Accuracy Based Feature Ranking Metric for Multi-Label Text Classification

Muhammad Nabeel Asim
Al-Khwarizmi Institute of Computer Science,
University of Engineering and Technology,
Lahore, Pakistan

Abdur Rehman
Department of Computer Science,
University of Gujrat, Pakistan

Umar Shoaib
Department of Computer Science,
University of Gujrat, Pakistan

***Abstract*—In many application domains, such as machine learning, scene and video classification, data mining, medical diagnosis and machine vision, instances belong to more than one categories. Feature selection in single label text classification is used to reduce the dimensionality of datasets by filtering out irrelevant and redundant features. The process of dimensionality reduction in multi-label classification is a different scenario because here features may belong to more then one classes. Label and instance space is rapidly increasing by the grandiose of Internet, which is challenging for Multi-Label Classification (MLC). Feature selection is crucial for reduction of data in MLC. Method adaptation and data set transformation are two techniques used to select features in multi label text classification. In this paper, we present dataset transformation technique to reduce the dimensionality of multi-label text data. We used two model transformation approaches: Binary Relevance, and Label Power set for transformation of data from multi-label to single label. The Process of feature selection is done using filter approach which utilizes the data to decide the importance of features without applying learning algorithm. In this paper we used a simple measure (ACC2) for feature selection in multi-label text data. We used problem transformation approach to apply single label feature selection measures on multi-label text data; did the comparison of ACC2 with two other feature selection methods, information gain (IG) and Relief measure. Experimentation is done on three bench mark datasets and their empirical evaluation results are shown. ACC2 is found to perform better than IG and Relief in 80% cases of our experiments.**

***Keywords*—Binary relevance (BR); label powerset (LP); ACC2; information gain (IG); Relief-F (RF)**

I. Introduction

A feature is a measurable characteristic or property of the observed process. Text data is high dimensional in nature, and a moderate sized dataset may contain thousands of features. Multi-label is another important property of text data; i.e. a document can belong to none, one or more than one classes. In single label classification, documents belong to only one label (class) but in multi-label classification, which is a case in real world scenario like web pages, newspapers, sports magazine, data mining etc., a document can belong to more than one class that has become recent research topic [1] . Feature selection (FS) is a data pre-processing step in many machine learning applications, which plays an important role in reduction of dimensionality [24]. It helps in mitigating the computational requirements and understanding data. FS removes dimensionality by filtering out irrelevant features, thus improving the prediction capability of a classifier. Researchers evaluate the integrity of feature selection in two ways, individual and subset evaluation [12], [5]. Individual evaluation is computationally efficient it evaluate and assign the weights (ranks) to features (variables) according to their prediction ability in classification. It ignores the inter-dependency of features and also incapable of removing redundant features [21]. Subset evaluation handles redundancy and relevance of features, but it requires higher computational power. The main objective of feature selection is to select subset of features having stronger discrimination power [19]. It reduces effects of redundancy and noise variables by keeping only the features which are efficient for prediction [3].

If two features are extremely correlated as to showing dependence on each other, only one feature is sufficient for data description [17]. Dependent features give no extra information about data. The goal of feature selection is to obtain total information from fewer unique features containing maximum discrimination about the classes. In some applications, due to lack of information about the observed process, features having no correlation with the class act as noise. Such feature produce bias in classification process. Classifier efficiency is enhanced by feature selection techniques which give some cognizance about data and the process being observed.

From machine learning perspective to remove irrelevant features, feature selection criterion is required, which takes into account relevance of each feature with the output class. Irrelevant features lead to poor generalization of the predictor. Feature selection is not some dimensionality extraction technique like principle component analysis (PCA) [2], [20]. Since discriminative features may be independent of all the data, so a procedure called pruning is introduced after feature selection to find the subset of optimal features. To evaluate all the subset of features of size 2N, problem become NP-hard which is difficult to solve in polynomial time that's why a sub-optimal solution is incorporated which can eliminate redundant features with malleable computations. Subset feature selection deals with the scenario that some subset of features are selected while all others are ignored.

Recent research categorizes the multi-label classification into two broad domains: problem transformation and method adaptation. The former first converts multi-label data into single label data and then single label classification techniques are applied, while in latter case single label classifiers are extended to cope with multi-label data.

In multi-label text classification domain for the first time

we introduce a well known feature selection technique ACC2, widely used in single label text classification for feature selection. In this paper e present a single label feature selection approach named ACC2 which is applied in conjunction with Binary Relevance and label power set. The presented technique is very fast and accurate compared to other two feature selection methods (IG, RF). To change the multi-label data into single label we use Binary Relevance (BR) and Label Powerset (LP) techniques.

BR transforms the original dataset into L datasets where L is the number of labels associated with the dataset. Each new dataset contains all the instances as in original dataset, but with only one class associated with each instance; and each of label value has only two states being either positive or negative. BR normally doesn't take into account the features correlation and fails to predict label ranking but it is light weight and reversible. Other advantage of BR is that independent features can be added or removed in model without disturbing rest of the model. In LP approach new classes are generated using possible combination of labels and then problem is solved using single label multi-class approach.

Remaining paper is distributed as: Related work is discuss in Section II. In Section II-B, we describe two label transformation methods and their basic theory. Basic concepts related to feature selection and its importance is discussed in Section III. Section VII introduces benchmark multi-label datasets and their statistics, while Section VI presents the most frequently used evaluation measures for multi-label learning. Results of feature selection algorithms on benchmark datasets are discuss in Section VIII.

II. Related Work

Feature selection is widely use to reduce the dimensionality of data. A number of comprehensive publications can be found on supervised, semi-supervised and non-supervised machine learning topics relating to features selection and classification domains [11], [12], [4], [18]. Multi-label feature selection approach using Relief and Information Gain (IG) is discussed in [13]. A novel approach which jointly performs feature selection with classification called the joint feature selection and classification for multi-label learning (JFSC) is proposed by [14]. Distribution based feature selection measure Chi square is used with label power set as a problem transformation technique [15]. Ensemble embedded feature selection (EEFS) a novel technique is propose by [16], , , , , they develop this method for the feature selection of multi-label clinical data.

To deal with multi-label classification variety of classifiers exist such as Ada-boost [26], BP-MLL [27], SVM [5], ML-KNN [25] each classifier has its own importance but ML-KNN is mostly preferred in most of the research work. In ML-KNN method Ecludean distance is measured between the unlabeled test example and the other instance of the training data set, then using the concept of maximum a posteriori (MAP) label for the test example is selected.

A. Multi-Label Learning

According to [5] multi-label learning has two categories: Multi-label Classification (MLC) and Label Ranking (LR). MLC is defined as a function $h_{MLC} : \chi \rightarrow 2^L$ where χ is an e-dimensional feature space and $L = \{\lambda_1, \lambda_2, \ldots, \lambda_r\}$ is an output space of $r > 1$ labels. Each subset of L is called label-set. If an input instance is given to classifier or predictor it will give a set of relevant labels, Y, and irrelevant labels, $\bar{Y}$. Hence, a bipartition of labels is obtained which is partitioning labels into relevant and irrelevant features. Generally speaking multi-class classification is a special case of MLC where $h_{MC} : \chi \rightarrow L$ while in binary classification $h_B : \chi \rightarrow \{0, 1\}$.

In Label ranking a function $f : \chi \times L \rightarrow R$ that returns ordering of all possible labels according to the relevance of labels in response to an input instance x. Thus a label λ_1 is ranked higher than other label λ_2 if it satisfies $f(x, \lambda_1) > f(x, \lambda_2)$. A rank function,$\tau_x$, maps the classifier real output values to the position of label in ranking, $\{1, 2, \ldots, r\}$. Hence, lower the position the better the label rank i.e. $f(x, \lambda_1) > f(x, \lambda_2) \Rightarrow \tau_x(\lambda_1) < \tau_x(\lambda_2)$. Fig. 1 [6] describes the basic taxonomy for feature selection in multi-label classification.

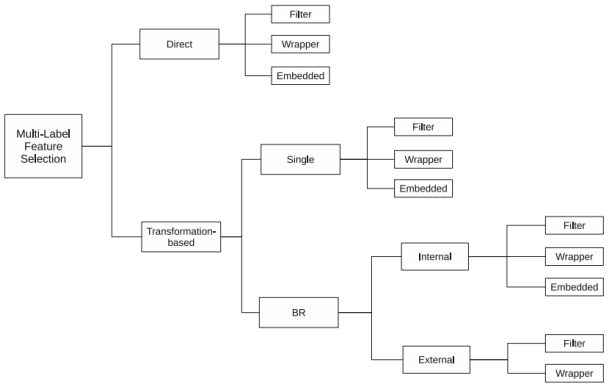

Fig. 1: Taxonomy for feature selection in multi-label classification.

B. Data Transformation Methods

Let X is an e-dimensional input space of numerical features. $L = \{\lambda_1, \lambda_2, \ldots, \lambda_r\}$ is an output space of $r > 1$ labels. A relation of features and labels is given as (x, Y) where $x = x_1, x_2, \ldots, x_e$, which is an e-dimensional instance associated to L set of labels as $Y_i \subseteq C$. Where $Y = \{y_1, y_2, \ldots, y_r\} = (0, 1)^r$ here Y is r-dimensional binary vector and label of each element is 1 if it is relevant, 0 otherwise. Table I shows the comparison of single label (binary, multi-class) data with multi-label one.

TABLE I: Single Label vs. Multi-label Dataset

Instances	Features	$Single-Label$ Binary	$Single-LabelMulti-Class$	$Multi-Label$				
		y ∈L ={0,1}	y ∈L ={λ_1,λ_2,λ_3,λ_4}	y_1	y_2	y_3	y_4	Y⊆L
1	f_1	0	λ_2	0	1	1	1	{λ_2,λ_3,λ_4}
2	f_2	1	λ_3	0	0	1	1	{λ_3,λ_4}
3	f_3	0	λ_4	1	1	0	0	{λ_1,λ_2}
4	f_4	1	λ_1	1	0	1	1	{λ_1,λ_3,λ_4}

Multi-label learning is categorized into two groups: method adaptation in which existing single label classifier models are enhanced to deal with multi-label data directly while second one is problem transformation methods which transform the multi-label problem into several binary classification problems (BR) or into different possible combinations of label set (LP).

C. Binary Relevance (BR)

It is like one-versus-all (OVA) approach, it generates one dataset for each label, in new generated dataset positive patterns represent the presence of a particular class label and all other patterns are set to negative. BR transforms the original dataset in to L datasets. Each new dataset contains all the instances as in original dataset, but with only one class; and each of feature value has only two states being either positive or negative. In the i_{th} dataset, if label set for an instance contains the i^{th} label then its label is positive otherwise negative. For classifying new pattern, it is assigned a class label by all the L datasets and the union of labels is the predicted label set. Although BR settles linearly with label set L of r dimensions; but it does not consider the correlation of labels.

Table III shows binary relevance (BR) based transformation of data from multi-label to single label when applied to the dataset of Table II.

TABLE II: Multi-label Dataset Example

Instances	Features	Label set
1	f_1	$\{\lambda_1,\lambda_3\}$
2	f_2	$\{\lambda_4\}$
3	f_3	$\{\lambda_1,\lambda_2,\lambda_3\}$
4	f_4	$\{\lambda_1,\lambda_2\}$

TABLE III: Dataset After BR Based Transformation

Instances	Label set
1	λ_1
2	- λ_1
3	λ_1
4	λ_1

Instances	Label set
1	- λ_2
2	- λ_2
3	λ_2
4	λ_2

Instances	Label set
1	λ_3
2	- λ_3
3	λ_3
4	- λ_3

Instances	Label set
1	- λ_4
2	λ_4
3	- λ_4
4	- λ_4

D. Label Power-set (LP)

In this approach each distinct combination of labels present in training set is treated as different class and then single-label classification is performed on the transformed data. Although this approach makes the task easy but with the increase in classes, label-set size also increases; hence increasing the computational cost and causes impediment in learning. The number of examples for training of each label set will be very small. To settle this problem, initial set of labels are split up into small random subsets of labels (label-sets). LP is performed on these label sets. This approach is called RAKEL, random k label sets, where k parameter specifies the size of label sets. Unlike BR, LP considers the correlation between labels. Table IV represents dataset formed after transformation using label power-set.

TABLE IV: Transformed Dataset using the Label Powerset Method

Instances	Labelset
1	$(\lambda_{1,3})$
2	(λ_4)
3	$(\lambda_{1,2,3})$
4	$(\lambda_{1,2})$

Let $C = \{\omega_i : i = 1, 2, \ldots, L\}$ be a finite set of classes, x_i is a, instance linked with set of labels Y_i where $Y_i \subseteq C$. A label set such that $S \subseteq C$ and $k = |S|$, is called $k - labelset$. Commonly used label sets are:

- Disjoint label-sets.
- Overlapping label-sets.

In disjoint label-set, each label set is of size k, and all label sets are disjoint; class label set C is randomly segregated into $l = [\frac{L}{k}]$ label sets, S. While in overlapping case label sets may overlap; C^k, overlapping label sets is the set of distinct $k - labelsets$ where $|C^k| = \binom{n}{k}$. To classify a new instance z, every classifier, h_i gives a binary prediction for each label in relative label-set, S_i. Nevertheless after the transformation it is possible to have limited number of combinations for new classes, hence producing sample imbalance issue.

III. Feature Selection

In this section basic concepts related to feature selection and its importance is discussed. In FS we cater the best features, which relatively provides more information of instance category to the classifier. In FS we find most suitable subset of features $X' \subseteq X$ that may enhance prediction capability of the classifier. There are basically three FS approaches: filter, wrapper and embedded. We discuss each one with detail.

A. Wrapper Approach

Wrapper methods find the most suitable subset of relevant features using the classification/learning algorithm; it offers high computation cost as it has to run classification task for each subset of features. As the number of features increases, the classification is required more often to find the suitable subset of features; thus giving arise to polynomial time tough scenario. To overcome the computational burden and to find most suitable subset of features, searching algorithm are incorporated.

There are different search algorithms for feature selection, each having its pros and cons. Tree structure is used in branch and bound approach [10] for selection of features; its complexity increases exponentially with increase in number of features. For large datasets with a huge number of features, exhaustive search approach is not appropriate. There are feasible linear approaches which yield good result with lesser computation cost i.e. sequential search, particle swarm optimization, genetic algorithm and heuristic search algorithms. Wrapper methods further split up into two categories: sequential search and heuristic search algorithms.

Sequential search algorithm continue to add/remove features until a maximum objective function is reached. A criterion is set whose objective is to maximize the objective function with minimal number of features. Sequential search algorithms are iterative in nature.

Sequential feature selection algorithm starts with an empty set; accumulate a single feature that yields maximal value for objective function. Wrapper approach necessitate the learning algorithm to find suitable set of features, but it is inclined towards finding the set of features which are more suitable for a particular learning algorithm; a rigorous computation power is required for the wrapper approach.

B. Embedded Approach

Embedded approach integrates feature selection with the training algorithm as some part of the process, like decision trees; selection of best features, having paramount discriminative power to differentiate among classes, at each stage.

C. Filter Methods

Filter methods selects sets of optimal features based on the peculiarity and idiosyncrasy of the dataset; irrelevant features are filtered out, this whole process is separate from the learning phase/algorithm. Variable ranking technique is the major method used in filters for feature selection in ordered form. Ranking methods are versatile thats why they hugely contribute to the practical applications. A particular ranking measure is used to rank the features with respect to some threshold; features below this threshold are discarded.

Basic trait of a relevant/distinctive feature is that it preserves the necessary information about classes present in the dataset. This trait is the relevance of feature necessary for segregation of distinct classes. But how could feature relevancy be described by current standards? Different researchers describe it differently. In [7] author defines an irrelevant feature as: "an irrelevant feature is conditionally independent of class labels". This fact depicts that a relevant feature can not be independent of class labels, but it can be independent of input data. This also suggests that relevant features have a certain amount of influence on the classes, if not then they should be considered as irrelevant. One most important parameter in determining the feature relevancy is feature correlation between features and classes; which describes a feature's importance to discriminate classes.

In this paper, we used ACC2 feature selection measure on multi-label text data and compared with two other well known filter based methods (Relief F and Information Gain). In Sections III-C1, III-C2 and III-D, we discuss these techniques in detail.

1) Relief F measure: It is heuristic approach developed by [8] removes the irrelevant features from the datasets. It is the extension of basic Relief algorithm [9]. Relief is capable of dealing with discrete as well as continuous attribute but it can't deal with multi-class problems. It estimates features on the basis of discrimination power value of attributes among the instances. Relief F seek for k nearest misses $M_j(C)$, $j = 1 \dots k$, for each class C. Calculate the weight/estimate by taking average contribution of each class.

$$W[A] = W[A] - \sum_{j}^{k} \frac{diff(A, R, H_j)}{n \times k} + \sum_{C \neq class(R)} \sum_{j}^{k} \left[\frac{P(C)}{1 - P(class(R))} \times \frac{diff(A, R, M_j)}{n \times k} \right] \quad (1)$$

In above equation R is a randomly selected instance, for which Relief searches for its two nearest neighbors: one from the same class, called nearest hit H, and the other from the different class, called nearest miss M. It updates the quality estimation W [A] for all attributes A depending on their values for instance R, M and H . If instances R and H have different values of the attribute A then the attribute A separates two instances with the same class which is not desirable so we decrease the quality estimation W [A]. In (1) different function calculates the difference between two instances on the basis of nearest hit and nearest miss.

Basic idea about the working is that it separates classes pair on the basis of features regardless the fact that which two classes are nearest to each other.

2) Information Gain (IG): Information gain represents dependency of input labels with the class labels. It is defined by well-known equation of Shannon's about entropy:

$$H_{entropy}(Y) = \sum_{y} p(y) log(p(y)) \quad (2)$$

Actually entropy is the uncertainty in output label Y. Hence entropy in output, given input labels is:

$$H_{entropy}(y|x) = \sum_{x} \sum_{y} p(x, y) log(P(y|x)) \quad (3)$$

By already knowing the input labels we can predict output label Y with more accuracy. Hence IG relates the dependency of input label X to output label Y given as:

$$I(X, Y) = H_{entropy}(Y) - H_{entropy}(y|x) \quad (4)$$

D. ACC2 Feature Selection Measure

Accuracy measure (ACC) is a well known feature selection technique widely used in single label text classification. It is simply the difference of true positives and false positives of a term. It works well in balanced dataset but perform poorly on unbalanced dataset because this algorithm is biased toward tp.

Balanced Accuracy measure($ACC2$) is an enhanced version of accuracy measure (ACC)[22]. ACC2 is the absolute difference of true positive rate (tpr) and false positive rate (fpr). As tpr is normalized; obtained after division with the class size; it solves the problem of biasing toward tp. In multi label text classification we, for the first time, use this simple technique for feature selection. Formulae of ACC and ACC2 are given in (5) and (6,) respectively.

$$Accuracy\ Measure = ACC = |t_p - f_p| \quad (5)$$

$$Balanced\ Accuracy\ Measure = ACC = |t_{pr} - f_{pr}| \quad (6)$$

$$tpr = \frac{t_p}{t_p + f_n} \quad (7)$$

$$tpr = \frac{t_p}{t_p + f_n} \quad (8)$$

IV. Proposed Methodology

In multi-label text classification, we present a well known feature selection measure ACC2; which is widely used in single label text classification. We compare the performance of ACC2 with two (Information gain, Relief-F) other feature selection measures. We first use Binary Relevance (BR) and Label Power-Set for data transformation. To reduce the dimensionality of data we did feature selection.

Description of feature selection methods with transformation techniques is given below:

1) ACC2-BR: ACC2 as feature selection measure based on BR
2) ACC2-LP: ACC2 as feature selection measure based on LP
3) RF-LP: RF as feature selection measure based on LP
4) RF-BR: RF as feature selection measure based on BR
5) IG-BR: IG as feature selection measure based on BR
6) IG-LP: IG as feature selection measure based on LP

Relief-F is a univariate feature selection measure; it demarcates or evaluate the quality of features of single label datasets. Relief-F award different score for features having different values on different classes but castigates features having different values for the same class.

Information gain used the entropy measure between labels and features showing dependency between features and labels (classes). Features having greater values of IG are ranked higher. Entropy is the impurity present in the instances/examples, while information gain is an average reduction in entropy in accordance with a given feature. Higher the value of IG, better is the dependence between features and classes.

Balanced accuracy measure is most widely used algorithm in single label text classification. It takes the absolute difference of true positive rate (tpr) and false positive rate (fpr). Detailed expressions of three feature selection measures are given in Section III.

RF-BR, IG-BR and ACC2-BR first transform the multi-label dataset into single label datasets using binary relevance transformation, then feature selection methods RF, IG and ACC2 are applied to select the highly discriminative features among the classes. But in these methods, as the BR does not consider the correlation between labels during transformation, the same problem exist in these approaches.

In RF-LP, IG-LP and ACC2-LP methods the process of feature selection is done after transformation of data from multi-label to singe label using label power-set technique. Data transformation techniques are described in Section II-B.

After feature selection the process of classification is done using ML-KNN classifier. We use four well known evaluation measures (Hamming Loss, Subset accuracy, Micro and Macro average F measure) to estimate the accuracy of three feature selection algorithms.

V. Motivation Example

This section discusses the working of six feature ranking metrics with the help of an example. Table V is a sampled dataset presented only for illustration and comparison of different metrics based on problem transformation. We have 15 documents belonging to 3 classes and 10 terms/features. We practically show that multi-label data after transformation to single label becomes highly unbalanced. It is not a problem in single label feature selection regime. In multi-label classification due to multi label to single label transformation problems do exist; as binary relevance does not take into consideration the label dependency. On the other hand, LP only considers the distinct label-sets. It is, therefore, unable to predict new label-sets, causing over-fitting of training data. However, these techniques are light weight giving results almost comparable to problem adaptation techniques.

Table VI shows comparison of six ranking metrics and scored assigned by these metrics to features. In multi-label datasets, features can have relevance with more than one classes. So it is very difficult to judge the discrimination power of particular feature with respect to class labels. So many factors are to consider in multi-label domain for rank assignment. As can be seen that IG-BR and ACC2-BR assigned first rank to f_{10} while RF-LP and RF-BR assigned first rank to f_4. From V, one can estimate that f_{10}, f_9 and f_8 are more important as they highly match with three classes. But RF-LP and RF-BR assigned the first rank to f_4. Other metrics assigned lower ranks to this feature. In multi-label domain, features correlation between themselves and with all the class labels should also be considered.

TABLE V: Artificially Sampled Dataset for Multi-label

S no	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8	f_9	f_{10}	C_1	C_2	C_3
1	1	0	1	1	0	0	1	0	1	1	1	1	1
2	1	1	1	1	0	0	1	0	0	1	1	1	0
3	0	0	1	1	1	0	1	0	0	0	0	0	0
4	1	0	1	0	0	1	1	0	0	1	1	1	0
5	1	0	1	0	0	1	0	1	1	1	0	0	1
6	1	0	0	1	1	0	0	0	0	0	0	0	0
7	1	1	1	0	0	0	0	1	0	0	1	1	1
8	1	0	0	0	1	1	0	1	1	0	1	1	1
9	0	0	0	1	0	0	0	1	1	1	0	0	0
10	1	1	0	0	0	1	0	1	1	0	1	1	1
11	0	1	0	0	0	1	0	1	1	1	1	1	1
12	0	1	1	0	0	0	1	1	1	1	1	1	1
13	1	1	1	0	0	1	0	1	1	1	1	1	1
14	1	1	1	0	1	0	0	1	1	1	1	1	1
15	1	1	0	0	0	1	1	1	1	0	0	1	1

TABLE VI: Comparison of Rank Assignee Metrics to Features on Sampled Dataset

term	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8	f_9	f_{10}
RF-LP	-0.0375	-0.066	0.0797	-0.02	0.0043	-0.066	0.181	0.291	0.538	0.516
RF-BR	0.0271	0.2525	-0.0348	0.2909	0.0293	-0.0053	-0.0388	0.137	0.137	-0.0303
IG-BR	0	0.1256	0	0	0	0	0	0	0.1216	0.3758
IG-LP	0	0	0	0	0	0	0	0	0.734	0.61
ACC2-BR	0.0742	0.0818	0.2454	0.1424	0.206	0.1651	0.1681	0.2424	0.556	0.6727
ACC2-LP	0.1674	0.167	0.2929	0.1565	0.1459	0.1224	0.2525	0.3314	0.3995	0.3641
Rank RF-LP	9	4	7	1	10	6	5	3	2	8
Rank RF-BR	6	2	9	1	5	7	10	4	3	8
Rank IG-BR	10	9	3	4	5	6	7	8	2	1
Rank IG-LP	9	10	3	4	5	6	7	8	1	2
Rank IG-ML	7	1	9	2	6	5	8	3	4	10
Rank Acc-BR	10	9	3	8	5	7	6	4	2	1
Rank Acc2-LP	6	7	4	8	9	10	5	3	1	2

VI. Evaluation Measures

Evaluation measures used for multi-label classification are different from those used for single label classification. Evaluation Measures fall into two categories: label based and example based. Label based is an extended form of evaluation measures used for single label classification domain. Example based is specifically built for multi-label domain [28]. Here we give the expressions of evaluation measures used for multi-label classification. In all below evaluation measures x is label predicted KNN classifier and y is actual or true label.

$$Hamming\,loss(x_i, y_i) = H_{loss} = \frac{1}{N}\sum_{i=1}^{N}\frac{|x_i \Delta y_i|}{L} = \frac{1}{N}\sum_{i=1}^{N}\frac{Xor(x_i, y_i)}{L} \quad (9)$$

Hamming loss is an average measure of difference between actual and predicted value for labels. A low value of hamming loss is required to show better classification performance.

$$Accuracy = \frac{1}{N}\sum_{i=1}^{N}|\frac{x_i \cap y_i}{x_i \cup y_i}| \tag{10}$$

Accuracy is the closeness of the measure value to the known standard value. It is a fraction of correctly classified instances to the total number of instances to be classified. In multi-label classification accuracy of a metric is measure using above equation.

$$Precision = \frac{1}{N}\sum_{i=1}^{N}|\frac{x_i \cap y_i}{x_i}| \tag{11}$$

Precision is the fraction of correctly classify instances to the total number of instances to be classify.

$$Recall = \frac{1}{N}\sum_{i=1}^{N}|\frac{x_i \cap y_i}{y_i}| \tag{12}$$

Recall shows the fraction of number of correct instances to the total number of retrieved instances.

$$Subset\ accuracy = \frac{1}{N}\sum_{i=1}^{N} I(x_i = y_i) \tag{13}$$

Subset accuracy or classification accuracy is defined by (10). It is very strict requirement, as it is the average of set of predicted labels exactly matching the set of actual labels.

$$F_1 - Measure = \frac{1}{N}\sum_{i=1}^{N} 2 \times \frac{|x_i \cap y_i|}{|x_i| + |y_i|}| \tag{14}$$

F_1 measure is a single measure obtained by combining two evaluation measures precision and recall. It is use to make trade off between precision and recall.

$$F_a(Macro\ averaged) = \frac{1}{q}\sum_{i=1}^{q}\frac{2t_p}{2t_p + f_p + f_n} \tag{15}$$

In macro F_1 measure we calculate the precision and recall of each set and take there average.

$$F_a(Micro\ averaged) = \frac{\sum_{i=1}^{q} 2t_p}{\sum_{i=1}^{q} 2t_p + \sum_{i=1}^{q} f_p + \sum_{i=1}^{q} f_n} \tag{16}$$

In micro F_1 measure we find the t_p, f_p and f_n of all the available sets and then apply them in (16) to calculate the final score. In equation q represents the available sets. A high value of accuracy and other evaluation criterion is required to show better classification performance, except for hamming loss metric.

VII. Experimental Setup and Datasets

We performed experiments on three benchmark text datasets given in Table VII. Preprocessing, such as stemming and stop word removal was already done on these data sets available at $(mulan\ dataset)$. We used Java platform for experimentation. Transformation of data from multi-label to single label is done using Binary Relevance (BR) and Label Powerset (LP) techniques. After data transformation feature selection algorithms are applied to reduce the dimensionality of data. The process of classification is done using ML-KNN classifier. The performance of feature selection algorithms is measure on percentage (10%, 20%, 30%, 40%, 50%,60%, 70%, 80%) of top ranked features selected by every algorithm. We used five (Hamming Loss, Ranking Loss, Subset accuracy, Micro and Macro average measure) evaluation measures to test the performance of six feature selection algorithms at different test points of data.

TABLE VII: Description of Datasets

Dataset	N	M	$\|L\|$	LC	LD	DC
bibtex	7395	1836 d	159	2.402	0.02	2856
Enron	1702	1001 d	53	3.38	0.06	753
medical	978	1449d	45	1.25	.03	94

Table VII shows benchmark datasets that are used in experimental evaluation for feature selection. Table also represents the characteristics of six datasets, such as number of instances (N); number of features (F); number of class labels (L); the label cardinality (LC); label density (LD); and distinct combinations of labels (DC).

$$Label\ Cardinality = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{L} y_j^{(i)} \tag{17}$$

Label cardinality (LC) shows the average number of labels per example/instance. It can be calculated using above equation. In (17) N is number of instances and L represent number of labels in a sample.

$$Label\ Density = \frac{\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{L} y_j^{(i)}}{L} \tag{18}$$

Label density (LD) is normalized form of LC shown in (18).

For each dataset D, feature reduction measure for feature selection can be calculated from (19).

$$Feature\ Reduction(D, X') = 100 - \frac{100 \times X'}{M} \tag{19}$$

Where, X' is the feature subset obtained after feature selection from dataset D; M is the number of examples. Six feature selection techniques are performed on each dataset. classifier response is evaluated for features that are selected.

VIII. RESULTS

We applied six FS methods and five evaluation measures on three benchmark datasets. Tables VIII, XII and XVI shows the hamming loss measure for described datasets. Hamming loss is the relative frequency of predicted and actual labels as previously shown in (9). Subset accuracy (10) is another measure, which tell that either a predicted label is the actual true label or not. Micro averaged precision results are shown in Tables X, XIV and XVIII. In micro averaged precision large classes dominate over small classes, as it is the fraction of true positives and $tp + tn$ of all concerned classes. F1 measure is the harmonic mean of precision and recall, it considers the true positives and ignores the true negatives but this measure assigns equal weight to precision and recall. Whereas precision is the number of actual correct results out of the marked correct results by the classifier $\frac{t_p}{t_p+f_p}$; and 'recall' is the fraction of correct results out of all the correct results $\frac{t_p}{t_p+f_n}$[23]. Macro average measure is more biased towards average recall than average precision. Label based micro average criterion is biased towards most populated labels, while macro average is the average of tp and fp for each class separately. Macro averaging is biased to least populated classes.

A. Enron Dataset

Enron dataset is a test bench dataset available at $(mulan\ dataset)$, having 1702 instances and 53 labels with cardinality 3.78. Tables VIII to XI show the experiments done on Enron dataset and in next subsections we discuss their results based on different measures.

1) Hamming Loss: Table VIII shows the hamming loss for six feature ranking measures based on filter approach on Enron dataset. Hamming loss is computed for different data test points for selected features. Least-BR shows those BR problem transformation based feature ranking measures having least hamming loss; Least-LP shows those feature ranking measures having least hamming loss among other measures for LP transformation case. As can be seen from Table VIII, ACC2 produces the least hamming loss both in BR as well as LP transformation case. It is the simplest technique among all described approaches.

TABLE VIII: Feature Ranking Metrics Having Least Hamming Loss using KNN Classifier

Hamming Loss								
Features	10%	20%	30%	40%	50%	60%	70%	80%
IG-BR	0.063	0.0646	0.066	0.068	0.0663	0.0645	0.0614	0.0635
RF-BR	0.0617	0.0598	0.0614	0.0635	0.0616	0.0613	0.0618	0.0614
ACC2-BR	0.0613	0.0605	0.0613	0.0633	0.0613	0.0637	0.063	0.0612
Least-BR	ACC2-BR	RF-BR	ACC2-BR	ACC2-BR	ACC2-BR	RF-BR	IG-BR	ACC2-BR
IG-LP	0.0613	0.059	0.0629	0.0624	0.0616	0.0619	0.0617	0.0618
RF-LP	0.0622	0.0604	0.0622	0.0623	0.0617	0.0618	0.0623	0.0623
ACC2-LP	0.0599	0.0625	0.0619	0.0618	0.0621	0.0625	0.063	0.0617
Least-LP	ACC2-LP	IG-LP	ACC2-LP	ACC2-LP	IG-LP	RF-LP	IG-LP	ACC2-LP

2) Subset Accuracy: Subset accuracy values of six feature ranking metrics are given in Table IX. Max-BR shows the occurrence of a measure, among three other measure based on BR problem transformation approach, having maximum subset accuracy value. In same way Max-LP shows a measure having the maximum subset accuracy value among other techniques based on LP transformation approach. Clearly, ACC2 measure subset accuracy is leading to all other techniques.

TABLE IX: Subset Accuracy Values for Feature Ranking Metrics using KNN Classifier on Enron Dataset

Subset Accuracy								
Features	10%	20%	30%	40%	50%	60%	70%	80%
IG-BR	0.0039	0.0235	0.0215	0.0196	0.0274	0.0274	0.0313	0.0352
RF-BR	0.0333	0.0215	0.0254	0.0215	0.0274	0.0215	0.0372	0.0411
ACC2-BR	0.045	0.0274	0.0274	0.0196	0.0313	0.0275	0.0294	0.0333
Max-BR	ACC2-BR	ACC2-BR	ACC2-BR	RF-BR	ACC2-BR	ACC2-BR	RF-BR	RF-BR
IG-LP	0.0391	0.0333	0.0254	0.0235	0.0313	0.0313	0.0254	0.0301
RF-LP	0.0333	0.0235	0.0196	0.0254	0.0274	0.0235	0.0254	0.0294
ACC2-LP	0.0294	0.0196	0.0294	0.0255	0.0294	0.0235	0.0372	0.0303
Max-LP	IG-LP	IG-LP	ACC2-LP	ACC2-LP	IG-LP	IG-LP	ACC2-LP	ACC2-LP

3) Micro and macro averaged F1-score: In Table X, we compare different feature ranking criterion based on BR and LP transformation approaches at different number of selected features. RF-BR performed better than IG-BR and ACC2-BR in micro-averaged case in BR domain. In LP domain, ACC2-LP is leading while IG-BR performed poorer. RF-BR approach outperformed in macro-averaged case while ACC2-LP outperformed in LP case. Whereas IG-BR as well as IG-LP underperformed in both micro and macro cases.

4) Ranking loss: Table XI shows ranking loss for different feature selection criteria. ACC2-BR has the least ranking loss for 30, 50, 60 and 80 percent of selected features. For 20, 40 and 70 percent of test points RF-BR has the least ranking loss; IG-BR has only least ranking loss at 10 percent of selected data points. For LP case, RF-LP and ACC2-LP has three times least ranking loss, while IG-LP has two times least ranking loss. Hence, overall ACC2 method outperformed for LP and BR cases.

B. Medical Dataset

Results for experiments of different measures on Medical dataset are presented in Tables XII to XV. Subsequent section present discussion of these measures.

1) Hamming loss for medical dataset: hamming loss for different metrics of medical dataset is given in Table XII. ACC2 has the least hamming loss for 9 out of 16 cases at different number of selected features. RF has least hamming loss for 4 cases, and IG has least hamming loss in 3 out of 16 cases.

2) Subset accuracy measure for medical dataset: For medical dataset, ACC2-BR has the maximum subset accuracy for 10% to 30% of total number of features (see Table XIII). While for 40% to 80% of total number of features, RF-BR has the maximum accuracy. In LP case, ACC2-LP gives the maximum subset accuracy only for 40% and 70% of features. IG underperformed in medical datasets, while RF technique take the maximum value in 10 out of 16 cases.

3) Micro and macro-average F1-score for medical dataset: In Table XIV combined values for micro and macro-averaged F1- score are given for medical dataset. Out of 16 calculations at different percentages ACC2 take the maximum micro-averaged value for 8 times while IG performed better than RF both in BR and LP case by taking 5 times max values of micro-averaged score. ACC2-BR becomes highest at 20% to 50% of selected features. ACC2-LP becomes highest at 30%, 50%, 70% and 80% of selected features. While IG-BR and IG-LP performed better than RF-BR and RF-LP in macro-average measure in medical dataset.

TABLE X: Micro and Macro-averaged F1-Score Values for Feature Ranking Metrics using KNN Classifier on Enron Dataset

Micro-averaged F-Measure								
Features	10%	20%	30%	40%	50%	60%	70%	80%
IG-BR	0.3066	0.3525	0.3722	0.3474	0.3781	0.4141	0.4405	0.4593
RF-BR	0.436	0.4813	0.4618	0.466	0.4868	0.4835	0.4879	0.4848
ACC2-BR	0.450	0.4413	0.4724	0.4534	0.4722	0.4659	0.4759	0.4871
Max-BR	ACC2-BR	RF-BR	ACC2-BR	RF-BR	RF-BR	RF-BR	RF-BR	ACC2-BR
IG-LP	0.4401	0.4738	0.4755	0.4841	0.4947	0.4853	0.4762	0.487
RF-LP	0.433	0.4721	0.4814	0.4967	0.5053	0.4845	0.492	0.4793
ACC2-LP	0.4504	0.4598	0.4857	0.4994	0.49	0.487	0.4788	0.4982
Max-LP	ACC2-LP	IG-LP	ACC2-LP	ACC2-LP	RF-LP	ACC2-LP	RF-LP	ACC2-LP
Macro-averaged F-Measure								
Features	10	20	30	40	50	60	70	80
IG-BR	0.0859	0.0928	0.0985	0.0915	0.102	0.1194	0.1217	0.1383
RF-BR	0.1212	0.135	0.134	0.133	0.1402	0.1414	0.1432	0.1494
ACC2-BR	0.1047	0.1106	0.1295	0.1363	0.1321	0.1376	0.1458	0.1464
Max-BR	RF-BR	RF-BR	RF-BR	ACC2-BR	RF-BR	RF-BR	ACC2-BR	RF-BR
IG-LP	0.109	0.1238	0.1272	0.1352	0.1434	0.1408	0.1369	0.137
RF-LP	0.1092	0.1243	0.133	0.1409	0.1469	0.1429	0.1371	0.1363
ACC2-LP	0.1145	0.1263	0.1406	0.1434	0.1399	0.1437	0.1409	0.1567
Max-LP	ACC2-LP	ACC2-LP	ACC2-LP	ACC2-LP	ACC2-LP	ACC2-LP	ACC2-LP	ACC2-LP

TABLE XI: Ranking Loss Values for Feature Ranking Metrics using KNN Classifier on Enron Dataset

Ranking Loss								
Features	10%	20%	30%	40%	50%	60%	70%	80%
IG-BR	0.0445	0.0533	0.0628	0.062	0.0618	0.0654	0.0654	0.0654
RF-BR	0.0429	0.0499	0.0592	0.0576	0.0567	0.0577	0.0567	0.0567
ACC2-BR	0.0584	0.0511	0.0581	0.063	0.0564	0.0574	0.0574	0.0564
Least-BR	IG-BR	RF-BR	ACC2-BR	RF-BR	ACC2-BR	ACC2-BR	RF-BR	ACC2-BR
IG-LP	0.0483	0.0536	0.0578	0.0627	0.0608	0.0674	0.0574	0.0574
RF-LP	0.0476	0.0555	0.061	0.0589	0.0604	0.0624	0.0545	0.0545
ACC2-LP	0.0768	0.0677	0.0651	0.0691	0.0601	0.0618	0.0518	0.0618
Least-LP	RF-LP	IG-LP	IG-LP	RF-LP	ACC2-LP	ACC2-LP	ACC2-LP	RF-LP

TABLE XII: Hamming Loss Values for Feature Ranking Metrics using KNN Classifier on Medical Dataset

Hamming Loss								
Features	10%	20%	30%	40%	50%	60%	70%	80%
IG-BR	0.0097	0.0098	0.0101	0.0101	0.0099	0.0095	0.0095	0.0095
RF-BR	0.0108	0.0103	0.0098	0.0102	0.0102	0.0102	0.0102	0.0102
ACC2-BR	0.0103	0.0096	0.001	0.01	0.0098	0.0094	0.0094	0.009
Least-BR	IG-BR	ACC2-BR	ACC2-BR	ACC2-BR	ACC2-BR	ACC2-BR	ACC2-BR	ACC2-BR
IG-LP	0.0098	0.0098	0.0097	0.01	0.0099	0.0096	0.0096	0.0096
RF-LP	0.0097	0.0099	0.0099	0.0096	0.0093	0.0094	0.0092	0.0092
ACC2-LP	0.0018	0.0106	0.0107	0.0103	0.01	0.01	0.01	0.0091
Least-LP	ACC2-LP	IG-LP	IG-LP	RF-LP	RF-LP	RF-LP	RF-LP	ACC2-LP

4) Ranking loss: ACC2-BR attains least value of ranking loss when we select top 50% to 80% of features in Table XV. RF-BR attains least value at approximately mid point of selected features. For LP case RF and ACC2 attains least values by going side by side while IG performance deteriorates both as compared to RF and ACC2 metrics.

TABLE XIII: Subset Accuracy Values for Feature Ranking Metrics using KNN Classifier on Medical Dataset

Subset Accuracy								
Features	10%	20%	30%	40%	50%	60%	70%	80%
IG-BR	0.6426	0.6689	0.6587	0.6621	0.6621	0.6689	0.6689	0.6689
RF-BR	0.6382	0.6553	0.6558	0.6655	0.6621	0.6655	0.6655	0.6655
ACC2-BR	0.6621	0.6724	0.6621	0.6587	0.6587	0.6587	0.6587	0.6587
Max-BR	ACC2-BR	ACC2-BR	ACC2-BR	RF-BR	RF-BR	RF-BR	RF-BR	RF-BR
IG-LP	0.6758	0.6758	0.6724	0.6587	0.6621	0.6621	0.6621	0.6621
RF-LP	0.6826	0.6826	0.6826	0.666	0.6962	0.6894	0.6797	0.6997
ACC2-LP	0.6041	0.6519	0.6519	0.6719	0.6519	0.6519	0.6819	0.6519
MAX-LP	RF-LP	RF-LP	RF-LP	ACC2-LP	RF-LP	RF-LP	ACC2-LP	RF-LP

C. Bibtex Dataset

Table XVI to XIX discusses results for different metrics on bibtex dataset. Bibtex is a benchmark dataset having 7395 documents and 1836 features having a total size of 7395 $\times$ 1836 with cardinality of 2.402.

1) Hamming loss measure of bibtex dataset: Least value of ACC2 occurred for BR transformation for initially 10% to 30% and then 70% to 80% features among IG-BR and RF-BR techniques. While for 40% to 60% of features, IG-BR attained

TABLE XIV: Micro and Macro-averaged Values for Feature Ranking Metrics using KNN Classifier on Medical Dataset

Micro-averaged F-Measure								
Features	10%	20%	30%	40%	50%	60%	70%	80%
IG-BR	0.8042	0.8254	0.819	0.8207	0.8218	0.8228	0.8318	0.8318
RF-BR	0.8006	0.8137	0.804	0.8179	0.8163	0.8179	0.8179	0.8179
ACC2-BR	0.8127	0.8267	0.8216	0.8186	0.8254	0.8254	0.8254	0.8254
MAX-BR	ACC2-BR	ACC2-BR	ACC2-BR	IG-BR	ACC2-BR	ACC2-BR	IG-BR	IG-BR
IG-LP	0.823	0.8254	0.8266	0.8202	0.8218	0.8206	0.8286	0.8286
RF-LP	0.827	0.8229	0.8243	0.8221	0.8342	0.8211	0.8358	0.8358
ACC2-LP	0.7776	0.8266	0.8082	0.8232	0.8211	0.8251	0.8211	0.8211
MAX-LP	IG-LP	ACC2-LP	IG-LP	ACC2-LP	RF-LP	ACC2-LP	RF-LP	RF-LP
Macro-averaged F-Measure								
Features	10	20	30	40	50	60	70	80
IG-BR	0.5618	0.5649	0.5745	0.5748	0.5761	0.5778	0.5778	0.5778
RF-BR	0.5139	0.5657	0.568	0.5714	0.5709	0.5714	0.5714	0.5714
ACC2-BR	0.5109	0.5679	0.5756	0.5757	0.5764	0.5764	0.5764	0.5764
MAX-BR	IG-BR	ACC2-BR	ACC2-BR	ACC2-BR	ACC2-BR	IG-BR	IG-BR	IG-BR
IG-LP	0.5419	0.5732	0.5769	0.5759	0.5761	0.577	0.577	0.577
RF-LP	0.5639	0.5719	0.5755	0.5765	0.5761	0.5776	0.5707	0.5707
ACC2-LP	0.4261	0.5176	0.5791	0.5662	0.5778	0.5738	0.5738	0.5738
MAX-LP	RF-LP	IG-LP	ACC2-LP	RF-LP	ACC2-LP	RF-LP	ACC2-LP	ACC2-LP

TABLE XV: Ranking Loss Values for Feature Ranking Metrics using KNN Classifier on Medical Dataset

Ranking Loss								
Features	10%	20%	30%	40%	50%	60%	70%	80%
IG-BR	0.0445	0.0533	0.0628	0.062	0.0618	0.0654	0.0654	0.0654
RF-BR	0.0429	0.0599	0.0592	0.0576	0.0567	0.0587	0.0567	0.0567
ACC2-BR	0.0584	0.0511	0.0581	0.063	0.0564	0.0574	0.0564	0.0564
Least-BR	RF-BR	ACC2-BR	ACC2-BR	RF-BR	ACC2-BR	ACC2-BR	ACC2-BR	ACC2-BR
IG-LP	0.0483	0.0536	0.0618	0.0627	0.0608	0.0574	0.0574	0.0574
RF-LP	0.0476	0.0555	0.061	0.0589	0.0607	0.0624	0.0545	0.0545
ACC2-LP	0.0768	0.0677	0.0601	0.0691	0.0605	0.0618	0.0508	0.0618
Least-LP	RF-LP	IG-LP	ACC2-LP	RF-LP	ACC2-LP	IG-LP	ACC2-LP	RF-LP

the least hamming loss value. On the other hand RF-BR did not take least value of hamming loss measure in BR domain. For LP case ACC2 and RF-BR generated the least hamming loss values, as shown in Table XVI.

TABLE XVI: Hamming Loss Values for Feature Ranking Metrics using KNN Classifier on Bibtex Dataset

Hamming Loss								
Features	10%	20%	30%	40%	50%	60%	70%	80%
IG-BR	0.0132	0.0137	0.0141	0.014	0.0143	0.0143	0.0147	0.0145
RF-BR	0.0141	0.0141	0.0145	0.0142	0.0146	0.0146	0.0146	0.0145
ACC2-BR	0.0124	0.0135	0.014	0.0146	0.0147	0.0148	0.0145	0.0142
Least-BR	ACC2-BR	ACC2-BR	ACC2-BR	IG-BR	IG-BR	IG-BR	ACC2-BR	ACC2-BR
IG-LP	0.0147	0.0149	0.0144	0.0148	0.0149	0.0148	0.0147	0.0146
RF-LP	0.0145	0.0148	0.0145	0.0147	0.0148	0.0148	0.0148	0.0148
ACC2-LP	0.0144	0.0154	0.0149	0.0153	0.0154	0.0147	0.015	0.0145
Least-LP	ACC2-LP	RF-LP	IG-LP	RF-LP	RF-LP	ACC2-LP	IG-LP	ACC2-LP

2) Subset accuracy measure for bibtex dataset: The comparison of IG, RF, ACC2 is shown in Table XVII for BR and LP transformation case on different percentages of total number of features. ACC2-BR took the lead in BR case attaining maximum values in five cases among other two techniques, while in LP case, RF took the same lead among other two techniques.

3) Micro and macro averaged F1-score for bibtex dataset: ACC2-BR acquired highest values of micro-averaged score for 50% to 80% of top selected features while top 10% and 20% of selected features, IG-BR attained higher values. RF-BR only attained max value on 30% of features. In LP case ACC2 and IG-LP attained maximum values in two cases of micro-averaged measure, while RF attained higher values than other measures in 4 cases (Table XVIII). In this case, for both LP and BR transformations RF and ACC2 attained maximum values in 5 out of 16 cases. IG attained maximum values in 6 out of 16 cases, in both transformation cases.

4) Ranking Loss for bibtex dataset: Table XIX shows the ranking loss of six different metrics; and the metrices attained the least ranking score among the six metrices. In bibtex case,

IG for BR and LP transformation cases attained least values for ranking loss measures in 7 cases, while RF remained highest in 5 out of 16 cases. ACC2 attained least ranking loss in 4 out of 16 cases.

TABLE XVII: Subset Accuracy Values for Feature Ranking Metrics using KNN Classifier on Bibtex Dataset

Subset Accuracy								
Features	10%	20%	30%	40%	50%	60%	70%	80%
IG-BR	0.1407	0.1542	0.1429	0.1402	0.1362	0.1333	0.1348	0.1321
RF-BR	0.101	0.1317	0.1298	0.1348	0.133	0.1327	0.133	0.138
ACC2-BR	0.0947	0.1082	0.1303	0.1407	0.1384	0.1347	0.1394	0.1387
Max-BR	IG-BR	IG-BR	IG-BR	ACC2-BR	ACC2-BR	ACC2-BR	ACC2-BR	ACC2-BR
IG-LP	0.0974	0.1136	0.1425	0.128	0.1389	0.1447	0.1317	0.1276
RF-LP	0.1001	0.1303	0.1434	0.1416	0.1407	0.1375	0.1353	0.1303
ACC2-LP	0.0947	0.1315	0.124	0.1204	0.1136	0.1244	0.1367	0.128
MAX-LP	RF-LP	ACC2-LP	RF-LP	RF-LP	RF-LP	IG-LP	ACC2-LP	RF-LP

TABLE XVIII: Micro and Macro-averaged F1-Score Values for Feature Ranking Metrics using KNN Classifier on Bibtex Dataset

Micro-averaged F-Measure								
Features	10%	20%	30%	40%	50%	60%	70%	80%
IG-BR	0.1879	0.2541	0.258	0.2765	0.262	0.2802	0.2707	0.2766
RF-BR	0.1544	0.2186	0.2617	0.2634	0.2634	0.2714	0.2682	0.2639
ACC2-BR	0.1311	0.209	0.2474	0.259	0.2692	0.2804	0.2728	0.2785
Max-BR	IG-BR	IG-BR	RF-BR	IG-BR	ACC2-BR	ACC2-BR	ACC2-BR	ACC2-BR
IG-LP	0.1213	0.1933	0.2336	0.2433	0.2599	0.2737	0.2752	0.277
RF-LP	0.1157	0.2063	0.246	0.2643	0.2742	0.2755	0.2725	0.2751
ACC2-LP	0.0913	0.1448	0.2002	0.2192	0.2368	0.2474	0.2768	0.2775
MAX-LP	IG-LP	IG-LP	RF-LP	RF-LP	RF-LP	RF-LP	ACC2-LP	ACC2-LP
Macro-averaged F-Measure								
Features	10%	20%	30%	40%	50%	60%	70%	80%
IG-BR	0.163	0.222	0.225	0.244	0.241	0.248	0.239	0.237
RF-BR	0.131	0.195	0.205	0.232	0.234	0.243	0.238	0.235
ACC2-BR	0.113	0.183	0.226	0.231	0.243	0.251	0.243	0.24
Max-BR	IG-BR	IG-BR	ACC2-BR	IG-BR	ACC2-BR	ACC2-BR	ACC2-BR	ACC2-BR
IG-LP	0.105	0.17	0.204	0.218	0.237	0.245	0.249	0.247
RF-LP	0.099	0.182	0.218	0.238	0.249	0.248	0.242	0.245
ACC2-LP	0.077	0.126	0.174	0.19	0.208	0.219	0.226	0.239
MAX-LP	IG-LP	RF-LP	RF-LP	RF-LP	RF-LP	RF-LP	IG-LP	IG-LP

TABLE XIX: Ranking Loss Values for Feature Ranking Metrics using KNN Classifier on Bibtex Dataset

Ranking Loss								
Features	10%	20%	30%	40%	50%	60%	70%	80%
IG-BR	0.213	0.1929	0.1825	0.1841	0.174	0.1761	0.1707	0.1747
RF-BR	0.248	0.2141	0.204	0.1863	0.1815	0.1807	0.1784	0.1761
ACC2-BR	0.2606	0.2201	0.202	0.1837	0.1813	0.1732	0.1701	0.1724
Least-BR	IG-BR	IG-BR	IG-BR	ACC2-BR	IG-BR	ACC2-BR	ACC2-BR	ACC2-BR
IG-LP	0.2744	0.2229	0.1979	0.1922	0.1749	0.1754	0.1662	0.1641
RF-LP	0.2727	0.2247	0.1997	0.1799	0.1696	0.1706	0.1627	0.1692
ACC2-LP	0.2791	0.2466	0.2149	0.2032	0.1816	0.1826	0.1712	0.1714
Least-LP	RF-LP	IG-LP	IG-LP	RF-LP	RF-LP	RF-LP	RF-LP	IG-LP

IX. Discussion

We present a feature selection technique in multi-label text classification. We demonstrated the comparative study of six feature selection metrices, three for BR and three for LP case, for multi-label text classification. ACC2 measure is very simple technique and requires less computations as compared to other metrices. Despite its simplicity, It's performance is comparable to other complicated metrices as shown in Tables XX to XXII. It can be seen from Table XX, least hamming and ranking loss for Enron dataset is attained by ACC2 measure. These are %age of number of selected features assigned to each case out of eight cases for each of BR and LP case. Hence in all three datasets, ACC2-BR has 70.8% least hamming loss among the six metrices; ACC2-LP attains least hamming loss in 33.33% cases. Overall, least ranking loss for three datasets for ACC2-BR is 58.33% and 25% for ACC2-LP case. While the subset accuracy, micro, macro-averaged measures are computed for maximum values among the six feature ranking metrices.

TABLE XX: Percentage of a Feature Ranking Metric Producing Highest Subset Accuracy, Micro, Macro Average F1 Measure and Producing Lowest Hamming and Ranking Loss Enron Dataset

EValuation MEasures	FR Metrics using BR Transformation			FR Metrics using LP transformation		
	IG	Rf	ACC2	IG	RF	ACC2
Hamming Loss	12.5	25	62.5	37.5	12.5	50
Subset Accuracy	0	37.5	62.5	50	0	50
Micro-averaged F-Measure	0	62.5	37.5	12.5	25	62.5
Macro-averaged F-Measure	0	75	25	0	0	100
Ranking Loss	12.5	37.5	50	25	37.5	37.5

TABLE XXI: Percentage of a Feature Ranking Metric Producing Highest Subset Accuracy, Micro, Macro Average F1 Measure and Producing Lowest Hamming and Ranking Loss Bibtex Dataset

Evaluation Measure	FR Metric using BR			FR Metric using LP transformation		
	IG-BR	Rf-BR	ACC2-BR	IG-LP	RF-LP	ACC2-LP
Hamming Loss	37.5	0	62.5	37.5	37.5	25
Subset Accuracy	37.5	0	62.5	12.5	62.5	25
Micro F	37.5	12.5	50	25	50	25
Macro f	37.5	0	62.5	37.5	62.5	0
Ranking Loss	50	0	50	37.5	62.5	0

TABLE XXII: Percentage of a Feature Ranking Metric Producing Highest Subset Accuracy, Micro, Macro Average F1 Measure and Producing Lowest Hamming and Ranking Loss Medical Dataset

Evaluation Measure	FR Metric using BR Transformation			FR Metric using LP Transformation		
	IG-BR	Rf-BR	ACC2-BR	IG-LP	RF-LP	ACC2-LP
Hamming Loss	12.5	0	87.5	25	50	25
Subset Accuracy	0	62.5	37.5	0	75	25
Micro-averaged F-Measure	37.5	0	62.5	2	37.5	37.5
Macro f	50	0	50	12.5	37.5	50
Ranking Loss	0	25	75	25	37.5	37.5

X. Conclusion

In this paper, we evaluate the performance of three feature ranking algorithms and two data transformation techniques by using five evaluation measures on three benchmark datasets. For data transformation techniques from multi-label to single label, we conclude that binary relevance doesn't take into consideration the label dependency. While on other hand LP only consider the distinct labelsets, hence unable to predict new labelsets causing over-fitting of training data.

In feature ranking algorithms Relief F measure does not deal with redundant features. Rather then converting a multinomial classification problem into binomial classification problem, RELIEFF searches for k near misses from each different class and averages their contributions for updating W, weighted with the prior probability of each class. Information gain capture the amount of information present in a feature for the purpose of automatic text classification. ACC2 select highly discriminative features which occur more time in one class but less times in other class. In future work, we will adopt ACC2 measure to directly deal with multi-label data.

Acknowledgment

We are thankful to Muhammad Salman Khalid and Shoaib Munir for their assistance with experimental setup, and valuable comments that greatly improved the manuscript.

References

[1] Chen, Weizhu, et al. "Document transformation for multi-label feature selection in text categorization." Data Mining, 2007. ICDM 2007. Seventh IEEE International Conference on. IEEE, 2007.

[2] Gibaja, Eva, and Sebastin Ventura. "A tutorial on multilabel learning." ACM Computing Surveys (CSUR) 47.3 (2015): 52.

[3] Park, Hee Jung, and John Guy. "Sixth nerve palsy post intravitreal bevacizumab for AMD: a new possibly causal relationship and complication?." Binocular Vision and Strabismus Quarterly 22.4 (2007).

[4] Kumar, Neeraj, et al. "Attribute and simile classifiers for face verification." Computer Vision, 2009 IEEE 12th International Conference on. IEEE, 2009.

[5] Tsoumakas, Grigorios, Ioannis Katakis, and Ioannis Vlahavas. "Mining multi-label data." Data mining and knowledge discovery handbook. Springer US, 2009. 667-685.

[6] Pereira, Rafael B., et al. "Categorizing feature selection methods for multi-label classification." Artificial Intelligence Review (2016): 1-22.

[7] Law, Martin HC, Mario AT Figueiredo, and Anil K. Jain. "Simultaneous feature selection and clustering using mixture models." IEEE transactions on pattern analysis and machine intelligence 26.9 (2004): 1154-1166.

[8] Kononenko, Igor, Edvard imec, and Marko Robnik-ikonja. "Overcoming the myopia of inductive learning algorithms with RELIEFF." Applied Intelligence 7.1 (1997): 39-55.

[9] Kira, Kenji, and Larry A. Rendell. "A practical approach to feature selection." Proceedings of the ninth international workshop on Machine learning. 1992.

[10] Kohavi, Ron, and George H. John. "Wrappers for feature subset selection." Artificial intelligence 97.1-2 (1997): 273-324.

[11] Guyon, Isabelle, and Andr Elisseeff. "An introduction to variable and feature selection." Journal of machine learning research 3.Mar (2003): 1157-1182.

[12] Liu, Huan, and Lei Yu. "Toward integrating feature selection algorithms for classification and clustering." IEEE Transactions on knowledge and data engineering 17.4 (2005): 491-502.

[13] Spolar, Newton, et al. "Filter approach feature selection methods to support multi-label learning based on relieff and information gain." Advances in Artificial Intelligence-SBIA 2012. Springer, Berlin, Heidelberg, 2012. 72-81.

[14] Huang, Jun, et al. "Joint Feature Selection and Classification for Multilabel Learning." IEEE Transactions on Cybernetics (2017).

[15] Trohidis, Konstantinos, et al. "Multi-Label Classification of Music into Emotions." ISMIR. Vol. 8. 2008.

[16] Guo, Yumeng, Fulai Chung, and Guozheng Li. "An ensemble embedded feature selection method for multi-label clinical text classification." Bioinformatics and Biomedicine (BIBM), 2016 IEEE International Conference on. IEEE, 2016.

[17] Jungjit, Suwimol, and Alex Freitas. "A lexicographic multi-objective genetic algorithm for multi-label correlation based feature selection." Proceedings of the Companion Publication of the 2015 Annual Conference on Genetic and Evolutionary Computation. ACM, 2015.

[18] Jungjit, Suwimol, et al. "Two extensions to multi-label correlation-based feature selection: A case study in bioinformatics." Systems, Man, and Cybernetics (SMC), 2013 IEEE International Conference on. IEEE, 2013.

[19] Khalid, Samina, Tehmina Khalil, and Shamila Nasreen. "A survey of feature selection and feature extraction techniques in machine learning." Science and Information Conference (SAI), 2014. IEEE, 2014.

[20] Lee, Jaesung, and Dae-Won Kim. "Feature selection for multi-label classification using multivariate mutual information." Pattern Recognition Letters 34.3 (2013): 349-357.

[21] Sechidis, Konstantinos, Nikolaos Nikolaou, and Gavin Brown. "Information theoretic feature selection in multi-label data through composite likelihood." Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR). Springer, Berlin, Heidelberg, 2014.

[22] Rehman, Abdur, Kashif Javed, and Haroon A. Babri. "Feature selection based on a normalized difference measure for text classification." Information Processing & Management 53.2 (2017): 473-489.

[23] Forman, George. "An extensive empirical study of feature selection metrics for text classification." Journal of machine learning research 3.Mar (2003): 1289-1305.

[24] Liu, Huan, and Hiroshi Motoda, eds. Computational methods of feature selection. CRC Press, 2007.

[25] Zhang, Min-Ling, and Zhi-Hua Zhou. "ML-KNN: A lazy learning approach to multi-label learning." Pattern recognition 40.7 (2007): 2038-2048.

[26] Schapire, Robert E., and Yoram Singer. "BoosTexter: A boosting-based system for text categorization." Machine learning 39.2-3 (2000): 135-168.

[27] Zhang, Min-Ling, and Zhi-Hua Zhou. "Multilabel neural networks with applications to functional genomics and text categorization." IEEE transactions on Knowledge and Data Engineering 18.10 (2006): 1338-1351.

[28] Maimon, Oded, and Lior Rokach. "Introduction to knowledge discovery and data mining." Data Mining and Knowledge Discovery Handbook. Springer US, 2009. 1-15.

AUTHORS' PROFILES

Muhammad Nabeel Asim received his Bachelor degree from University of Management and Technology (UMT) and Masters degree in Electrical Engineering from University of Engineering and Technology, Lahore, Pakistan. Currently working as Research Officer at Al-Khawarizmi Institute of Computer Science (KICS) University of Engineering and Technology (UET), Lahore, Pakistan. His research interests are Bioinformatics, Artificial Intelligence, Information Retrieval,Natural Language Processing and Text classification.

Dr. Abdur Rehman completed his PhD from UET Lahore in computer science. During his PhD, he has been part of Al-Khawarizmi Institute of Computer Science, UET Lahore for 8 year working as a researcher on different posts. His core expertise are in the field of machine learning and his focused area of research is "Text Classification". He is currently working as Assistant Professor in Department of Computer Science, University of Gujrat, Pakistan.

Dr. Umar Shoaib did his PhD in Department of Computer and Control Engineering Politecnico di Torino, Italy. His current research interests include Machine Learning, Robotics, Text Mining, Scalable Networks, Interfaces, Cloud Computing, Natural Language Processing, Text mining and Internet of Things.