Bayesian Chain Classifier with Feature Selection for Multi-label Classification

Ricardo Benítez Jiménez, Eduardo F. Morales, and Hugo Jair Escalante.

Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE)
Sta. María Tonantzintla, CP 72840, Puebla, México.
{ricardo.benitez, emorales, hugojair}@inaoep.mx

Abstract. Multi-label classification task has many applications in Text Categorization, Multimedia, Biology, Chemical data analysis and Social Network Mining, among others. Different approaches have been developed: Binary Relevance (BR), Label Power Set (LPS), Random k label sets (RAkEL), some of them consider the interaction between labels in a chain (Chain Classifier) and other alternatives around this method are derived, for instance, Probabilistic Chain Classifier, Monte Carlo Chain Classifier and Bayesian Chain Classifier (BCC). All previous approaches have in common and focus on is in considering different orders or combinations of the way the labels have to be predicted. Given that feature selection has proved to be important in classification tasks, reducing the dimensionality of the problem and even improving classification model's accuracy. In this work a feature selection technique is tested in BCC algorithm with two searching methods, one using Best First (BF-FS-BCC) and another with GreedyStepwise (GS-FS-BCC), these methods are compared, the winner is also compared with BCC, both tests are compared through Wilcoxon Signed Rank test, in addition it is compared with others Chain Classifier and finally it is compared with others approaches (BR, RAkEL, LPS).

Keywords: Multi-label classification · Chain classifier· BCC · Feature Selection.

1 Introduction

Classical classification task consists of determining a label or class for an instance, which is characterized by n features, on the other hand, multi-label classification consists in predicting a subset $l \subseteq L$ of binary labels (0,1). The task of multi-label classification has many applications in Text Categorization, Multimedia, Biology, Chemical data analysis and Social Network Mining, among others. This task has to deal with problems like high computational cost, derived from considering dependencies between labels and the complexity of the resulting model.

There are two well-known approaches for multi-label classification task: Binary Relevance (BR) and Label Power Set (LPS), the BR approach creates n binary

classifiers, one per label and determines the subset of labels according to each binary classifier[17]. In contrast, LPS produces n classifiers according to all the possible combinations of binary labels existing on the training data [12], which in the worst case yields 2^n classifiers. These methods have some disadvantages, for instance, BR does not consider the relationship between labels and LPS might suffer from a high computational complexity depending on the number of n different subset of labels.

With the intention to deal with the complexity of the models and the interaction between labels, other approaches have been developed. One strategy is Random k label sets (RAkEL), which by following the LPS logic creates n classifiers but considering only a subset of labels with a size of k [11]. On the other hand, the Chain Classifier (CC) has been proposed by considering the interaction between labels, it uses a classifier for each label, considers the previous prediction l_i to classify the next label l_j [6] and others alternatives around this method are derived, for instance, Probabilistic Chain Classifier (PCC) which estimates the optimal chain but at a high computational cost, therefore, it is recommended to use it in datasets with no more than 15 labels [1], Monte Carlo Chain Classifier (MCC) estimates the optimal chain through a Monte Carlo approach [5]. Furthermore, there is also the Bayesian Chain Classifier (BCC) which builds a Bayesian Network and then a undirected tree using Chow and Liu's algorithm, then it chooses a node as the root and starting off from there it creates a directed tree [15].

Nevertheless, one thing all previous approaches have in common and focus on is in considering different orders or combinations of the way the labels have to be predicted. In this work, a feature selection technique is applied in the middle of the BCC algorithm, the idea is to improve its performance by improving separately the individual internal classifiers with the feature selection.

2 Related Work

Feature selection has proved to be important in classification tasks, reducing the dimensionality of the problem and even improving classification model's accuracy.

Recently, strategies been developed to reduce the number of features used in the multi-label classification task. In [8] they used feature selection with two methods, RefielF and Information Gain, in a BR strategy, in both cases using a filter approach for each label in two different scenarios, one with C4.5 decision trees and the other one with Support Vector Machine (SVM) as internal classifier, in [9] they extend the work with the LPS approach using lazy k Nearest Neighbor algorithm as internal classifier.

Fast Correlation-Based Filter is also explored in [4] by creating a Maximum Expanding Tree between each label and each of the features, again implemented along with a BR strategy. Other measures have been explored in [14], in order to determine relevant features, including a fast calculation method for chi-square statistics is developed in an attempt to reduce the time this schema requires on BR.

Furthermore, in [16] they compare Principal Component Analysis (PCA) and Genetic Algorithms (GA) as a wrapper style. Naive Bayes is assigned as internal classifier and using transformed algorithms for multi-label classification like Adaboost and Rank-SVM. Although the base approach, Multi-Label Naive Bayes (MLNB) maintains better results than MLNB-PCA and MLNB-GA, however, these are competitive with a smaller amounts of features.

Hence, this work proposes a basic approach to select features for the BCC algorithm using a search of the subset of features in two different ways, Best First (BF) and GreedyStepwise (GS), while using CfsSubsetEval as the evaluator of the subset, these techniques are implemented within the BCC algorithm. BF, GS and Correlation-based Feature Selection[3] (CfsSubsetEval) are implemented in Weka[13].

3 Methodology and development

Two searching methods are applied to find the subset of features to be used in the BCC algorithm (FS-BCC), the first one: with a BF technique (BF-FS-BCC). The second one: GreedyStepwise Technique (GS-FS-BCC), this approaches are tested with the Naive Bayes as the internal classifier.

3.1 Feature Selection in BCC

BCC algorithm[10] is modified with the intention to introduce a feature selection technique in this case BF or GS resulting on FS-BCC.

Given the structure and the chain sequence of the BCC algorithm, for each label a subset of features X is selected and a classifier is built. This can be summarize as following:

1. Obtain the undirected tree of the classes using Chow and Liu's algorithm.
2. Create a directed tree selecting randomly one class as a root node.
3. For each class C_j in the chain do:
   (a) Do until the worth of the subset doest not improve:
      i. Search with BF or GS the subset of features using CfsSubsetEval as heuristic.
      ii. Eval the worth of a subset with respect to accuracy.
   (b) Build a Navie Bayes classifiers for the class C_j with the selected features and its parent $Pa(C_j)$.
4. To determinate the subset of classes that correspond a new instance, the output of the each internal classifier is concatenated.

Figure 1 illustrates this approach using a Naive Bayes as internal classifier. On the left, original BCC algorithm building a classifier with all the features and its parent for each class (C_{j1},C_2...,C_5), on the right, FS-BCC build a classifier with a subset of the features for each class, hence each internal classifier can contain different features. It is important to mention that this approach continues using the parent in the tree as an additional feature.

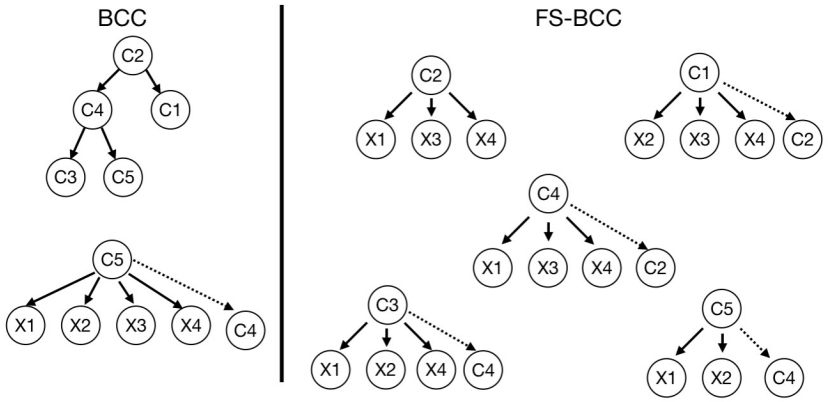

Fig. 1. Original BCC in the left and BCC with Feature Selection (FS-BCC) in the right.

FS-BCC has been implemented in Meka [7], a software for multi-label classification that includes algorithms implemented in Weka [13] like BF and GS, furthermore include CfsSubsetEval evaluator. BF and GS are used with the forward variant, this variant begins with a subset of one feature and continues adding other new feature until the worth of the subset stops improving[1].

3.2 Datasets

In table 1 is possible to see number of labels (classes), instances, features and the domain of each dataset. It can be observed that the domains are Music, Image, Text, and Biology, the number of labels and features goes from 6 to 159 and 71 to 1836 respectively, on the other hand, the number of instances goes from 593 to 120,919. These datasets are available in http://mulan.sourceforge.net/datasets-mlc.html and http://waikato.github.io/meka/datasets/.

[1] The implementation code is available in: https://github.com/R-Benitez-J/FS-BCC

Table 1. Description of each dataset.

Num.	Dataset	Labels	Instances	Features	Domain
1	Music	6	593	71	Music
2	Scene	6	2,407	294	Image
3	Slashdot	22	3,782	1,079	Text
4	Yeast	14	2,417	1,03	Biology
5	20NG	20	19,300	1,006	Text
6	Enron	53	1,702	1,001	Text
7	LangLog	75	1,460	1,004	Text
8	Medical	45	978	1,449	Text
9	Ohsumed	23	13,929	1,002	Text
10	IMDB	28	120,919	1,001	Text
11	Bibtex	159	7,395	1,836	Text

3.3 Evaluation Measures

Four evaluation measures are considered in this work, Multi-label Accuracy also called Accuracy, Macro Average (by example), Hamming Score (Mean Accuracy) and elapsed time, the last one with the intention to have more information when choosing BF-FS-BCC or GS-FS-BCC, these measures can be found in Meka[7].

Accuracy: Known as multi-label accuracy is given by the equation 1, where c_j is the real value of the label in the vector of subset of labels and c'_j is the predicted value for the classification model.

$$Accuracy = \frac{1}{N}\sum_{i=i}^{N}\frac{|c_i \wedge c'_i|}{|c_i \vee c'_i|} \tag{1}$$

Macro-Average (by example): Corresponds to a half of F1-Score but in this case evaluated by example.

$$MacroAverage = \frac{1}{2}\sum_{1}^{N}(2 * Precision * Recall/(Precision + Recall)) \tag{2}$$

Hamming Score (Mean Accuracy): Refers to the average of the accuracy for each label, where, c'_{ij} corresponds to the prediction given by the model and c_{ij} to the real value (Equation 3).

$$HammingScore = \frac{1}{d}\sum_{j=1}^{d}\frac{1}{N}\sum_{i=1}^{N}\delta\left(c'_{ij}, c_{ij}\right) \tag{3}$$

Time: The time is presented in seconds and represents the average elapsed time per fold, in this case, all the experiments were tested with 10-Cross-Fold Validation.

Percentage of Selected Features: Indicates the percentage of the features that an internal classifier is built, in equation 4, N_{sf} corresponds to the selected features through FS-BCC and N_f to the original number of attributes.

$$\%SelectedFeatures = \frac{N_{sf} \times 100}{N_f} \tag{4}$$

4 Experiments and results

In the same way BCC algorithm was proposed [15], [10], Naive Bayes is used as the internal classifier.

The experiments were development in four phases, in the first phase, BF-FS-BCC and GS-FS-BCC are compared looking for significance difference between them. In the second phase the winner of the previous phase was compared with the original BCC algorithm, in both cases (first and second phase) the method to compared the differences was Wilcoxon signed rank [2] with a significance level of 0.05. In the third phase FS-BCC and BCC are compared against Chain Classifiers approaches like CC, MCC and PCC (limited to 15 labels). At the end, in the fourth phase BCC and FS-BCC is compared against other approaches like LPS and RAkEL with k=3 also used in [11].

4.1 BF-FS-BCC compared to GS-FS-BCC

In this phase, four Wilcoxon Signed Rank tests were developed, one for each evaluation measure and also the time, in table 2, it is possible to observe the Wilcoxon Signed Rank test for Hamming Score and Macro Average, the first two columns correspond to BF-FS-BCC, the next two columns to GS-FS-BCC and the last two correspond to the absolute difference and rank. Table 3 presents Accuracy and Elapsed Time in the same form.

Wilcoxon Signed Rank test indicates that there are no significant differences in Macro Average, Accuracy and elapsed time between GS-FS-BCC and BF-FS-BCC, only in Hamming Score GS-FS-BCC is significantly better than BF-FS-BCC, however in average BF-FS-BCC is better than GS-FS-BCC, hence BF-FS-BCC is chosen to be compared with BCC algorithm and other approaches.

Selected Features: Table 4 indicates the average (Avg.) of number and percentage of selected features for each dataset, including the corresponding standard deviation (STD). These values correspond to the BS-FS-BCC method. In summary, this method uses only the 5.63% of the total number of features with a STD of 1.47%.

Table 2. Hamming Scores (Ham. Score) and Macro Average (Ma. Avg.) of BF-FS-BCC and GS-FS-BCC.

		BF-FS-BCC		GS-FS-BCC		Abs. Diff. (Rank)	
Num.	Dataset	Ham. Score	Ma. Avg.	Ham. Score	Ma. Avg.	Ham. Score	Ma. Avg.
1	Music	0.7473	**0.601**	**0.7482**	0.597	0.0009 (9)	0.004 (6)
2	Scene	0.8142	0.614	0.8142	0.614	0 (n/a)	0 (n/a)
3	Slashdot	0.944	**0.363**	**0.9443**	0.361	0.0003 (6)	0.002(3)
4	Yeast	**0.7473**	**0.572**	0.7472	0.571	0.0001 (1.5)	0.001 (1.5)
5	20NG	0.9471	0.494	**0.9474**	0.494	0.0003 (5)	0 (n/a)
6	Enron	0.9264	0.471	**0.9268**	0.471	0.0004 (7)	0 (n/a)
7	LangLog	0.9609	0.176	**0.961**	**0.179**	0.0001 (1.5)	0.003 (4.5)
8	Medical	0.9871	**0.772**	0.9871	0.767	0 (n/a)	0.005 (7)
9	Ohsumed	**0.9270**	0.478	0.9269	0.478	0.0001 (3)	0 (n/a)
10	IMDB	0.9181	**0.077**	**0.9183**	0.074	0.0002 (4)	0.003 (4.5)
11	Bibtex	0.9423	0.417	**0.9429**	**0.418**	0.0006 (8)	0.001(1.5)
	Avg.	0.8965	**0.4577**	**0.8968**	0.4567		

Table 3. Accuracy and Elapsed Time of BF-FS-BCC and GS-FS-BCC

		BF-FS-BCC		GS-FS-BCC		Abs. Diff. (Rank)	
Num.	Dataset	Accuracy	Time	Accuracy	Time	Accuracy	Time
1	Music	**0.508**	0.186	0.505	**0.137**	0.003 (6.5)	0.049 (1)
2	Scene	0.51	**19.444**	0.51	21.726	0 (n/a)	2.282 (3)
3	Slashdot	**0.324**	294.425	0.323	**68.729**	0.001 (2.5)	225.696(9)
4	Yeast	0.453	**1.657**	0.453	1.751	0 (n/a)	0.094(2)
5	20NG	0.44	**893.808**	**0.441**	961.932	0.001 (2.5)	68.124(7)
6	Enron	0.357	146.746	**0.358**	**133.658**	0.001 (2.5)	13.087(4)
7	LangLog	0.272	280.189	**0.275**	**256.435**	0.003 (6.5)	23.757(6)
8	Medical	**0.725**	41.987	0.721	**27.99**	0.004 (8)	13.997(5)
9	Ohsumed	0.393	661.65	0.393	**517.866**	0 (n/a)	143.784(8)
10	IMDB	**0.06**	5779.67	0.058	**5137.666**	0.002 (5)	642.004(10)
11	Bibtex	0.323	**16866.793**	**0.324**	18375.316	0.001 (2.5)	1508.523(11)
	Avg.	**0.3968**	**2271.505**	0.3965	2318.4731		

4.2 BCC compared to BF-FS-BCC

Table 5, shows the Hamming Score and Macro Average of BCC and BF-FS-BCC, the Wilcoxon Signed rank indicates a significant difference in Hamming Score where BF-FS-BCC is the winner, on the other hand there is no difference in Macro Average and Accuracy (Table 6). Nevertheless, BF-FS-BCC is better than BCC at least on average.

4.3 BF-FS-BCC compared to others Chain Classifiers

This section presents the results of the comparative of BF-FS-BCC with CC, MCC and PCC multi-label classifiers, in these results only the nine first datasets are used given the required time to test the IMDB and Bibtex datasets.

Table 4. Average (Avg.) and standard deviation (STD) of Selected Features expressed in number and percentage.

Num.	Dataset	Avg. and STD of Selected Features	Avg. and STD. of Selected Features in %
1	Music	19.48 (±5.69)	3.29% (±0.96%)
2	Scene	69.43 (±22.45)	23.62% (±7.63%)
3	Slashdot	21.37 (±12.97)	1.98% (±1.2%)
4	Yeast	18.68 (±9.91)	18.13% (±0.51%)
5	20NG	21.76 (±0.89)	2.17% (±0.89%)
6	Enron	12.39 (±8.48)	1.24% (±0.85%)
7	LangLog	18.19 (±14.09)	1.81% (±1.4%)
8	Medical	8.94 (±6.43)	0.62% (±0.44%)
9	Ohsumed	20.98 (±3.05)	2.09% (±0.3%)
10	IMDB	13.127 (±5.223)	1.311% (±0.522%)
	Avg.	22.43 (±8.918)	5.626% (±1.47%)

Table 5. Hamming Score, Macro Average and Wilcoxon Test of BCC and BF-FS-BCC

		BCC		BF-FS-BCC		Abs. Diff. (Rank)	
Num.	Dataset	Ham. Score	Ma. Avg.	Ham. Score	Ma. Avg.	Ham. Score	Ma. Avg.
1	Music	**0.7513**	**0.636**	0.7473	0.601	0.004 (1)	0.035 (2)
2	Scene	0.7597	0.566	**0.8142**	**0.614**	0.0545 (7)	0.048 (4)
3	Slashdot	0.9306	**0.448**	**0.9442**	0.363	0.0134 (2)	0.085 (6)
4	Yeast	0.6982	0.54	**0.7473**	**0.572**	0.0491 (6)	0.032 (1)
5	20NG	0.8988	0.457	**0.9471**	**0.494**	0.0483 (5)	0.037 (3)
6	Enron	0.8038	0.34	**0.9264**	**0.471**	0.1226 (10)	0.131 (10)
7	LangLog	0.7252	0.078	**0.9609**	**0.176**	0.2357 (11)	0.098 (8)
8	Medical	0.967	0.636	**0.9871**	**0.772**	0.0201 (3)	0.136 (11)
9	Ohsumed	0.8575	0.399	**0.927**	**0.478**	0.0695 (8)	0.079 (5)
10	IMDB	0.8757	**0.194**	**0.9181**	0.077	0.0424 (4)	0.117 (9)
11	Bibtex	0.8597	0.319	**0.9423**	**0.417**	0.0826 (9)	0.098 (7)
	Avg.	0.8298	0.4194	**0.8965**	**0.4577**		

In table 7, Hamming Score is presented, Accuracy in table 8 and Macro Average in table 9, BF-FS-BCC is better on average in at least two measures, Accuracy and Hamming Score, in Macro Average, MCC is better but BF-FS-BCC is competitive.

4.4 BF-FS-BCC compared to others Approaches

BCC algorithm is highly competitive compared to other approaches (BR, RAkEL and LPS), hence given that BF-FS-BCC is better than BCC algorithm at least on average, this method is compared with these other approaches.

From table 10, 11 and 9 we can determine, BF-FS-BCC is highly competitive with other approaches, even outperforms some of them (BR and RAkEL) in

Table 6. Accuracy and Wilcoxon Test of BCC and BF-FS-BCC

Num.	Dataset	BCC	BF-FS-BCC	Abs. Diff. (Rank)
1	Music	**0.534**	0.508	0.026 (1)
2	Scene	0.453	**0.51**	0.057 (3)
3	Slashdot	**0.389**	0.324	0.065 (4)
4	Yeast	0.421	**0.453**	0.032 (2)
5	20NG	0.37	**0.44**	0.07 (5)
6	Enron	0.227	**0.357**	0.13 (10)
7	LangLog	0.183	**0.272**	0.089 (7)
8	Medical	0.537	**0.725**	0.188 (11)
9	Ohsumed	0.292	**0.393**	0.101 (8)
10	IMDB	**0.144**	0.06	0.084 (6)
11	Bibtex	0.21	**0.323**	0.113 (9)
	Avg.	0.3418	**0.3968**	0.055

Table 7. Hamming Score of BF-FS-BCC compared to other chain classifiers

Dataset	BCC	BF-FS-BCC	CC	MCC	PCC
Music	0.7513	0.7473	**0.7522**	0.7512	0.751
Scene	0.7597	**0.8142**	0.7630	0.7630	0.763
Slashdot	0.9306	**0.9440**	0.9357	0.9343	–
Yeast	0.6982	**0.7473**	0.6952	0.6944	–
20NG	0.8988	**0.9471**	0.9135	0.9134	–
Enron	0.8038	**0.9264**	0.7962	0.7960	–
LangLog	0.7252	**0.9609**	0.6723	0.6724	–
Medical	0.9670	**0.9871**	0.9680	0.9679	–
Ohusumed	0.8575	**0.9270**	0.8627	0.8624	–
Avg.	0.8213	**0.889**	0.8176	0.8172	–

Table 8. Accuracy of BF-FS-BCC compared to other chain classifiers

Dataset	BCC	BF-FS-BCC	CC	MCC	PCC
Music	**0.534**	0.508	0.532	0.532	0.531
Scene	0.453	**0.51**	0.458	0.458	0.458
Slashdot	0.389	0.324	0.422	**0.434**	–
Yeast	0.421	**0.453**	0.418	0.42	–
20NG	0.37	**0.44**	0.399	0.4	–
Enron	0.227	**0.357**	0.228	0.228	–
LangLog	0.183	**0.272**	0.18	0.18	–
Medical	0.537	**0.725**	0.549	0.547	–
Ohsumed	0.292	**0.393**	0.294	0.294	–
Avg.	0.3784	**0.4424**	0.3866	0.3881	–

some cases. LPS was not tested in all the dataset given the number of labels in the dataset.

Table 9. Macro Average of BF-FS-BCC compared to other chain classifiers

Dataset	BCC	BF-FS-BCC	CC	MCC	PCC
Music	**0.636**	0.601	0.634	0.633	0.633
Scene	0.566	**0.614**	0.571	0.571	0.571
Slashdot	0.448	0.363	0.48	**0.494**	–
Yeast	0.54	**0.572**	0.532	0.534	–
20NG	0.475	**0.494**	0.482	0.483	–
Enron	0.34	**0.471**	0.34	0.34	–
LangLog	0.078	**0.176**	0.071	0.071	–
Medical	0.636	**0.772**	0.643	0.642	–
Ohsumed	0.399	0.077	**0.401**	**0.401**	–
Avg.	0.4575	0.46	0.4615	**0.4632**	–

Table 10. Hamming Score of BF-FS-BCC, BCC and other approaches

Num.	Dataset	BCC	BF-FS-BCC	BR	RAkEL	LPS
1	Music	0.7513	0.7473	0.7463	0.7025	**0.766**
2	Scene	0.7597	0.8142	0.7582	0.7908	**0.863**
3	Slashdot	0.9306	0.944	0.9318	0.93	**0.95**
4	Yeast	0.6982	0.7473	0.6983	0.6774	**0.759**
5	20NG	0.8988	0.9471	0.8971	0.9162	**0.964**
6	Enron	0.8038	**0.9264**	0.8037	0.8932	–
7	LangLog	0.7252	**0.9604**	0.7246	0.9251	–
8	Medical	0.967	**0.9871**	0.9671	0.9697	–
9	Ohsumed	0.8575	**0.927**	0.8577	0.8764	–
	Avg.	0.8213	**0.8889**	0.8205	0.8534	

Table 11. Accuracy of BF-FS-BCC, BCC and other approaches

Num.	Dataset	BCC	BF-FS-BCC	BR	RAkEL	LPS
1	Music	**0.534**	0.508	0.526	0.503	0.507
2	Scene	0.453	0.51	0.452	0.507	**0.615**
3	Slashdot	0.389	0.324	0.377	0.247	**0.51**
4	Yeast	0.421	0.453	0.422	0.426	**0.473**
5	20NG	0.37	0.44	0.367	0.357	**0.548**
6	Enron	0.227	**0.357**	0.227	0.046	–
7	LangLog	0.183	**0.272**	0.183	0.172	–
8	Medical	0.537	**0.725**	0.541	0.36	–
9	Ohsumed	0.292	**0.393**	0.293	0.208	–
	Avg.	0.3784	**0.4424**	0.3764	0.314	

5 Conclusions and future work

BF-FS-BCC was tested using 10-Cross-Fold validation and through these results, it is possible to determine that BF-FS-BCC is highly competitive compared to other Chain Classifier and even better in some cases than BR, RAkEl, using only in average 5.6% of the features. That is an important reduction of features, nevertheless, as a disadvantage, the time required to find the subset of features

Table 12. Macro accuracy of BF-FS-BCC, BCC and other approaches

Num.	Dataset	BCC	BF-FS-BCC	BR	RAkEL	LPS
1	Music	**0.636**	0.601	0.629	0.626	0.595
2	Scene	0.566	0.614	0.567	0.621	**0.641**
3	Slashdot	0.448	0.363	0.435	0.288	**0.537**
4	Yeast	0.54	**0.572**	0.54	0.553	0.568
5	20NG	0.457	0.494	0.454	0.424	**0.652**
6	Enron	0.34	**0.471**	0.34	0.074	–
7	LangLog	0.078	**0.176**	0.078	0.051	–
8	Medical	0.636	**0.772**	0.638	0.429	–
9	Ohsumed	0.399	**0.478**	0.4	0.281	–
	Avg.	0.4555	**0.5045**	0.4534	0.3718	

needs to be considered, although this elapsed time affects only in the model building phase.

As future work in this same approach, another internal classifier can be considered for more test. Also to compare BF-FS-BCC with transformed algorithms for multi-label classification like Adaboost and Rank-SVM, among others.

Other techniques of feature selection can apply with the intention to reduce the time of building model and improve the FS-BCC approach.

References

1. Dembczynski, K., Cheng, W., Hüllermeier, E.: Bayes optimal multilabel classification via probabilistic classifier chains. In: ICML. vol. 10, pp. 279–286 (2010)
2. Demšar, J.: Statistical comparisons of classifiers over multiple data sets. Journal of Machine learning research **7**(Jan), 1–30 (2006)
3. Hall, M.A.: Correlation-based feature selection for machine learning (1999)
4. Lastra, G., Luaces, O., Quevedo, J.R., Bahamonde, A.: Graphical feature selection for multilabel classification tasks. In: International Symposium on Intelligent Data Analysis. pp. 246–257. Springer (2011)
5. Read, J., Martino, L., Luengo, D.: Efficient monte carlo optimization for multi-label classifier chains. In: Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on. pp. 3457–3461. IEEE (2013)
6. Read, J., Pfahringer, B., Holmes, G., Frank, E.: Classifier chains for multi-label classification. Machine learning **85**(3), 333 (2011)
7. Read, J., Reutemann, P., Pfahringer, B., Holmes, G.: MEKA: A multi-label/multi-target extension to Weka. Journal of Machine Learning Research **17**(21), 1–5 (2016), http://jmlr.org/papers/v17/12-164.html
8. Spolaôr, N., Cherman, E.A., Monard, M.C., Lee, H.D.: Filter approach feature selection methods to support multi-label learning based on relieff and information gain. In: Advances in Artificial Intelligence-SBIA 2012, pp. 72–81. Springer (2012)
9. SpolaôR, N., Cherman, E.A., Monard, M.C., Lee, H.D.: A comparison of multi-label feature selection methods using the problem transformation approach. Electronic Notes in Theoretical Computer Science **292**, 135–151 (2013)

10. Sucar, L.E., Bielza, C., Morales, E.F., Hernandez-Leal, P., Zaragoza, J.H., Larrañaga, P.: Multi-label classification with bayesian network-based chain classifiers. Pattern Recognition Letters **41**, 14–22 (2014)
11. Tsoumakas, G., Katakis, I., Vlahavas, I.: Random k-labelsets for multilabel classification. IEEE Transactions on Knowledge and Data Engineering **23**(7), 1079–1089 (2011)
12. Tsoumakas, G., Katakis, I., et al.: Multi-label classification: An overview. International Journal of Data Warehousing and Mining (IJDWM) **3**(3), 1–13 (2007)
13. Witten, I.H., Frank, E., Hall, M.A., Pal, C.J.: Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann (2016)
14. Xu, H., Xu, L.: Multi-label feature selection algorithm based on label pairwise ranking comparison transformation. In: Neural Networks (IJCNN), 2017 International Joint Conference on. pp. 1210–1217. IEEE (2017)
15. Zaragoza, J.H., Sucar, L.E., Morales, E.F., Bielza, C., Larranaga, P.: Bayesian chain classifiers for multidimensional classification. In: IJCAI. vol. 11, pp. 2192–2197 (2011)
16. Zhang, M.L., Peña, J.M., Robles, V.: Feature selection for multi-label naive bayes classification. Information Sciences **179**(19), 3218–3229 (2009)
17. Zhang, M.L., Zhou, Z.H.: Ml-knn: A lazy learning approach to multi-label learning. Pattern recognition **40**(7), 2038–2048 (2007)