Moreover, the 5-fold validation mechanism is adopted on all data sets. Table 2 gives the running time and Fig. 2 shows the predictive performances of multi-label feature selection methods.

(1) ML-OSMI vs. F-Score. As is shown in Table 2, F-Score takes fewer time on 8 of 12 data sets except for *CAL500*, *enron*, *genbase* and *medical*. However, Fig. 2 shows that ML-OSMI achieves higher Average Precision on 11 of 12 except for the *bibtex*. There has no significant difference on Coverage among all feature selection methods. For Hamming Loss, ML-OSMI delivers better results on

[2] https://github.com/KKimura360/MLC_toolbox.

Table 2. Running time (Seconds)

Ind	F-Score	MIFS	RFS	MLMRMR	Proposed
1	**0.013**	0.117	0.795	0.026	0.102
2	**3.719**	29.751	603.866	30.022	18.655
3	0.141	0.897	0.529	**0.011**	0.041
4	**7.294**	424.569	3989.808	96.614	45.698
5	0.650	1.837	20.059	2.209	**0.388**
6	0.361	0.529	3.103	0.882	**0.025**
7	**1.074**	1.978	13.013	2.575	3.441
8	0.733	1.157	6.733	1.558	**0.532**
9	**0.029**	0.717	34.849	1.044	2.388
10	**0.499**	27.580	21669.986	20.382	1.142
11	**0.523**	19.950	6544.833	14.217	24.936
12	**0.027**	0.568	34.542	0.285	0.545

enron, *genbase*, *languagelog*, *medical* and *scene*. On other 7 data sets, ML-OSMI and F-Score perform equally well. Besides, ML-OSMI obtains better performance on 9 out of 12 data sets for One Error and 10 out of 12 data sets for Ranking Loss.

(2) ML-OSMI vs. MIFS. Table 2 says that ML-OSMI uses fewer time to select features on 9 out of 12 data sets than MIFS. Figure 2 shows that ML-OSMI performs better than MIFS on all data sets but the *scene* on Average Precision, Hamming Loss and Ranking Loss. For Coverage, neither of them shows superiority. Moreover, except for *scene* and *languagelog*, ML-OSMI gains better results of One Error than MIFS.

(3) ML-OSMI vs. RFS. The comparisons between ML-OSMI and RFS in Table 2 show that ML-OSMI achieves better time efficiency on all data sets. For the predictive performances, Fig. 2 indicates that ML-OSMI gets better results evaluated by Average Precision, One Error and Ranking Loss on all data sets except for the *emotions* and *languagelog*. Besides, ML-OSMI outperforms RFS on 8 out of 12 data sets on Hamming Loss and delivers the same results on 3 of the remaining 4 data sets. For Coverage, ML-OSMI and RFS perform almost equally well.

(4) ML-OSMI vs. MLMRMR. Table 2 shows that MLMRMR takes less time than ML-OSMI on *emotions*, *CAL500*, *languagelog*, *scene*, *20NG* and *yeast*, while ML-OSMI takes less time than MLMRMR on the other 6 data sets. As Fig. 2 shows, ML-OSMI performs better than MLMRMR on *enron* and *scene* and MLMRMR performs better than ML-OSMI on *enron* and *bibtex*. On the remaining 9 data sets, ML-OSMI performs as good as MLMRMR.

4.3 Comparisons with OSFS Methods in Streaming Feature Scenario

We also compare ML-OSMI with two state-of-the-art OSFS algorithms, Alpha-investing [14] and SAOLA [22]. To evaluate the effectiveness of the proposed multi-label online streaming feature selection algorithm, we choose 8 data sets with extreme high dimensionality to simulate the streaming feature selection scenario. Average Precision and Hamming Loss are used as the criterions to demonstrate the performance of the algorithms. Figure 3 reports the performances of LP-SAOLA, LP-alpha-investing and ML-OSMI with the features flowing in continuously over time. Table 3 gives the running time.

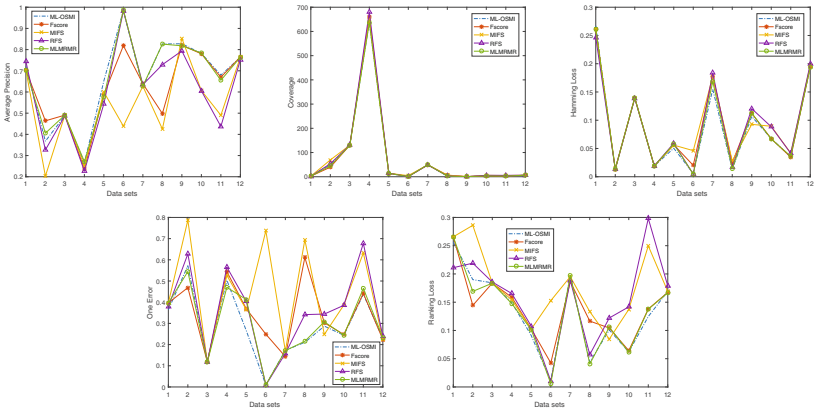

Fig. 2. Comparisons with multi-label feature selection methods

Table 3. Running time (Seconds)

Dataset	lp-alpha-investing	lp-saola	Proposed
emotions	**0.004**	0.154	0.102
bibtex	**15.331**	435.656	18.655
CAL500	**0.003**	0.193	0.041
delicious	**6.180**	43.281	45.698
enron	0.416	109.994	**0.388**
genbase	1.137	1.068	**0.025**
languagelog	**0.875**	106.720	3.441
medical	**0.481**	150.697	0.532
scene	**0.211**	1.153	2.388
tmc2007	18.408	52.927	**1.142**
20NG	41.580	157.684	**24.936**
yeast	**0.007**	0.027	0.545

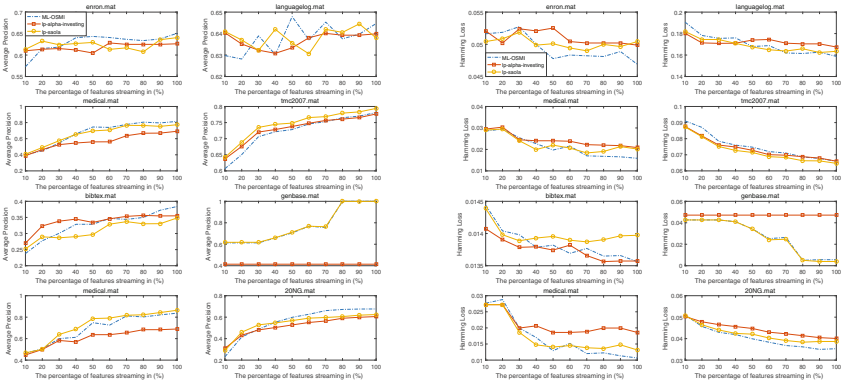

Fig. 3. The predictive performance changes with features streaming in

(1) ML-OSMI vs. LP-alpha-investing. Figure 3 shows that the proposed algorithm outperforms LP-alpha-investing on 6 out of 8 data sets evaluated by Average Precision and Hamming Loss. For *mc2007*, LP-alpha-investing generates better results on the prior 80% of features than ML-OSMI. However, with new features continuously flow in, ML-OSMI performs better than LP-alpha-investing. Table 3 says that *LP-alpha-investing* takes less time dealing with 8 out of 12 data sets. It should be noted that LP-alpha-investing transforms the whole label set into a single multi-class label, as makes it more time efficiency.

(2) ML-OSMI vs. LP-saola. On *enron*, *medical*, *bibtex* and *20NG*, ML-OSMI gets better Average Precision and Hamming Loss with features streaming flowing in. Besides, compared to LP-SAOLA, the proposed algorithm gains better time efficiency on 9 out of 12 data sets except for *delicious*, *scene* and *yeast*. Especially, on six relatively higher dimensional data sets with thousands of features, *bibtex*, *enron*, *genbase*, *medical*, *tmc2007* and *20NG*, the proposed algorithm shows better efficiency for taking relative less time.

5 Conclusion

In this paper, we propose a multi-label online streaming feature selection algorithm to address multi-label feature selection with dynamic features. The proposed method first granulates the labels. Labels in the same granula are high correlated and labels in different granula are mutually independent or weakly correlated. Then, transforming each granula of labels into a multi-class label, the original labels is converted into a new space with much lower dimensionality, taking high-order correlations into consideration. Moreover, the relevance and redundancy of features are redefine based on mutual information to guide feature selection procedure. Finally, the features are selected with the new label space in online manner. Comprehensive experiments are conducted to verify the effectiveness of the proposed method, comparing it with traditional multi-label feature selection methods and online streaming feature selection methods. Results have

shown that the proposed multi-label online feature selection algorithm can effectively solve multi-label feature selection with dynamic features. In our future work, we will study how to deliver feature selection with features and labels flow in simultaneously.

Acknowledgements. This work was supported by the National Key Research and Development Program of China (Grant no. 2016YFB1000900), the National Natural Science Foundation of China (Grant nos. 61572091, 61772096), Chongqing Basic and Frontier Research Project (cstc2015jcyjA40018) and The Science and Technology Project Affiliated to the Education Department of Chongqing Municipality (KJ1500438).

References

1. Hua, X.S., Qi, G.J.: Online multi-label active annotation: towards large-scale content-based video search. In: International Conference on Multimedia 2008, Vancouver, British Columbia, Canada, pp. 141–150, October 2008
2. Lai, H., Yan, P., Shu, X., Wei, Y., Yan, S.: Instance-aware hashing for multi-label image retrieval. IEEE Trans. Image Process. **25**(6), 2469 (2016)
3. Trohidis, K., Tsoumakas, G., Kalliris, G., Vlahavas, I.P.: Multi-label classification of music into emotions. In: ISMIR 2008, 9th International Conference on Music Information Retrieval, Drexel University, Philadelphia, PA, USA, 14–18 September 2008, pp. 325–330 (2008)
4. Wu, B., Lyu, S., Hu, B.G., Ji, Q.: Multi-label learning with missing labels for image annotation and facial action unit recognition. Patt. Recogn. **48**(7), 2279–2289 (2015)
5. 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)
6. Tsoumakas, G., Katakis, I., Vlahavas, I.P.: Mining multi-label data. In: Data Mining and Knowledge Discovery Handbook, 2nd edn., pp. 667–685 (2010)
7. Jian, L., Li, J., Shu, K., Liu, H.: Multi-label informed feature selection. In: Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016, New York, NY, USA, 9–15 July 2016, pp. 1627–1633 (2016)
8. Lee, J., Kim, D.W.: Mutual information-based multi-label feature selection using interaction information. Expert Syst. Appl. **42**(4), 2013–2025 (2015)
9. Li, F., Miao, D., Pedrycz, W.: Granular multi-label feature selection based on mutual information. Patt. Recogn. **67**, 410–423 (2017)
10. Wu, X., Yu, K., Wang, H., Ding, W.: Online streaming feature selection. In: Proceedings of the 27th International Conference on Machine Learning (ICML 2010), 21–24 June 2010, Haifa, Israel, pp. 1159–1166 (2010)
11. Wu, X., Yu, K., Ding, W., Wang, H.: Online feature selection with streaming features. IEEE Trans. Patt. Anal. Mach. Intell. **35**(5), 1178 (2013)
12. Wang, J., et al.: Online feature selection with group structure analysis. IEEE Trans. Knowl. Data Eng. **27**(11), 3029–3041 (2016)
13. Perkins, S., Theiler, J.: Online feature selection using grafting. In: Machine Learning, Proceedings of the Twentieth International Conference (ICML 2003), 21–24 August 2003, Washington, DC, USA, pp. 592–599 (2003)

14. Zhou, J., Foster, D.P., Stine, R.A., Ungar, L.H.: Streaming feature selection using alpha-investing. In: Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Chicago, Illinois, USA, 21–24 August 2005, pp. 384–393 (2005)
15. Cherman, E.A., Monard, M.C., Lee, H.D.: A comparison of multi-label feature selection methods using the problem transformation approach. Electr. Notes Theor. Comput. Sci. **292**, 135–151 (2013)
16. Spolaôr, N., Monard, M.C., Lee, H.D.: Feature selection for multi-label learning. In: Proceedings of the 24th International Conference on Artificial Intelligence, Series, IJCAI 2015, pp. 4401–4402. AAAI Press (2015)
17. Lin, Y., Hu, Q., Liu, J., Duan, J.: Multi-label feature selection based on max-dependency and min-redundancy. Neurocomputing **168**, 92–103 (2015)
18. Kimura, K., Sun, L., Kudo, M.: MLC toolbox: A MATLAB/OCTAVE library for multi-label classification. CoRR, abs/1704.02592 (2017). http://arxiv.org/abs/1704.02592
19. Peng, H., Long, F., Ding, C.: Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans. Patt. Anal. Mach. Intell. **27**(8), 1226 (2005)
20. Nie, F., Huang, H., Cai, X., Ding, C.H.Q.: Efficient and robust feature selection via joint $l_{2,1}$-norms minimization. In: Advances in Neural Information Processing Systems 23: 24th Annual Conference on Neural Information Processing Systems 2010. Proceedings of a meeting held 6–9 December 2010, Vancouver, British Columbia, Canada, pp. 1813–1821 (2010)
21. Lin, Y., Hu, Q., Zhang, J., Wu, X.: Multi-label feature selection with streaming labels. Inf. Sci. **372**, 256–275 (2016)
22. Yu, K., Wu, X., Ding, W., Pei, J.: Towards scalable and accurate online feature selection for big data. In: 2014 IEEE International Conference on Data Mining, ICDM 2014, Shenzhen, China, 14–17 December 2014, pp. 660–669 (2014)
23. Sun, L., Kudo, M., Kimura, K.: Multi-label classification with meta-label-specific features. In: 23rd International Conference on Pattern Recognition, ICPR 2016, Cancún, Mexico, 4–8 December 2016, pp. 1612–1617 (2016)
24. Zhang, M.L., Zhou, Z.H.: ML-KNN: a lazy learning approach to multi-label learning. Patt. Recogn. **40**(7), 2038–2048 (2007)
25. Kong, D., Ding, C.H.Q., Huang, H., Zhao, H.: Multi-label reliefF and F-statistic feature selections for image annotation. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012, pp. 2352–2359 (2012)