International Journal of Applied Engineering Research ISSN 0973-4562 Volume 12, Number 24 (2017) pp. 14774-14782

Feature Selection for Improving Multi-Label Classification using MEKA

Susan Koshy
Research Scholar, Bharathiar University, Coimbatore, Tamil Nadu, India.
Assistant Professor, Department of Computer Science
St.Thomas College of Arts and Science, Chennai, Tamil Nadu, India.
Orcid Id: 0000-0001-8972-685X

Dr. R.Padmajavalli
Associate Professor, Department of Computer Applications,
Bhaktavatsalam Memorial College for Women, Chennai, Tamil Nadu, India.

Abstract

The extensive dimensionality in multi-label classification can be overcome by selecting representative words that describe an instance and removing the redundant and insignificant ones. The popular technique of feature selection when applied reduces the size of the dataset and hence speeds up and improves the accuracy of the learning process of classification. This paper looks at the performance of Correlation feature subset selection (CFS) on two datasets which have a large number of features to reduce the dimension of the dataset. Eventually multi-label classification using four problem transformation methods binary relevance, classifier chains, pruned sets and RAkEL and seven different single label base classifiers on the same data sets are evaluated for their performance metrics.

Keywords: Feature selection, multi-label text classification, problem transformation, algorithm adaptation, Correlation based feature subset selection

INTRODUCTION

Feature selection is an essential preprocessing method to solve the problem of dimensionality. The presence of redundant and irrelevant features hampers the learning process and involving feature selection greatly improves the learning process. 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 according to M.A Hall (1999) [26].

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. Similarly a news article about a conference on climate change can be labeled both politics and environment [25][2].

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

The technique of feature selection is to identify representative words and remove unimportant ones, which could make learning faster and even improve learning performance.

Feature selection will search the feature space $X = \{X1,X2,...,XM\}$ to find a good subset of features $X' \subseteq X$ which describes the dataset as well as the original set of features X

does. Such selection can help in building faster, cost effective and accurate models for data processing.

A feature selection process primarily consists of four steps namely subset generation, subset evaluation, stopping criterion, and result validation. Subset generation generates a feature subset which is the candidate for evaluation. The attribute subset selection may either start at a full, empty or random feature set. The generated subset is then evaluated by using an evaluation criterion which determines the worthiness of the subset. The subset generated is then compared with the best subset generated earlier. Based on a stopping criterion, further process of generation and evaluation of subsets is not continued. Last, the selected feature subset is validated with different tests using artificial or real world data.

A filter model for feature selection is independent of a classification algorithm and uses general characteristics of data for evaluating and selecting features. The filter approach works independently of the learning algorithm and irrelevant features are removed. It analyses the properties of a dataset and in this way chooses the appropriate features. Filters may not choose the best features for specific learning algorithms but they are fast and simple to implement. The filter is popular due to its less complex nature in multi-label feature selection. The wrapper approach requires a specific learning algorithm to determine and evaluate which features are selected. It has a high computational cost as it has to call the learning algorithm for whenever each feature set is considered, although it finds features which are better suited for the specific learning algorithm.

A hybrid or embedded model combines the two models into one framework. For particular learning algorithms where feature selection is also a part of the training process, the embedded model is beneficial. Feature selection methods based on a wrapper model are not suitable in large scale problems. Most of the feature selection methods are based on a filter model which evaluates each feature independently.

Some of the feature evaluation metrics to evaluate the goodness of features for classification are Fisher score, Chi square, ReliefF, Gini Index, Information Gain, 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. This paper discusses about correlation feature subset selection and it is the technique used for feature selection of the two datasets. Although it is a lesser preferred technique as it ignores label correlation it gives better results according to this paper.

The paper is discusses some of the related work in multi-label feature selection and the different techniques used. The next section gives the key issues followed by the general frame work of multi-label feature selection. The next section discusses two feature selection techniques correlation based feature selection and information gain. A broad outline of multi-label classification and the evaluation metrics are discussed. This is followed by the experimental results, discussion and conclusion.

DEFINITION OF THE PROBLEM

Consider the instance space of the d-dimension $X = \mathrm{R}^d$(or Z^d), and $Y = \{y_1, y_2, \ldots, y_q\}$ which represents the label space that has q possible class labels. The main objective of the multi-label learning here is learning a function $h{:}X \rightarrow 2^Y$that is from the training of the multi-label set $D = \{(\boldsymbol{x}_i, Y_i) \mid 1 \leq i \leq m\}$. For each multi-label instance $(\boldsymbol{x}_i, Y_i)$, $\boldsymbol{x}_i \in X$ is a d-dimensional feature of vector $(x_{i1}, x_{i2}, \ldots, x_{id})$and $Y_i \subseteq Y$ denotes the set of the labels that are associated with $\boldsymbol{x}_i$. Feature selection searches the feature vector to find a good subset of features which describes the dataset as well as the original set of features. For any particular instance that is unknown $\boldsymbol{x} \in X$, is the multi-label classifier $h(\cdot)$ that predicts $h(\boldsymbol{x}) \subseteq Y$ as the proper set of labels for $\boldsymbol{x}$.

RELATED WORKS

Min-Ling Zhang, Jose M. Pena and Victor Robles in their paper, 'Feature selection for multi-label naive Bayes' classification used a method called MLNB which are naive Bayes' classifiers incorporating two stage filter wrapper feature selection to handle multi-label instances in the year 2009[28].

Gauthier Doquire, Michel Verleysen in their work 'Mutual information-based feature selection for multi-label classification' proposed a method that uses the multivariate mutual information criterion along with a problem transformation and a pruning strategy in 2013. 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 [27].

Rafael B. Pereira, Alexandre Plastino, Bianca Zadrozny,Luiz H. C. Merschmann in their paper 'Information Gain Feature Selection for Multi-Label Classification' used Information Gain feature selection along with problem transformation techniques for multi-label classification in 2015 [29].

Yaping Cai, Ming Yang, Yang Gao and Hujun Yin in their paper 'ReliefF-based Multi-label Feature Selection' in 2015 have used a strategy ML-ReliefF, to select distinguishing features to improve multi-label classification accuracy along with the ML-KNN classifier [30].

Suwimol Jungjit and Alex A. Freitas in their paper 'A New Genetic Algorithm for Multi-Label Correlation-Based Feature

Selection'(2015) have proposed a new genetic algorithm in order to search for highly relevant subsets [31].

Suwimol Jungjit, Alex A. Freitas, M. Michaelis and J. Cinatl in their paper 'Two Extensions to Multi-label Correlation-Based Feature Selection: a case study in bioinformatics(2015)' have proposed two extensions to the correlation based subset feature selection and have incorporated mutual information for finding the weights of class labels [32].

Key issues

The output size of the multi-label dataset is normally very huge and the sets of labels grow exponentially and this is the prime challenge that is faced in multi label learning. For any label space that has class labels that are 20 in number, where q = 20, the possible label set number may exceed one million which will be 2^{20}. In order to cope with this kind of a challenge of output space that is exponentially sized, it becomes necessary that correlations or dependency factors among labels are considered [7]. The current strategies here are grouped into three different families which are on the basis of the order of correlation that is needed for the techniques of learning that are used. The multi label learning task in first order is managed label-by-label and the coexistence has been ignored with that of other labels [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 modeling than the second and first order strategies but the computation is more strenuous and not easily scalable [8][16].

Multi-label feature selection for classification

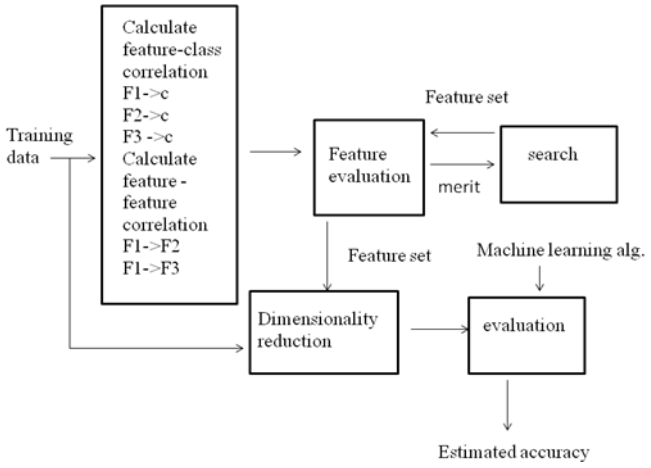

Figure 1: A General Framework of Feature Selection for Classification

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].

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].

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 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].

Multilabel 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].

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].

Evaluation metrics for multi label classification

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

$$\text{Hamming Loss}=\frac{1}{N}\sum_{i=1}^{N}\frac{\left|Y_i \Delta Z_i\right|}{M} \quad (4)$$

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

$$\text{Accuracy}=\frac{1}{N}\sum_{i=1}^{N}\frac{\left|Yi \cap Zi\right|}{\left|Yi \cup Zi\right|} \quad (5)$$

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.

$$\text{Precision}=\frac{1}{N}\sum_{i=1}^{N}\frac{\left|Yi \cap Zi\right|}{\left|Zi\right|} \quad (6)$$

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

$$\text{Recall}=\frac{1}{N}\sum_{i=1}^{N}\frac{\left|Yi \cap Zi\right|}{\left|Yi\right|} \quad (7)$$

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

$$\text{F-Measure}=\frac{1}{N}\sum_{i=1}^{N}\frac{2\left|Yi \cap Zi\right|}{\left|Zi\right|-\left|Yi\right|} \quad (8)$$

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.

$$\text{Subset Accuracy}=\frac{1}{N}\sum_{i=1}^{N} I\left(\left|Zi\right|=\left|Yi\right|\right) \quad (9)$$

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 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].

EXPERIMENTAL RESULTS

MEKA, an extension of WEKA framework is used to evaluate two different multi-label datasets ENRON and YELP. MEKA which is based on WEKA Toolkit of Machine Learning further includes many multi-label methods from the literature. WEKA developed by Waikato University is open source software that is issued under General Public License or GNU. ENRON [18] is a subset of Enron text corpus of email. This is based on an email collection that was duly exchanged between the employees of Enron Corporation which were made publically available during a legal investigation. It contains emails 1702 in number which have 1054 attributes and 53 labels. The YELP [17] dataset is about reviews of food in restaurants which contains 1960 instances with 676 attributes and 8 labels. Both these sets of data are in ARFF (Attribute rich file format). ENRON is downloaded from mulan.sourceforge.net and the Yelp from yelp.com and both these datasets belong to the text domain.

These experiments are run on 32bit machine that has a clock speed of 3.40 GHz and due to this factor other datasets from different domains and feature selection methods such as information gain and gain ratio could not be evaluated. The methods of problem transformation such as RAkEL, classifier chains, Pruned sets and binary relevance are used along with different base single label classifiers like BN or Bayesian Network, J48, NB or Naïve Bayes, kNN or K nearest Neighbor, AdaB or AdaBoost, ZR or ZeroR and RF or Random Forest. The metrics of evaluation like AC or accuracy from the equation 2, HL or Hamming loss from the equation 1 and ZO or Zero One Loss, RL or Rank Loss, AP or Average precision and TT or total time is taken for the purpose of analysis.

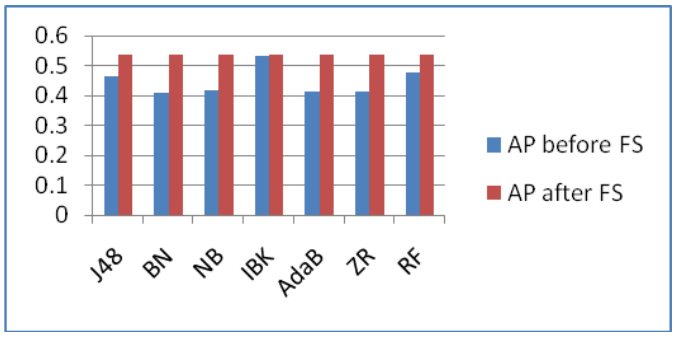

Figure 2: Average precision of various base classifiers with pruned sets multi-label classifier before and after feature selection on ENRON datasets

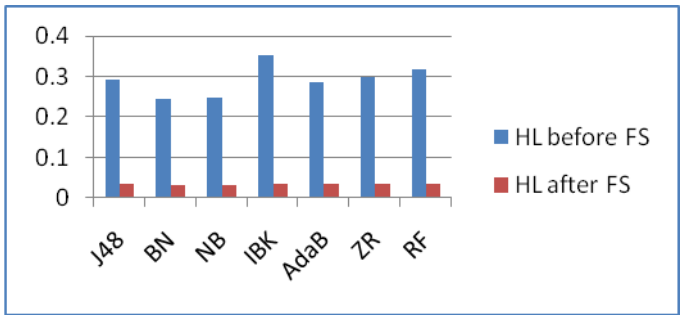

Figure 3: Hamming loss of various base classifiers with pruned sets multi-label classifier before and after feature selection on ENRON datasets

DISCUSSION

Feature selection of multi-label instances is gaining importance and relevance with the enormous application scope and in the light of this need a systematic review has been done regarding the availability of scientific publication. It has been found that nearly 61.2% of the publications have been on feature selection using the filter approach and there is a lack of publication regarding label dependence and correlation among the labels [23]. In this paper seven single label base classifiers of WEKA available on the MEKA framework namely BN or Bayesian Network, J48, NB or Naïve Bayes, kNN or K nearest neighbor, AdaB or AdaBoost, ZR or ZeroR and RF or Random Forest are taken for evaluation on four methods of problem transformation which are BR or Binary Relevance, PS or Pruned sets, RAkel and CC or Classifier Chains. Binary relevance method is the simplest multi-label classification technique which ignores the correlations of the labels and has the problem of class imbalance [1]. A base classifier from every category namely functional, neural and statistical is chosen and their performance is duly observed.

Table 1 to 8 show the evaluation metrics of these datasets which were classified after feature selection and the evaluation measures prior to feature selection are indicated in brackets. The figures 2 and 3 indicate the difference in evaluation metrics of average precision and hamming loss respectively on the dataset ENRON before and after feature selection. Attribute feature selection was used and the evaluation method was correlation based subset feature selection and the search technique is best fit. Evaluation of the accuracy of all classifiers is based on cross validation. In our earlier paper the same datasets have been used and evaluation without feature selection resulted in high time taken factor to build the model but after applying feature selection it has drastically reduced. The number of features after applying CFS feature selection has reduced from 1054 to 39 attributes in the case of ENRON dataset and from 676 to 27 attributes in the case of Yelp dataset which means that 94% of the features having been removed. Yet the accuracy of the classifier has not diminished and irrespective of the base classifier used

there is uniformity in the metrics for every multi-label classifier used. The best performance is with pruned sets multi-label classifier for both the datasets, Enron and Yelp with respect to the total time to build the model (TT).

Binary relevance on YELP				
	AC	**HL**	**AP**	**TT**
J48	0.455(.41)	0.033(.28)	0.569(.42)	0.086(26)
BN	0.456(.44)	0.033(.24)	0.575(.4)	0.094(2)
NB	0.456(.42)	0.033(.25)	0.575(.4)	0.084(1.4)
IBK	0.456(.28)	0.033(.35)	0.576(.46)	1.078(65)
AdaB	0.456(.41)	0.033(.3)	0.57(.43)	0.451(3.5)
ZR	0.454(.39)	0.033(.3)	0.577(.43)	0.078(.25)
RF	0.456(.47)	0.033(.22)	0.576(.43)	0.08(21)

Table 1: Evaluation metrics of various single label classifiers for binary relevance problem transformation on YELP

Pruned Set on YELP				
	AC	**HL**	**AP**	**TT**
J48	0.455(.36)	0.034(.29)	0.537(.46)	0.013(3.3)
BN	0.456(.44)	0.033(.24)	0.537(.41)	0.016(4.8)
NB	0.456(.43)	0.033(.24)	0.537(.41)	0.013(4.1)
IBK	0.455(.22)	0.034(.35)	0.537(.53)	0.045(8.1)
AdaB	0.455(.35)	0.034(.28)	0.537(.41)	0.047(.12)
ZR	0.454(.33)	0.034(.29)	0.537(.41)	0.009(.05)
RF	0.455(.32)	0.034(.31)	0.537(.47)	0.006(.16)

Table 2: Evaluation metrics of various single label classifiers for pruned set problem transformation on YELP

Classifier Chains on YELP				
	AC	**HL**	**AP**	**TT**
J48	0.456(.44)	0.033(.23)	0.537(.43)	0.404(22)
BN	0.454(.45)	0.033(.24)	0.537(.42)	0.131(1.4)
NB	0.454(.44)	0.033(.24)	0.537(.42)	0.081(1.1)
IBK	0.456(.25)	0.034(.33)	0.537(.49)	2.807(64)
AdaB	0.456(.39)	0.033(.22)	0.537(.45)	0.554(3.3)
ZR	0.454(.33)	0.033(.23)	0.537(.45)	0.045(.16)
RF	0.456(.34)	0.034(.29)	0.537(.45)	0.137(.44)

Table 3: Evaluation metrics of various single label classifiers for classifier chain problem transformation on YELP

RAkEL on YELP				
	AC	**HL**	**AP**	**TT**
J48	0.455(.47)	0.033(.29)	0.537(.39)	0.033(61)
BN	0.456(.46)	0.033(.23)	0.537(.41)	0.034(4)
NB	0.456(.45)	0.033(.24)	0.537(.4)	0.028(4)
IBK	0.456(.24)	0.033(.35)	0.537(.5)	0.418(80)
AdaB	0.456(.4)	0.033(.22)	0.537(.45)	0.083(.81)
ZR	0.454(.33)	0.033(.23)	0.025(.45)	0.454(.45)
RF	0.456(.38)	0.033(.28)	0.537(.42)	0.027(.9)

Table 4: Evaluation metrics of various single label classifiers for RAkEL problem transformation on YELP

Binary Relevance on ENRON				
	AC	**HL**	**AP**	**TT**
J48	0.338(.37)	0.063(.06)	0.509(.11)	0.13(169)
BN	0.339(.14)	0.063(.09)	0.423(.12)	0.15(20)
NB	0.339(.15)	0.063(.1)	0.423(.1)	0.13(18)
IBK	0.339(.31)	0.063(.07)	0.422(.07)	0.927(61)
AdaB	0.339(.38)	0.063(.05)	0.423(.1)	0.646(80)
ZR	0.315(.32)	0.069(.07)	0.422(.09)	0.123(7)
RF	0.339(.48)	0.063(.05)	0.422(.1)	0.126(198)

Table 5: Evaluation metrics of various single label classifiers for binary relevance problem transformation on ENRON

Pruned sets on ENRON				
	AC	**HL**	**AP**	**TT**
J48	0.341(.35)	0.062(.06)	0.414(.07)	0.016(12)
BN	0.341(.35)	0.062(.05)	0.414(.07)	0.025(60)
NB	0.341(.33)	0.062(.06)	0.414(.07)	0.022(40)
IBK	0.341(.32)	0.062(.06)	0.414(.07)	0.036(1)
AdaB	0.341(.27)	0.062(.06)	0.414(.07)	0.062(.43)
ZR	0.315(.17)	0.069(.07)	0.414(.07)	0.008(.1)
RF	0.341(.33)	0.062(.06)	0.414(.07)	0.011(1)

Table 6: Evaluation metrics of various single label classifiers for pruned set problem transformation on ENRON

Classifier Chain on ENRON				
	AC	**HL**	**AP**	**TT**
J48	0.34(.37)	0.063(.05)	0.41(.07)	1.17(19)
BN	0.341(.19)	0.063(.22)	0.414(.08)	0.212(27)
NB	0.341(.2)	0.063(.2)	0.414(.08)	0.159(21)
IBK	0.34(.32)	0.063(.06)	0.414(.07)	0.803(66)
AdaB	0.341(.31)	0.062(.05)	0.415(.07)	1.209(116)
ZR	0.315(.14)	0.062(.06)	0.414(.07)	0.079(7)
RF	0.34(.31)	0.063(.07)	0.414(.07)	0.318(11)

Table 7: Evaluation metrics of various single label classifiers for classifier chain problem transformation on ENRON

RAkEL on ENRON				
	AC	**HL**	**AP**	**TT**
J48	0.332(.02)	0.065(.06)	0.415(.07)	0.033(56)
BN	0.332(.04)	0.065(.11)	0.415(.07)	0.037(13)
NB	0.332(.04)	0.065(.1)	0.415(.07)	0.028(16)
IBK	0.332(.03)	0.065(.06)	0.415(.07)	0.295(28)
AdaB	0.332(.01)	0.065(.06)	0.415(.07)	0.087(34)
ZR	0.315(0)	0.069(.06)	0.415(.07)	0.023(32)
RF	0.332(.03)	0.065(.07)	0.415(.07)	0.025(1.8)

Table 8 Evaluation metrics of various single label classifiers for RAkEL problem transformation on ENRON

CONCLUSION

Feature selection has reduced the dimension of the dataset by 94% and has improved the evaluation parameters in multi-label classification. In this paper seven base classifiers BN or Bayesian Network, J48, NB or Naïve Bayes, kNN or K nearest neighbor, AdaB or AdaBoost, ZR or ZeroR and RF or Random Forest are taken for evaluation on the four methods of transformation which are BR or Binary Relevance, PS or Pruned sets, RAkEL and Classifier Chains have been used. All the evaluation metrics for multi-label classification with feature selection have improved after the dimensionality reduction technique. A correlation feature subset selection evaluation and a best fit search technique gave improved evaluation metrics for all the four multi-label problem transformation methods uniformly.

REFERENCES

[1] Zhang, M. L., & Zhou, Z. H. (2014). A review on multi-label learning algorithms. IEEE transactions on knowledge and data engineering, 26(8), 1819-1837.

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

[3] Han, J., & Kamber, M. (2001). Simon Fraser University "Data Mining Concepts and Techniques".

[4] Cherman, E. A., Monard, M. C., & Metz, J. (2011). Multi-label problem transformation methods: a case study. CLEI Electronic Journal, 14(1), 4-4.

[5] Feldman, R., & Sanger, J. (2007). The text mining handbook: advanced approaches in analyzing unstructured data. Cambridge university press.

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

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

[8] Tsoumakas, G., Katakis, I., & Vlahavas, I. (2009). Mining multi-label data. In Data mining and knowledge discovery handbook (pp. 667-685). Springer US.

[9] McCallum, A. (1999, July). Multi-label text classification with a mixture model trained by EM. In AAAI'99 workshop on text learning (pp. 1-7).

[10] Read, J., Pfahringer, B., Holmes, G., & Frank, E. (2009, September). Classifier chains for multi-label classification. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases (pp. 254-269). Springer Berlin Heidelberg.

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

[12] Tsoumakas, G., Katakis, I., & Vlahavas, I. (2011). Random k-labelsets for multilabel classification. IEEE Transactions on Knowledge and Data Engineering, 23(7), 1079-1089.

[13] Zhang, M. L., & Zhou, Z. H. (2007). ML-KNN: A lazy learning approach to multi-label learning. Pattern recognition, 40(7), 2038-2048.

[14] Madjarov, G., Kocev, D., Gjorgjevikj, D., & Džeroski, S. (2012). An extensive experimental comparison of methods for multi-label learning. Pattern Recognition, 45(9), 3084-3104.

[15] Zhang, M. L., & Zhang, K. (2010, July). Multi-label learning by exploiting label dependency. In Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 999-1008). ACM.

[16] Gibaja, E., & Ventura, S. (2014). Multi-label learning: a review of the state of the art and ongoing research. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 4(6), 411-444.

[17] http://www.ics.uci.edu/~vpsaini/-yelp dataset challenge

[18] https://www.cs.cmu.edu/~./enron

[19] Spolaôr, Newton, and Grigorios Tsoumakas. "Evaluating feature selection methods for multi-label text classification." BioASQ workhsop (2013).

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

[21] Zhao, Z., Morstatter, F., Sharma, S., Alelyani, S., Anand, A., Liu, H.: Advancing feature selection research - ASU feature selection repository. Technical Report (2011), Arizona State University

[22] Saeys, Yvan, Iñaki Inza, and Pedro Larrañaga. "A review of feature selection techniques in bioinformatics." bioinformatics 23.19 (2007): 2507-2517

[23] Spolaôr, N., M. C. Monard, and H. D. Lee. "A systematic review to identify feature selection publications in multi-labeled data." Relatório Técnico do ICMC No 374.31 (2012):

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

[25] Azam, Nouman, and JingTao Yao. "Comparison of term frequency and document frequency based feature selection metrics in text categorization."Expert Systems with Applications 39.5 (2012): 4760-4768.

[26] Hall, Mark Andrew. "Correlation-based feature selection for machine learning." (1999).

[27] Doquire, Gauthier, and Michel Verleysen. "Mutual information-based feature selection for multilabel classification." Neurocomputing 122 (2013): 148-155.

[28] Zhang, Min-Ling, José M. Peña, and Victor Robles. "Feature selection for multi-label naive Bayes classification." Information Sciences 179.19 (2009): 3218-3229.

[29] Pereira, Rafael B., et al. "Information gain feature selection for multi-label classification." Journal of Information and Data Management 6.1 (2015): 48.

[30] Cai, Yaping, Ming Yang, and Hujun Yin. "Relieff-based multi-label feature selection." International Journal of Database Theory and Application 8.4 (2015): 307-318.

[31] Jungjit, Suwimol, and Alex A. Freitas. "A New Genetic Algorithm for Multi-Label Correlation-Based Feature Selection." Proceedings. Presses universitaires de Louvain, 2015.

[32] 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.

[33] J.-J. Lamben. “Strategic Marketing. The European Perspective (Translated from French)”. St. Petersburg: Nauka, 1996.

[34] E.V. Tolkacheva. “Strategic Controlling in the Enterprise Management System”, Management in Russia and Abroad, vol. 4, pp. 109-118, 2004.

[35] S.M. Rezer. “Fundamentals of modeling optimal logical systems delivery. Transport Innovations”, Scientific and Technical Journal, vol. 3(18), 2014.