Journal of Theoretical and Applied Information Technology
30th April 2018. Vol.96. No 8

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

EVALUATING THE INFLUENCE OF FEATURE SELECTION TECHNIQUES ON MULTI-LABEL TEXT CLASSIFICATION METHODS USING MEKA

[1]SUSAN KOSHY, [2]DR. R.PADMAJAVALLI
[1]Research Scholar, Bharathiar University, Coimbatore, Tamil Nadu, India,
Assistant Professor, Department of Computer Science,
St.Thomas College of Arts and Science, Chennai, India.
[2]Associate Professor, Department of Computer Applications,
Bhaktavatsalam Memorial College for Women, Chennai, India
Email: [1]susanabraham90@gmail.com,[2]padmahari2002@yahoo.com

ABSTRACT

Multi-label classification has generated a lot of interest with its useful applications in real world situations as against traditional single label classification. Feature selection has a positive affect on the performance of multi-label classification as it elevates the performance of learning algorithms, reducing the storage requirement and complexity of the multidimensional space. There are many multi label algorithms to handle the problem of multi-label classification and the issue of dimensionality is overcome by feature selection. There are several feature selection techniques and the right combination of multi-label classification and multi-label feature selection will help in building an efficient model of classification for a given dataset. This paper uses the available algorithms and evaluates the influence of filter feature selection methods and multi-label classification for two standard text datasets drawn from real domains. Multi-label Problem transformation transform the multi-label dataset into single label. Binary Relevance, Classifier Chains, Pruned Sets and an ensemble method called RAkEL, multi-label classifiers with two single label classifier namely J48 and Naïve base are used. Feature selection is followed by multi-label classification. This paper uses five standard techniques namely correlation feature subset selection, correlation feature selection, gain ratio, information gain and ReliefF to evaluate the relationship between feature selection, multi-label classification and single label base classifier in order to obtain enhanced multi-label evaluation metrics.

Keywords: *Correlation Based Feature Subset Selection, Correlation Feature Selection, Multi-Label Text Classification, Gain Ratio, Information Gain, Relieff.*

1. INTRODUCTION

Feature selection is an essential preprocessing method to solve the problem of dimensionality in multi-label learning. The presence of redundant and irrelevant features hampers the learning process and involving feature selection greatly improves learning. Feature selection has fourfold benefit a) requiring less amount of data needed to achieve learning , b)improved accuracy of prediction c)reduction in the time required for execution d) knowledge learned is more easily understood. The algorithms of Feature selection can be either filter, wrappers or embedded. Wrappers use the technique where the classification algorithm is used to select the relevant features while filters work independently and select features not based on the classification algorithm. The former gives good prediction results but are time consuming, while the later is quicker and can be used before the prediction model. A good subset of features should be highly correlated to the predicted or target class and yet the subset of features should be uncorrelated of each other.

In single-label classification each example is associated with a single class label and a classifier learns to associate each new test example with one of these known class labels and when it is associated with multiple labels it is multi-label classification just as the human brain can associate one idea with multiple concepts. A news article about a conference on climate change can be labeled both politics and environment [1] [2].

The multi-label context contains an extra dimension and this additional dimension affects both the learning and evaluation processes. The

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

evaluation process is no longer straight forward as in single label learning, since a simple correct/incorrect evaluation is insufficient to convey the comparative predictive power of a given classifier. Thus, different evaluation methods are needed. Learning is affected by label correlations, or label relationships, that occur in the multi-label dimension. Instead of choosing a single class label from a label set, a multi-label classifier must consider combinations of labels. This situation is aggravated as the quantity of data grows. Two unique approaches are put forth to handle multi-label classification problems one is to adapt the algorithm to handle the classification such as support vector machines, AdaBoost, KNN. The second approach which this paper deals with is converting the multi-label problem to several single label problems such as Binary relevance, classifier chains, pruned sets and Random K label sets [1].

Some of the feature evaluation metrics to evaluate the goodness of features for classification are Fisher score, Chi square, ReliefF, Gain Ratio, Information Gain(IG), Correlation Feature Subset selection(CFS) and Rough Set. Chi square is a common statistical measure and is designed for discrete variables and requires an extra step of discretising the features. It behaves erratically for small counts of rarely occurring features in text categorization. Mutual information is a symmetric measure about the information one variable has about another.

The paper discusses some of the related work in multi-label feature selection. The next section gives the key issues followed by the general frame work of multi-label feature selection. The next section discusses feature selection techniques. In the next section, problem transformation multi-label classification methods and the evaluation metrics are discussed. This is followed by the experimental results, discussion and conclusion.

2. MOTIVATION

The problem of dimensionality in solving multi-label classification and to evaluate how different multi-label classification algorithms perform for different types of feature selection methods has motivated this study on different text datasets. The filter approach of feature selection is used because it is fast, simple and costs less in terms of computation. Further the filter methods do not depend upon the classification algorithm. Five filter feature selection approaches correlation based subset feature selection, correlation, Information gain, gain ratio and ReliefF which have been used in literature have been selected for evaluation. CFS and ReliefF are multivariate but IG and Gain ratio are univariate. Problem transformation methods, transform the multi-label dataset into single label and these are selected for classification because it is less complex such as Binary relevance, Classifier Chain, Pruned sets and Rakel. The forte of this evaluation is that we have been able to look at several results for two text datasets before narrowing on the feature selection method and the multi-label classification method.

3. RELATED WORKS

Min-Ling Zhang, Jose M. Pena et al. [28] used a method called MLNB which are naive Bayes' classifiers incorporating two stage filter wrapper feature selection to handle multi-label instances.

Gauthier Doquire, Michel Verleysen et al.[27] proposed a method that uses the multivariate mutual information criterion along with a problem transformation and a pruning strategy. The earlier works use the univariate Chi Square statistics to select features which does not consider redundancy between feature and the other disadvantage is that they are designed for discrete variables but when continuous variables are used they have to be discretized.

Rafael B. Pereira, Alexandre Plastino et al.[29] used Information Gain feature selection along with problem transformation techniques for multi-label classification.

Yaping Cai, Ming Yang et al.[30] have used a strategy ML-ReliefF, to select distinguishing features to improve multi-label classification accuracy along with the ML-KNN classifier.

Suwimol Jungjit and Alex A. Freitas[31] have proposed a new genetic algorithm in order to search for highly relevant subsets.

4. KEY ISSUES

The multi label learning task is managed label-by-label and the coexistence has been ignored and is the first order strategy [8]. Multi-label learning is handled by taking into consideration relations among pairs of labels in the second order strategy and gives a rank between relevant and irrelevant labels. This type of a strategy gives a better performance [8]. In the high order strategy relations that exist between labels are considered along with the influence of one label on another. The high order strategy has a better correlation

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

modeling than the second and first order strategies but the computation is more strenuous and not easily scalable [8][16].

5. MULTI-LABEL FEATURE SELECTION

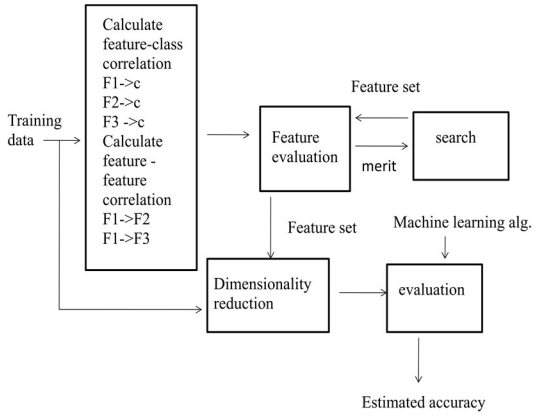

Figure1: A General Framework of Feature Selection for Classification

The learning of a good classifier is hindered by the presence of unwanted features due to the huge size of the data. The number of irrelevant or redundant features when removed can drastically reduce the running time of the learning algorithms and yield a better classifier. The feature selection algorithms address few basic issues namely a) Starting point of the search of features will affect the search strategy which means a forward search, a backward search or a mid way search can be used b) Search organization is based on the starting point and some of the search techniques are greedy hill climbing, best first search and genetic search to name a few c) How the features are evaluated is the next important step either as filter which are independent of the learning algorithm or may iteratively utilize the performance of the learning algorithms to evaluate the quality of the selected features as in wrapper models d) the last step is a stopping criterion where the feature selector has to stop searching in the space of feature subset when none of the alternates improves the merit of a subset of features[20] (Figure 1).

With the final selection of features, a classifier is induced for the prediction phase. Only the minimally sized subset of features are selected according to the following criteria, a) classification accuracy is not reduced b) the final class distribution with the selected features, is as close as possible to the original class distribution with all features.

The standard method of feature selection is to search through the subsets of features and try to find the best ones among the competing 2^m candidate subsets according to some evaluation function. It is expensive and computationally difficult, even for a medium-sized feature set of size m. Methods based on heuristic or random search methods try to reduce computational complexity by compromising on performance. The stopping criterion will prevent an exhaustive search of subsets when further searching does not improve its quality[21][24].

5.1 Multi-Label Feature Selection Methods

Correlation based feature selector(CFS) algorithm is an algorithm that ranks the feature subsets according to a correlation based evaluation function that is heuristic and is based on the Pearson coefficient. Subsets that are highly correlated with the class but not correlated with each other are selected. The redundant features are the ones that are highly correlated among themselves and have to be discarded. When the features or predictors are selected we need to choose features that measure various aspects of the target variable. The Pearson's coefficient shows that the correlation between a set of predictors and a target variable is a function of the number of predictor variables and the magnitude of inter-correlations among them together with the magnitude of the correlations between the predictors and the outside variable.

$$. x_{zc} = \frac{k\, x_{zi}}{\sqrt{k + k(k-1) x_{ii}}} \quad (1)$$

x_{zc} is the correlation between the summed predictors and the target variable(merit or goodness of a feature), x_{zi} is the average correlations between the predictors and the target variable(feature and class correlation) and x_{ii} is the average inter correlations between the predictor(correlations among the features)[26][20].

The search for the best features or predictors can be forward selection, backward elimination or best fit search. The forward selection is a strategy where the search starts with no features and greedily selects the next feature until no further addition will improve the evaluation method. The backward elimination starts with the entire feature set and removes one feature at a time till the evaluation does not deteriorate. The best fit search can select the entire feature set and remove redundant features one by one or start with none at all and add features one at a time. In this paper the

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

best fit search method has been used.

Information gain: The concept of entropy is used in information gain as a measure to decide on the best splitting criteria (partitioning the dataset on a particular feature) which can be used in decision tree classification algorithms. Entropy is a measure of the amount of indecision in a dataset or how much information is required to explain an item or also how many bits are required to portray all the classes a feature belongs to. The information gain of a particular attribute from a set of training instances can be defined as the difference in the entropy of the entire training set and the sum of the entropy of the subsets partitioned on the values of that particular feature.

$$\text{InfoGain}(Z, F) = \text{Entropy}(Z) - \sum_{x \in F} \frac{|Z_x|}{|Z|} * \text{Entropy}(Z_x) \quad (2)$$

Z is the set of training examples, F is the attribute considered, and x is the value of the attribute F.

The entropy of an entire dataset for multi-label instances is calculated as the number bits to describe whether an instance belongs to a class or does not to a belong to a class using probability or relative frequency.

$$\text{Entropy}(Z) = -\sum_{i=1}^{N}((p(c_i)\log p(c_i)) + ((q(c_i)\log q(c_i))) \quad (3)$$

N is the number of classes

$p(c_i)$ is the probability of class c_i

$q(c_i)=1- p(c_i)$ is the probability of not being a member of class c_i

A high value of information gain indicates a strong correlation between the feature and the particular class [19][22].

ReliefF according to Yaping Cai et al considers the effect of interacting features or correlations. ReliefF employs a statistical method to select the appropriate features which are relevant. ReliefF randomly selects a sample of instances, and for each instance in it finds Near Hit and near Miss instances on the basis of Euclidean distance measure which is the distance between two points in Euclidean space. An instance is near hit if it has least Euclidean distance among all instances of the same class as that of the chosen instance and near Miss is the instance which has smallest Euclidean distance among all instances of different class. The weights of the features are updated which are zero at the start and is on the assumption that a feature is more relevant if it differentiates between an instance and its near Miss, and less relevant if it differentiates between an instance and its near Hit. When all instances in the sample are evaluated, it chooses all features having a weight higher than or equal to a threshold [25]. An instance is chosen from the dataset, and the nearest neighboring sample that belongs to the same class (nearest hit) and the nearest neighboring sample that belongs to a differing class (nearest miss) are identified. A change in attribute value accompanied by a change in class increases the weight of the attribute based on the belief that the attribute change could cause a class change. Alternately, a change in attribute value followed by no change in class leads to decrease in weight of the attribute based on the fact that the attribute change had no effect on the class. This procedure of refreshing the value of the weight of the attribute is performed for a random set of samples in the dataset or for every sample in the dataset. The new weights are then averaged so that the final weight is in the range [−1, 1]. The advantage of ReliefF is that it works for noisy and correlated features and for nominal and continuous data according to literature[30].

Gain ratio: Information gain is biased towards attributes or features which have a large number of unique values which will produce a large number of partitions which is overcome by gain ratio. Gain ratio applies normalization on information gain using a split information value.

A decision tree has non-terminal nodes which are tests on one or more attributes or features and terminal nodes represent the outcome of decisions. The information gain measure is used to select the test attribute at each node of the decision tree. The information gain is biased towards attributes having a large number of values. The ID3 decision tree induction algorithm is enhanced by C4.5 and uses an extension of information gain known as gain ratio.

The expected information needed to classify a tuple in D is given by

$$\text{Info}(D) = -\sum_{i-1}^{m} \log_2(p_i) \quad (4)$$

p_i is the probability that a tuple in D belongs to class C_i. To arrive at a perfect classification more details will be needed based on the partitioning on attribute A

$$\text{Info}_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} * \text{info}(D_j) \quad (5)$$

Information gain is the difference between original and new information

$$\text{Gain}(A) = \text{Info}(D) - \text{Info}_A(D) \quad (6)$$

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Gain(A) tells how much would be gained by branching on the feature A.

The split information value is the normalized information gain which represents the information generated by splitting the training data set into partitions based on the outcomes of the test on a particular attribute and it considers the total number of tuples.

$$\text{Splitinfo}_A(D) = -\sum_{j=1}^{v} \frac{|Dj|}{|D|} * \log_2 \frac{|Dj|}{|D|} \quad (7)$$

A is the attribute with distinct values $\{a_1,a_2,a_3 \ldots a_v\}$,that is used to split the partition of tuple D into v partitions or subsets $\{D_1,D_2,D_3 \ldots D_v\}$ where D_j contains tuples that have outcome a_j of A.

The gain ratio is defined as

$$\text{GainRatio}(A) = \frac{\text{Gain}(A)}{\text{Splitinfo}(A)} \quad (8)$$

The attribute with the maximum gain ratio is selected as the splitting attribute [3].

6. MULTI-LABEL CLASSIFICATION METHODS

The methods of Multi-label classification are grouped as follows:

(a) Methods of problem transformation

(b) Methods of algorithm adaptation.

The methods of problem transformation can transform classification problems that are multi-label into a single label classification or into problems of regression and further fit those into the algorithms that currently exist. The method of algorithm adaptation can extend algorithms of specific learning in order to handle data that is multi-label directly and ensures that it fits the algorithm to the data [5],[7].

Problem transformation methods

BR or Binary Relevance is a common method of problem transformation owing to its simplicity [16]. This considers the prediction of every label as a task of binary classification that is independent. It builds its own binary classifiers for every label set. The BR predicts the union of these labels which are predicted positively by each of the classifiers. The main limitation of this method is that an assumption that the assigned labels for each example are independent is made and the correlation aspect among labels is completely ignored [4][15].

LP or Label Power-Set This is a problem transformation method [1]. It considers every unique label set in the set of training as one of them of a newly brought about classification task with a single label. This classifier of LP predicts the label that is most likely that is a set of labels. The correlations of labels are taken into consideration in this but it is much more complex [16].

RAkEL or Random k-label-sets these make a construction of an ensemble of classifiers of label power sets [12]. Each of the classifiers of LP is trained in different subsets randomly made in a set of labels. A decision is calculated averagely for every label and finally is taken as a positive for a particular label and if the decision is larger than that of a particular threshold value that is given as the result. This method also considers the label correlation problems.

CC or Classifier Chains An improved version of binary relevance where a chain of binary classifiers are created and every classifier will be responsible for learning as well as predicting and takes into account the predicted labels of the previous classifier thus forming a chain [10]

PS or Pruned Sets this treats label sets as single ones and allows the process of classification to take into account the correlations that exist between labels. PS generally focuses on the important correlations that brings down the complexity and at the same time increases accuracy [11].

Algorithm Adaptations Methods

This type of method adapts its internal mechanism to permit multi label problems like lazy learning and its associative methods, support vector machines, neural networks, probabilistic methods and decision trees [1][9][13].

6.1 Evaluation Metrics For Multi-Label Classification

In multi-label datasets, the number of labels associated with an instance may be small when compared to the total number of possible labels |L|. This factor can influence the performance of different multi-label methods namely cardinality and density and cause different behaviors in multi-label learning methods. Cardinality of a dataset is the mean of the number of labels which belong to the instances of the dataset and density is the mean of the number of labels which belong to the instances of the dataset divided by the number of labels.

The multi label classifier evaluation needs other measures compared to the problems of single label [6]. While classifying these examples the classification result can be either partially right or wrong. This takes place when there is a correct

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

assigning of an example to the minimum number of labels it belongs to, but it does not assign all the labels that it actually belongs. So a classifier can also assign one or even more labels to which it does not belong. The evaluation measures can be grouped broadly into two, based on example and based on label. The former makes an evaluation of the average difference between the labels that are predicted and their actual labels for every instance or example. The latter on the other hand, is a metric that ensures each label is being evaluated initially and then an averaging is done for all of the labels that are given for consideration [16].If a dataset for evaluation in multi-labeled examples is shown as (x_i, Y_i), i=1… N, in which $Y_i \subseteq L$ denotes the actual set of true labels and $L=\{\lambda_j: j=1 \dots M\}$ denotes the actual set of all labels. If an example x_i is given then the label set which is predicted by a means of a multi-label method is shown as Z_i, when the rank that is predicted for a label λ is shown as $r_i(\lambda)$. The label that is most relevant gets the highest rank (1), and the one that is least relevant gets the lowest rank (M) [16].

Example-based Measures

Hamming Loss: The Hamming Loss makes an evaluation of the frequency in a given example and is associated to labels that may be wrong or one that belongs to an instance which is not predicted correctly. An ideal performance is got when the loss of hamming is equal to 0. The loss of hamming being lower the result will be a classifier that performs better.

$$\text{Hamming Loss}=\frac{1}{N}\sum_{i=1}^{N}\frac{|Y_i \Delta Z_i|}{M} \quad (9)$$

Accuracy: Accuracy is the one that measures whether a true label Y_i is close to a label that is predicted Z_i. It denotes the ratio of union as well as the intersection of the label sets both predicted and actual which are taken for every example and further averaged considering a number of different examples

$$\text{Accuracy}=\frac{1}{N}\sum_{i=1}^{N}\frac{|Yi \cap Zi|}{|Yi \cup Zi|} \quad (10)$$

Precision: Precision may be defined as that percentage of positive examples that are true belonging to all examples that are classified under the category of positive by a classification model.

$$\text{Precision}=\frac{1}{N}\sum_{i=1}^{N}\frac{|Yi \cap Zi|}{|Zi|} \quad (11)$$

Recall: Recall denotes that percentage of examples that are categorized by a positive model of classification that is true as well as positive.

$$\text{Recall}=\frac{1}{N}\sum_{i=1}^{N}\frac{|Yi \cap Zi|}{|Yi|} \quad (12)$$

F-Measure: The F-Measure or the F-Score is a proper combination of both Precision as well as Recall. It is nothing but the harmonic average of the precision and the metrics of recall that is aggregated to the score performance

$$\text{F-Measure}=\frac{1}{N}\sum_{i=1}^{N}\frac{2|Yi \cap Zi|}{|Zi|-|Yi|} \quad (13)$$

Subset Accuracy: The Subset Accuracy is one very restrictive metric of accuracy that considers one classification as right if all the predicted labels by classifier is right.

$$\text{Subset Accuracy}=\frac{1}{N}\sum_{i=1}^{N} I\left(|Zi|=|Yi|\right) \quad (14)$$

Exact Match is defined as the accuracy of each example where all label relevancies must match exactly for an example to be correct.

$$\text{Exact match}=\frac{1}{N}\sum_{i=1}^{N} I\left(Y_i=Z_i\right) \quad (15)$$

Label-based Measures

The precision known as Micro-averaged precision denotes the example ratio that is rightly classified as either true positives or as false positives incorrectly. The Micro-averaged recall denotes the ratio of the examples that are classified rightly as 1 and all the other examples that actually belong to class 1 which is the false negative. The micro-averaged F-measure denotes a mean that is harmonic belonging to both Micro-Recall and Micro-Precision.

The precision that is Macro-average is first computed by duly computing the precision for every label separately and further averaging this over other labels. This procedure is used for macro-averaged recall as well. The F-measure that is macro-averaged is the harmonic mean of the Macro-recall and the Macro-precision [14].

One-error: This measure makes an evaluation of the frequency of all labels that are top-ranked and not in a true label set. Its best performance is got only when one error equals 0. The lower the one error values, the better the performance.

Coverage: Coverage may be defined as the distance that covers all the labels possible that are duly assigned to a sample *x*. If the value of the

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

coverage is smaller the performance is better.

Average Precision: This denotes the average precision that is taken for all labels possible and can evaluate the algorithms completely. It measures the labels that are ranked above another label $l \in Yi$ that actually is in *Yi*. The ideal performance is got only when the average precision equals 1 [14].

7. DESCRIPTION OF DATASETS

Name	N	M	q	LC	LD	DC
Medical	978	1449	45	1.25	0.03	94
Langlog	1460	1004	75	1.66	0.07	1147

Table 1: Dataset description

The description of the two data sets Medical and Language log are given in Table1, where N indicates the number of instances, M is the number of features, q is the number of labels, the label cardinality LC is the average number of labels associated with each instance, the label density LD is a normalized version of label cardinality divided by the total number of labels, and the number of unique combinations DC of labels. The datasets were obtained from Mulan and Meka repositories [18].

8. EXPERIMENTAL RESULTS

MEKA is an extension of WEKA framework and is used to evaluate two different multi-label datasets MEDICAL and LANGUAGE LOG. MEKA is based on WEKA Toolkit and includes many multi-label learning methods from literature. WEKA developed by Waikato University is open source software that is issued under General Public License.

Language Log is a multi-label text dataset from the language forum discussion and is available from MEKA or R Ultimate Multi-label dataset repository. It contains 1460 instances, 1004 features and 75 labels.

Medical is a text multi-label dataset, obtained from MULAN a java library for multi-label learning and is an anonymous free text about patient symptoms. It contains 978 instances, 1449 features and 45 labels which are codes from the international classification of diseases. Both these sets of data are in ARFF (Attribute rich file format) and preprocessed [18].

The evaluations are performed on 32bit machine with a clock speed of 3.40 GHz. Feature selection techniques namely correlation feature subset selection(CFS), correlation(correl), gain ratio(GR), information gain(IG) and ReliefF are used. After feature selection is performed the filtered dataset is sent to the classifier for multi-label classification problem transformation such as RAkEL, Classifier chains, Pruned sets and Binary relevance and evaluated for each single label base classifiers, J48 and NB or Naïve Bayes. The performance of each of the five feature selection on each of the four classification methods (BR ,CC, PS, Rakel) and for two different single label classifier for two datasets is analysed which means 16 classification evaluations before feature selection(BFS) and 80 classifications after feature selection have been done to study the relationship between feature selection method, problem transformation algorithm and base classifier. The metrics chosen for comparison are hamming loss (Equation 9), total time to build the model, average precision (Equation 11), accuracy (Equation 10) and exact match (Equation 15)

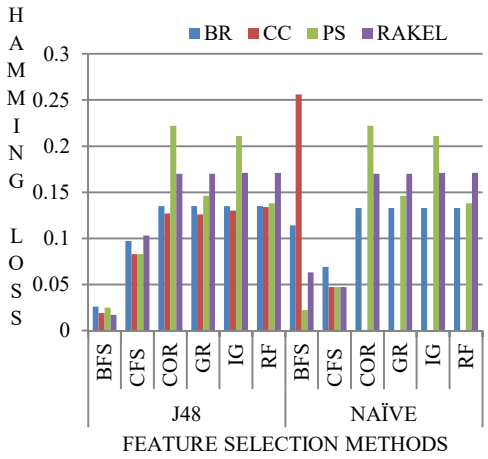

Figure 2: Hamming loss for Language Log dataset

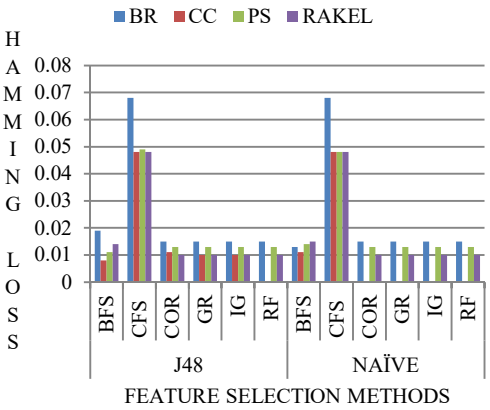

Figure 3: Hamming loss for Medical dataset

Journal of Theoretical and Applied Information Technology
30th April 2018. Vol.96. No 8

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

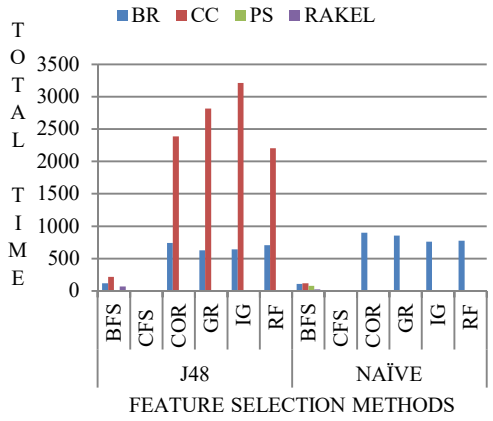

Figure 4: Total time to build the model for Language Log dataset

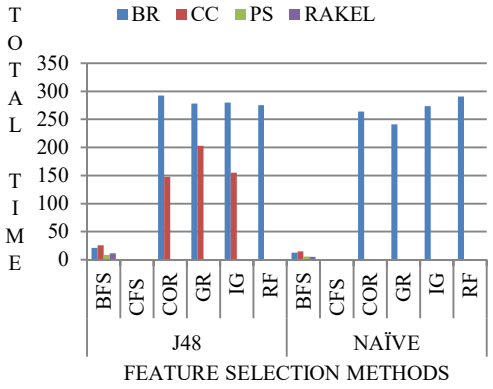

Figure 5: Total time to build the model for Medical dataset

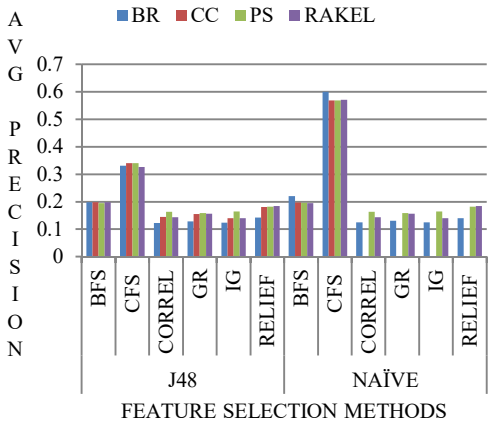

Figure 6: Average precision for Language Log

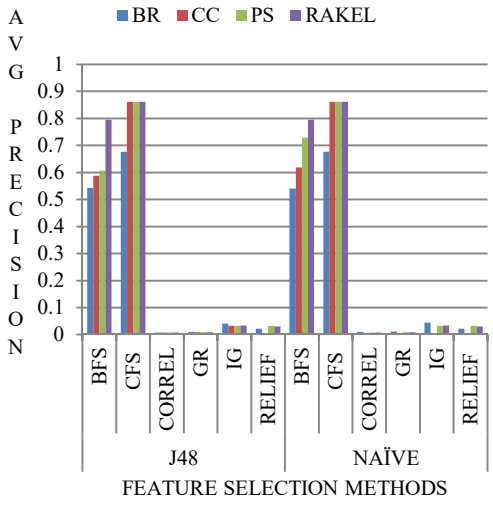

Figure 7: Average precision for Medical dataset

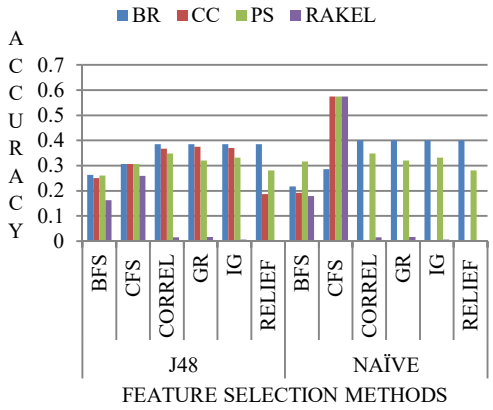

Figure 8: Accuracy for Language Log dataset

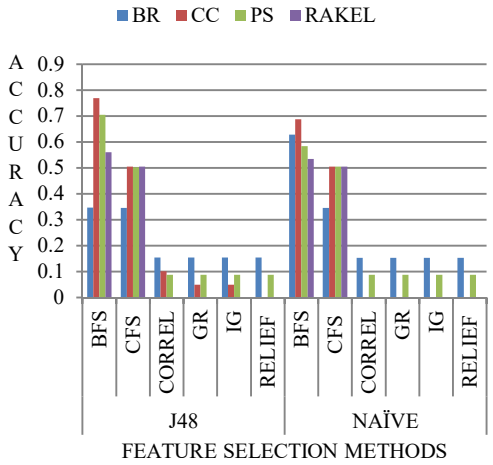

Figure 9: Accuracy for Medical dataset

Journal of Theoretical and Applied Information Technology
30th April 2018. Vol.96. No 8

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

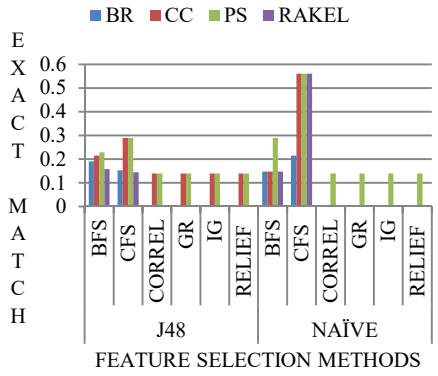

Figure 10: Exact Match for Language Log dataset

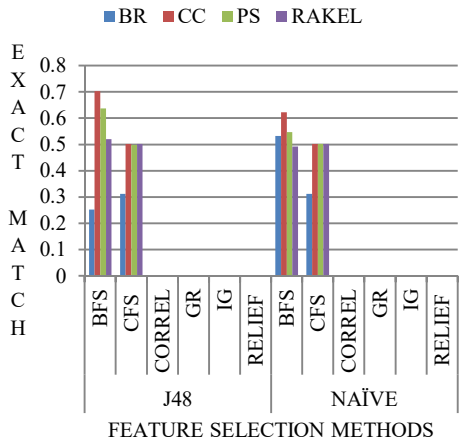

Figure 11: Exact Match for Medical dataset

9. DISCUSSION

The Filter Feature selection is done before the process of classification, unlike wrapper feature selection. The wrapper approach is more complex but gives better performance. The filter approach is independent of the learning algorithm, computationally simple, fast and scalable. Using filter method, feature selection is done and the resultant dataset is given as input to various multi-label classifiers. Nearly 60% of the literature publications have been on filter feature selection approach [23]. Due to the availability of several algorithms for feature selection and multi-label classification this paper intends to verify the influence of different problem transformation classifiers with different feature selection techniques. The main aim of the paper is to evaluate which among the feature selection method chosen improves the classification metrics for specific problem transformation methods. Five filter feature selection methods CFS, Correlation, Gain Ratio, Information gain and ReliefF and four problem transformation methods are selected for observation.

According to literature, filter feature selection methods such as information gain, gain ratio do not take into account the interactions between features, and this is overcome by multivariate filters such as CFS. The individual predictive ability of each feature along with the degree of redundancy between them is evaluated by CFS and proves the worth of the subset [26][27][30].This has been validated in our evaluations of datasets.

Post MJ et al have run 12 algorithms on 400 data sets to understand whether feature selection improves classification accuracy for a given model and have observed that 41 per cent have improved results but only in 10 per cent there was statistical significance [32]. From this it may be assumed that features have already been carefully selected by domain experts in the datasets used for experimentation. This is the case with our selection of datasets as there have not been significant improvements in evaluation metrics even after feature selection. Specifically there is not much improvement in Average precision and accuracy after applying feature selection except with correlation feature subset selection. Thus evaluation metrics may differ when raw data are used.

In our evaluation after applying a Feature selection technique the resulting dataset are classified using multi-label classification by Binary Relevance, Classifier Chains, Pruned Sets and Rakel using J48 and Naïve Base single label classifier. Thus 16 classifications are done before feature selection (BFS) and 80 classifications after feature selection have been performed on the two text datasets Medical and Language Log for multi-label classification. The medical dataset gives good hamming loss value for all feature selection techniques (Figure 2). The Language Log dataset gives higher hamming loss after feature selection which is contrary to rule (Figure 3) The total time taken to build the model after feature selection is reduced for both Pruned set and Rakel classifiers for both the datasets which is a positive result but Binary relevance and Classifier chain have very high time to build the model irrespective of the feature selection method used (Figure 4 and 5). Average precision is high for both datasets when Correlation Feature subset selection (CFS) is used for all the Multi-label classifiers (Figure 6 and 7). Similarly accuracy also is high for both datasets when Correlation Feature subset selection (CFS) is

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

used for all the Multi-label classifiers and base classifiers (Figure 8 and 9). Exact match metric is high for both datasets when Correlation Feature subset selection (CFS) is used for CC, PS and Rakel Multi-label classifiers and J48 and Naïve base classifiers(Figure 10 and 11).

10. CONCLUSION

In this paper we have compared five filter feature selection methods (CFS, Correlation, Gain Ratio, Information Gain and ReliefF) on four problem transformation methods (BR, CC, PS, Rakel) using two base classifiers J48 and Naïve Base on two text multi-label datasets (Medical and Language Log). The main aim of the paper is to evaluate which among the feature selection method chosen improves the classification metrics for specific problem transformation methods. The evaluation of the classification results, 96 in all reveal that Correlation feature subset selection is a better feature selection and the multi-label classifiers pruned sets and Rakel give efficient results along with the naïve base single label classifier which concurs with results with Read J et al [12][33].

There is not much improvement in Average precision and accuracy after applying feature selection except with correlation feature subset selection. It may be attributed to the fact that the features of the datasets are selected by domain experts and experimental setting may differ when raw data are used.

In our earlier paper only one feature selection technique was used namely correlation feature subset selection evaluation and a best fit search technique. Improved multi-label evaluation metrics were obtained for all four problem transformation methods (BR, CC, PS, Rakel) uniformly for two different multi-label datasets [17]. In future we need to experiment with raw text data to further validate the results put forth in this paper.

REFERENCES

[1] Zhang ML, Zhou ZH. A review on multi-label learning algorithms. IEEE transactions on knowledge and data engineering. 2014 Aug;26(8):1819-37.

[2] Gupta V, Lehal GS. A survey of text mining techniques and applications. Journal of emerging technologies in web intelligence. 2009 Aug 1;1(1):60-76.

[3] Han J, Kamber M. Simon Fraser University. Data Mining: Concepts and Techniques. 2001.

[4] Cherman EA, Monard MC, Metz J. Multi-label problem transformation methods: a case study. CLEI Electronic Journal. 2011 Apr;14(1):4-.

[5] Feldman R, Sanger J. The Text Mining Handbook: Advanced Approaches to Analyzing Unstructured Data.

[6] Korde V, Mahender CN. Text classification and classifiers: A survey. International Journal of Artificial Intelligence & Applications. 2012 Mar 1;3(2):85.

[7] Tsoumakas G, Katakis I. Multi-label classification: An overview. International Journal of Data Warehousing and Mining. 2006;3(3).

[8] Tsoumakas G, Katakis I, Vlahavas I. Mining multi-label data. InData mining and knowledge discovery handbook 2009 (pp. 667-685). Springer, Boston, MA.

[9] McCallum A. Multi-label text classification with a mixture model trained by EM. InAAAI workshop on Text Learning 1999 Jul 18 (pp. 1-7).

[10] Read J, Pfahringer B, Holmes G, Frank E. Classifier chains for multi-label classification. Machine learning. 2011 Dec 1;85(3):333.

[11] Read J, Pfahringer B, Holmes G. Multi-label classification using ensembles of pruned sets. InData Mining, 2008. ICDM'08. Eighth IEEE International Conference on 2008 Dec 15 (pp. 995-1000). IEEE.

[12] Tsoumakas G, Katakis I, Vlahavas I. Random k-labelsets for multilabel classification. IEEE Transactions on Knowledge and Data Engineering. 2011 Jul;23(7):1079-89.

[13] Zhang ML, Zhou ZH. ML-KNN: A lazy learning approach to multi-label learning. Pattern recognition. 2007 Jul 1;40(7):2038-48.

[14] Madjarov G, Kocev D, Gjorgjevikj D, Džeroski S. An extensive experimental comparison of methods for multi-label learning. Pattern recognition. 2012 Sep 1;45(9):3084-104.

[15] Zhang ML, Zhang K. Multi-label learning by exploiting label dependency. InProceedings of the 16th ACM SIGKDD

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

international conference on Knowledge discovery and data mining 2010 Jul 25 (pp. 999-1008). ACM.

[16] Gibaja E, Ventura S. Multi-label learning: a review of the state of the art and ongoing research. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 2014 Nov 1;4(6):411-44.

[17] Koshy S, Padmajavalli R. Feature Selection for Improving Multi-Label Classification using MEKA. International Journal of Applied Engineering Research. 2017;12(24):14774-82.

[18] http://meka.sourceforge.net/~Language Log

[19] Spolaôr N, Tsoumakas G. Evaluating feature selection methods for multi-label text classification. BioASQ workhsop. 2013 Sep 27.

[20] Liu H, Motoda H, editors. Computational methods of feature selection. CRC Press; 2007 Oct 29.

[21] Liu H, Motoda H, editors. Computational methods of feature selection. CRC Press; 2007 Oct 29.

[22] Saeys Y, Inza I, Larrañaga P. A review of feature selection techniques in bioinformatics. bioinformatics. 2007 Oct 1;23(19):2507-17.

[23] Spolaôr N, Monard MC, Lee HD. A systematic review to identify feature selection publications in multi-labeled data. Relatório Técnico do ICMC No. 2012;374(31):3.

[24] Forman G. An extensive empirical study of feature selection metrics for text classification. Journal of machine learning research. 2003;3(Mar):1289-305.

[25] Spolaôr N, Cherman EA, Monard MC, Lee HD. Filter approach feature selection methods to support multi-label learning based on relieff and information gain. InAdvances in Artificial Intelligence-SBIA 2012 2012 (pp. 72-81). Springer, Berlin, Heidelberg.

[26] Hall MA. Correlation-based feature selection for machine learning.

[27] Doquire G, Verleysen M. Mutual information-based feature selection for multilabel classification. Neurocomputing. 2013 Dec 25;122:148-55.

[28] Zhang ML, Peña JM, Robles V. Feature selection for multi-label naive Bayes classification. Information Sciences. 2009 Sep 9;179(19):3218-29.

[29] Pereira RB, Plastino A, Zadrozny B, Merschmann LH. Information gain feature selection for multi-label classification. Journal of Information and Data Management. 2015 Oct 12;6(1):48..

[30] Cai Y, Yang M, Yin H. Relieff-based multi-label feature selection. International Journal of Database Theory and Application. 2015;8(4):307-18.

[31] Jungjit S, Freitas AA. A new genetic algorithm for multi-label correlation-based feature selection. In23rd European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning 2015 (pp. 285-290).

[32] Post MJ, van der Putten P, van Rijn JN. Does feature selection improve classification? a large scale experiment in OpenML. InInternational Symposium on Intelligent Data Analysis 2016 Oct 13 (pp. 158-170). Springer, Cham.

[33] Read J, Pfahringer B, Holmes G. Multi-label classification using ensembles of pruned sets. InData Mining, 2008. ICDM'08. Eighth IEEE International Conference on 2008 Dec 15 (pp. 995-1000). IEEE.