DOI: 10.21917/ijsc.2014.0122

APPLICATION OF RANKING BASED ATTRIBUTE SELECTION FILTERS TO PERFORM AUTOMATED EVALUATION OF DESCRIPTIVE ANSWERS THROUGH SEQUENTIAL MINIMAL OPTIMIZATION MODELS

C. Sunil Kumar[1] and R.J. Rama Sree[2]
[1]*Research and Development Center, Bharathiar University, India*
E-mail: sunil_sixsigma@yahoo.com
[2]*Rashtriya Sanskrit Vidyapeetha, India*
E-mail: rjramasree@yahoo.com

Abstract

In this paper, we study the performance of various models for automated evaluation of descriptive answers by using rank based feature selection filters for dimensionality reduction. We quantitatively analyze the best from amongst five rank based feature selection techniques, namely Chi Squared filter, Information Gain filter, Gain Ratio filter, Relief filter and Symmetrical Uncertainty filter. We use Sequential Minimal Optimization with Polynomial kernel to build models and we evaluate these models in terms of various parameters such as Accuracy, Time to build the models, Kappa, Mean Absolute Error and Root Mean Squared Error. For all except the Relief filter, the accuracies obtained are at least 4% better than the accuracies obtained with the same models without any filters applied. We found that the accuracies recorded are same for Chi Squared filter, Information Gain filter, Gain Ratio filter and Symmetrical Uncertainty filter. Therefore accuracy alone is not the sole determinant in selecting the best filter. The time taken to build the models, Kappa, Mean Absolute Error and Root Mean Squared Error play a major role in determining the effectiveness of these filters. The overall rank aggregation metric of Symmetrical Uncertainty filter is 45 and this is better by 1 rank than the rank aggregation metric of Information gain filter. Symmetric Uncertainty filter's rank aggregation metric is better by 3, 6, 112 ranks respectively when compared to the rank aggregation metrics of Chi Squared filter, Gain Ratio filter and Relief filters. Through these quantitative measurements, we conclude that Symmetrical Uncertainty attribute evaluation is the overall best performing rank based feature selection algorithm applicable for auto evaluation of descriptive answers.

Keywords:

Descriptive Answers, Text Classification, Rank Based Filters, Feature Selection, Dimensionality Reduction

1. INTRODUCTION

Evaluation of answers and providing a scoring is a challenging task. Assignment of only a single category to each answer makes this task a hard classification task. The human evaluator or the system is supposed to interpret the answer and classify the answer into one of the possible rubrics pre-allocated for the answer. We believe supervised learning method can be applied to classify the answers into appropriate rubrics based on the likelihood suggested by training samples. The supervised learning process requires extracting various text features from the documents meant as training set and then train using a sophisticated machine learning algorithm. One particular problem with this text classification task is that depending on the document size, the number of features can be very large, sometimes spanning into thousands too! The huge number of features is a major problem for a training algorithm to perform effective learning and prediction. Feature selection helps sort out this problem of huge number of features. There are two ways available to reduce the number of features, the first one is Feature selection which eliminates the un-required features from the complete set of features - this means only some of the key features contributing significantly to the model's performance are chosen and used [1]. The other approach is Feature transformation which computes new features that are functions of the old features i.e., the reduced new features somehow inherently represent the old features [2]. Techniques such as principal component analysis do the task of identifying patterns within high dimension data and then compressing i.e. by reducing the number of dimensions, without much loss of information [3] [4].

For the scope of this paper, we focused our research on feature selection only and not feature transformation. The context is to identify and eliminate irrelevant or redundant features in the data that makes the knowledge discovery process during training more difficult, thereby making the data used in training less noisy and more reliable. We have employed various rank based feature selection or attribute selection techniques that utilize a combination of search and attribute utilization estimation to rank the attributes in the data. Application of any feature selection filter assigns a significance score and a rank to each of the features used in experimentation. Significance score represents the importance of the feature in the model prediction task. Feature significance score is directly proportional to the rank assigned. In other words, a rank of 1 is assigned to the feature that got the highest significance score, rank 2 is assigned to the feature with second highest score and so on. Feature selection filter also allows a threshold to be specified and this threshold controls the number of features to be retained post the application of the filter. In our experiments, we set this threshold to 0, which means any feature that is assigned a 0 significance score is eliminated from the output obtained. The output thus obtained is the reduced feature set which is further consumed in the experiments. Models were built using the reduced datasets obtained and the performance of the models were measured from different perspectives viz., memory occupied by final training set, time taken for training and of course the correct number of predictions. Finally, the best performing model and contributing feature selection method was selected based on the metrics obtained.

The rest of this paper is organized as follows. Section 2 details the previous work undertaken in this area of research and

the research motivation that makes the research covered under this paper unique from others. Section 3 discusses the data used, experimental setup, the preliminaries of the tools and techniques used. Section 4 describes the models built and measurements made during the experiments. Analysis of experimental results is dealt with in section 5. Finally, concluding remarks and further research plans are indicated in section 6.

2. RELATED WORK AND RESEARCH MOTIVATION

Multiple research works were previously carried out on dimensionality reduction in text classification. They range from application of filters, wrappers to data from various domains to developing hybrid filtering techniques for effective selection of attributes. Some key details on such research contributions are introduced in this section. Details on differentiating factors for the research covered under this paper with the previous research is also described in this section.

The work in [5] discusses various developments in machine learning to the problem of selecting relevant features, and the problem of selecting relevant examples which is otherwise the feature selection problem currently dealt in this research. Various feature selection methods were surveyed in detailed in [6] [7]. In [8], the authors discuss an unsupervised feature selection algorithm that makes use of a randomly selected sampling technique. In [9], the authors present a comparative study of few alternatives of five most prevalent feature selection methods. In [10], the authors compared chi-square, information gain, document frequency, mutual information and term strength on the Reuters and Ohsumed datasets. K-Nearest neighbours unsupervised learning algorithm and Linear Least Square Fit classification algorithms were used to obtain the classification. From the research, it was confirmed that information gain and chi-square filters are the best filters. In [11], the authors also reveal that information gain and chisquared attribute evaluation filters as best filters among the 12 filters the experiments were conducted with. This conclusion was derived based on generalization of results obtained from 19 multi-class data sets that contained 229 binary text classification problems. In [12], the authors detail the results obtained from fifteen standard machine learning data sets from the UCI collection and it suggests that Relief rank based filter is one of the best overall performer apart from wrappers. Interestingly, Information Gain filter was also part of this research and this is not the filter that was proved as the best in this research.

All the results reported in the previous research efforts focused on accuracy for concluding the best filter and used a variety of filters rather than comparing a specific category of filters. We believe a comparison is required between filters of the same category i.e., rank based filters. Hence our motivation for the research is covered under this article. We choose to compare Chi-square (CS), Gain Ratio (GR), Info gain (IG), Relief, Symmetrical Uncertainty (SU) rank based filters for our research as all of these rank features for their selection. Another perspective to our research is to evaluate and compare the rank based filters not just based on classification accuracy obtained but also based on model training time and other error metrics such as kappa, Root Mean Squared Error, Mean Absolute Error. Also, rather than generalizing the results from various domains and deriving a conclusion, we want to explore the behaviour of results for our specialized domain i.e., descriptive answers. All these research motivations make our research standout from previous researches done in this area.

3. EXPERIMENTAL SETUP

The setup in which the experiments are conducted for this paper are specified in this section.

3.1 DATA COLLECTION

In February 2012, The William and Flora Hewlett Foundation (Hewlett) sponsored the Automated Student Assessment Prize (ASAP) [26] to machine learning specialists and data scientists to develop an automated scoring algorithm for student-written essays. As part of this competition, the competitors are provided with hand scored essays under 8 different prompts which are questions to which answers were obtained from Students. These answers were the datasets. 5 of the 8 essays prompts are used for the purpose of this research.

All the graded essays from ASAP are according to specific data characteristics. All responses were written by students of Grade 10. On an average, each essay is approximately 50 words in length. Some are more dependent upon source materials than others. The number of essays obtained for each prompt vary, for example the lowest amount of documents among the training data sets is 1190 whereas the highest is 1982 [26]. All the documents are in ASCII text followed by a human score, a resolved final score was given in cases there is a variance found with scores provided by two human scorers [27]. For the purpose of evaluation of the performance of the model, we considered the score predicted by the model to comply with the resolved human score in training example.

The data used for training, validation and testing the models are answers written by students for 5 different questions. Data for a question is considered as one unique dataset. So, we have a total of 5 datasets. The questions that students are asked to provide responses to are from Chemistry, English Language Arts and Biology.

3.2 DATA CHARACTERISTICS

In each of the 5 training data sets used for our research, the training set is 900 samples in size. These 900 training samples were randomly picked from the total available samples under each prompt. Our previous research for determining appropriate sample size for automated essay scoring using Sequential Minimal Optimization (SMO) [13] revealed that using 900 samples for training proved to yield slightly better results than using other sample sizes [14] therefore the decision to use 900 samples as the training sample size.

3.3 WEKA WORKBENCH

For the purpose of designing and evaluating our experiments, we have used a machine learning workbench called Weka. Weka stands for “Waikato Environment for Knowledge Analysis” and it is a free offering from University of Waikato, New Zealand. This workbench has a user-friendly interface and it incorporates

numerous options to develop and evaluate machine learning models [15] [16]. These models can be utilized for a variety of purposes, including automated essay scoring.

All experiments performed were executed on a Dell Latitude E5430 laptop. The laptop is configured with Intel Core i5 - 3350M CPU @ 2.70 GHz and with 4 GB RAM however Weka workbench is configured to use a maximum of 1 GB. The laptop runs on Windows 7 64 bit operating system.

3.4 STATISTICAL FEATURE EXTRACTION

The features enumerated below are extracted from the input training data set to build feature tables –

- Unigrams - An n-gram of size 1 is referred to as a "unigram" [28].
- Bigrams - An n-gram of size 2 is a "bigram" or "digram" [28].
- Trigrams - An n-gram of size 3 is a "trigram" [28].
- Stop words - The most common, short function words, such as the, is, at, which, and on.
- Stemming - It is a process of reducing inflected (or sometimes derived) words to their stem, base or root form-generally a written word form [17]. Porter stemmer is used for stemming purpose in this research.
- Punctuations - unigrams representing items such as periods, commas, or quotation marks.

In the final dataset used for training and testing, we included Unigrams, Bigrams and Trigrams and implemented stemming on them. Stop words and Punctuations are excluded from the training, test datasets.

3.5 FEATURE SELECTION

A feature is selected based on how it affects the predictive capability of the models. If a feature contributes positively in predicting the outcome from the model then such feature is considered relevant for the model. Another perspective that is employed in feature selection is that the feature in consideration should not be highly correlated with another one i.e. it should not be an indirect representative of another feature in the model. Therefore identifying a good feature set involves finding those features that are not highly correlated as well as features that contribute towards prediction task.

Feature selection can be accomplished through wrapper and filter methods. Wrappers depend heavily on classification algorithm to measure the prominence of a feature to be included in the model. Feature selection through wrappers generally performs better than filters because the filter selection is optimized for the particular learning algorithm to be used [18]. The downside of using wrappers is that one needs to know the classification algorithm to be used prior to implementation of feature selection through wrappers. Another downside is that wrappers are very time taking and they are computationally expensive as features are evaluated with the chosen classification algorithm prior to finalizing the worthiness of features. Filters based feature selection evaluate the usefulness of features in prediction independent of any learning algorithm. Filters are fast and are computationally more efficient but totally ignore the dependency of features' worthiness on learning algorithms [18].

Most attribute evaluation filters work in conjunction with rank searching. Features are ranked and a specific number of features falling below the user specified threshold are discarded from the feature set included for the purpose of model building.

With the availability of multiple wrappers and filters there are numerous permutations possible to derive features that are appropriate for model building. To reduce the number of permutations possible, we focused on rank based individual feature evaluating filters. Other factor that influenced this decision is the computational efficiency and faster processing that the filters offer.

Entropy forms the basis for IG, GR, and SU filters used in this research. Information theory commonly makes use of the concept of Entropy [17]. Entropy represents a measure of the system's randomness. Entropy is generally represented by *H* which stands for the Greek Alphabet Eta. The authors of [18] and [29] give a detailed review of the concept of entropy as used in information theory.

The filters enumerated below are used for the purpose of this research:

1) Chi Squared (*CS*) Attribute Evaluation: This filter computes the chi-squared statistic of each attribute with respect to the class. Additional inputs about *CS* Attribute Evaluation filter are available in [19] [30].
2) Information Gain (*IG*) Attribute Evaluation: This filter evaluates the worth of an attribute by measuring the *IG* with respect to the class [20] [31].

   Information Gain is given by the Eq.(1) –

   $$IG(Class, Attribute) = H(Class) - H(Class|Attribute) \quad (1)$$
3) Gain Ratio (*GR*) Attribute Evaluation: *GR* filter evaluates the worth of an attribute by measuring the gain ratio with respect to the class. It is a non-symmetrical filter that compensates for certain bias issues found with information gain attribute selection filter as described in [21] [32].

   Gain Ratio is given by the Eq.(2) –

   $$GR(Class, Attribute) = \left(H(Class) - \frac{H(Class \mid Attribute)}{H(Attribute)} \right) \quad (2)$$
4) Symmetrical Uncertainty (*SU*) Attribute Evaluation: *SU* filter evaluates the worth of an attribute by measuring the symmetrical uncertainty with respect to the class. This attribute evaluation filter compensates for the bias in Information Gain [33]. In [22], the author explains that symmetrical uncertainty of a class and an attribute can be availed through the Eq.(3) -

   $$SU(X,Y) = \frac{2 * GAIN(X \mid Y)}{H(X) + H(Y)} \quad (3)$$

   where, $H(X)$ is the entropy of the discreet random variable.

   $$X.GAIN(X|Y) = H(X) - H(X|Y) \quad (4)$$

   where, $H(X|Y)$ is the conditional entropy which quantifies the remaining uncertainty of a random variable given that the value of another random variable is known.
5) Relief Attribute Evaluation: This filter evaluates the worth of an attribute by repeatedly sampling an instance and

considering the value of the given attribute for the nearest instance of the same and different class [23]. This attribute evaluation filter can operate on both discrete and continuous class data [34].

3.6 MODEL BUILDING

All models are built using John Platt's sequential minimal optimization algorithm [24] for training a support vector classifier and polynomial kernel is used along with other default parameters as available in Weka.

4. MODELS BUILT AND MEASUREMENTS

Various models are built during the experiments, the measurements obtained and various conclusions made through analysis of the measurements done during the experiments are described in this section.

The 5 attribute evaluation filters were applied on the 5 datasets separately with a threshold of 0. Threshold of 0 actually eliminates all features that are of no or less significance. Reduced feature sets for all 5 datasets are independently arrived at by eliminating all features which fall below the threshold value of 0.

Now that the reduced feature sets are arrived at, models are built on Weka workbench, we used randomized 10-fold cross-validation in order to testing performance the models.

The Models' reliability is captured through Kappa, Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) metrics.

Kappa statistic is used to measure the agreement between predicted and observed categorizations of a dataset, while correcting for an agreement that occurs by chance. However, like the plain success rate, it does not take costs into account [25] [35] [36].

To define Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE), assume the predicted values on the test instances are $p_1, p_2, \ldots, p_n$; the actual values are $a_1, a_2, \ldots, a_n$.

MAE is given by the Eq.(5) [37] –

$$MAE = |p_1 - a_1| + |p_2 - a_2| + \ldots + |p_n - a_n| \quad (5)$$

RMSE is given by the Eq.(6) [35] –

$$RMSE = \sqrt{(p_1 - a_1)^2 + (p_2 - a_2)^2 + \ldots + (p_n - a_n)^2} \quad (6)$$

Better models will have Kappa closer to 1 [36], MAEs and RMSEs closer to zero [25] [37]. Time to build the models is also captured across the models for comparison purposes.

4.1 REDUCTION OF ATTRIBUTES WITH RANK BASED FILTERS

For each of the 5 data sets, we measured the number of features that got retained when a filter is applied. For comparison purpose we also retained the initial number of attributes that were present in the dataset without any attribute selection filter application on the dataset. Table.1 shows the number of attributes retained in each of the datasets with the application of each Attribute selection filter. The table also shows the number of attributes in each dataset without the application of an attribute selection filter application on the datasets.

4.2 ACCURACY OBTAINED WITH MODELS ON DATASETS WITH 10 FOLD CROSS VALIDATION

For each of the reduced data sets, we built models using SMO. In Table.2, 10 fold cross validation is performed so as to obtain the accuracy. For ease of comparison and to make the accuracy percentages meaningful, the accuracy percentages were rounded to the nearest integer. Table.2 shows the accuracies obtained using 10 fold cross validation with SMO models. The individual reduced datasets obtained by application of attribute selection filters and the dataset with no attribute selection were used to build the models. Accuracy in this context is the percentage of correct scores predicted by the model using the 10 fold cross validation technique.

4.3 TIME TAKEN TO BUILD THE MODELS

We captured the time taken to build and test the models through 10 fold cross validation on Weka workbench for comparison purposes. Again, we retained the models built with no feature selection applied to contrast between the models. Table.3 shows the time taken to build models using the reduced datasets obtained by application of attribute selection filters and dataset where no attribute selection filter is applied.

4.4 ERROR METRICS MEASURED FOR MODELS PERFORMANCE

Kappa statistic, MAE and RMSE are captured for models built for data sets where feature selection is not applied as well as for models built using data sets where feature selection is applied to use them for comparison purposes. Table 4 shows the Kappa, MAE and RMSE recorded with 10 fold cross validation on the various datasets that were considered for experimentation.

Table.1. Reduction in number of features with attribute selection filter application on datasets

Data Set	Number of features with no attribute selection applied	Number of features retained with application of attribute selection filters				
		Chi square	Gain Ratio	Info gain	Relief	Symmetrical
1	25190	254	254	254	437	254
2	22847	126	126	126	462	126
3	29475	400	400	400	406	400
4	20915	378	378	378	519	378
5	19599	373	373	373	559	373

Table.2. Accuracy (%) obtained with 10 fold cross validation on datasets

Data Set	Accuracy obtained with dataset with no attribute selection and with 10 fold cross validation	Accuracies obtained with datasets using attribute Selection Filters and with 10 fold cross validation				
		Chi Square	Gain Ratio	Info Gain	Relief	Symmetrical
1	52	60	60	60	57	60
2	50	64	64	64	48	64
3	72	77	77	77	69	77
4	80	87	87	87	81	87
5	86	90	90	90	87	90

Table.3. Time taken (in seconds) to build the models

Data Set	Model building time with data set where no attribute selection is applied	Model building times with datasets obtained from application of various attribute selection filters				
		Chi Square	Gain Ratio	Info Gain	Relief	Symmetrical
1	2.6	0.68	0.5	0.61	0.73	0.71
2	2.95	0.23	0.24	0.25	1.3	0.27
3	2.69	0.34	0.3	0.3	0.49	0.3
4	2.02	0.26	0.27	0.26	0.33	0.24
5	1.47	0.15	0.16	0.17	0.24	0.17

Table.4. Error metrics recorded with rank-based feature selection techniques for different datasets

Data Set	Error metrics	Kappa, MAE and RMSE metrics obtained with various datasets and various attribute selection filters					
		No Attribute selection	Chi square	Gain Ratio	Info gain	Relief	Symmetrical
Data Set 1	Kappa statistic	0.346	0.4599	0.4598	0.4616	0.4222	0.4613
	Mean Absolute Error (MAE)	0.3056	0.292	0.2919	0.2919	0.2974	0.2919
	Root Mean Squared Error (RMSE)	0.3902	0.3724	0.3722	0.3722	0.379	0.3721
Data Set 2	Kappa statistic	0.1233	0.3003	0.3016	0.2946	0.0925	0.2985
	MAE	0.3763	0.3274	0.3277	0.3277	0.3822	0.3274
	RMSE	0.4764	0.4225	0.4229	0.423	0.4811	0.4225
Data Set 3	Kappa statistic	0.4711	0.5862	0.5862	0.5881	0.4407	0.5862
	MAE	0.298	0.278	0.2783	0.2778	0.2988	0.278
	RMSE	0.3871	0.3602	0.3605	0.3598	0.3864	0.3602
Data Set 4	Kappa statistic	0.3933	0.6374	0.6374	0.6374	0.4997	0.6374
	MAE	0.2707	0.2623	0.2623	0.2623	0.2681	0.2623
	RMSE	0.3435	0.3309	0.3309	0.3309	0.3394	0.3309
Data Set 5	Kappa statistic	0.4275	0.6278	0.6266	0.6266	0.5129	0.6266
	MAE	0.2676	0.2623	0.2623	0.2623	0.265	0.2623
	RMSE	0.3388	0.3306	0.3306	0.3306	0.3346	0.3306

5. EXPERIMENTAL RESULTS DISCUSSION

In order to objectively compare the performance of various models built using the feature selection techniques, we ranked each the measurements separately across each of the six datasets comparing across the five feature selection algorithms used for this research purpose. The ranking mechanism and the ranks are described below.

5.1 RANKING THE RETAINED FEATURE SETS IN EACH DATASET BY FEATURE SELECTION ALGORITHM USED

The number of features retained in each of the six datasets is compared across the five feature selection methods and the feature selection method in each of the dataset which yielded in retention of least number of features is ranked with rank 1 and next least number of features is ranked rank 2 etc., In cases where same number of features are retained, same rank is

assigned to both however the next rank is skipped for the next lower valued feature retention. The lowest number of retained features is ranked as 1 because the algorithm runs more efficiently due to less in-memory space requirement to store the reduced feature set. Table.5A shows the ranks based on the number of features retained post application of the various attribute selection filters.

5.2 RANKING ACCURACIES OBTAINED IN EACH DATASET

The highest accuracy across each dataset is given rank 1 and lowest accuracy is given rank 6. In situations of same accuracy, same rank is assigned to both; however the next rank is skipped for the purpose of assigning the next rank. Table 5B shows the ranks based on the accuracies obtained through 10 fold cross validation with SMO models on the reduced datasets.

5.3 RANKING BASED ON THE TIME TAKEN TO BUILD MODELS

The lowest time to build the model across each dataset is given rank 1 and lowest time to build the model is given rank 6. In situations of same time to build the model, same rank is assigned to both however the next rank is skipped for purpose of assigning the next rank. Table 5C shows the ranks based on the time taken to build models with SMO algorithm using the reduced datasets obtained from application of various attribute selection filters.

5.4 RANKING THE KAPPA STATISTIC IN EACH DATASET

The highest Kappa across each dataset is given rank 1 and lowest kappa is given rank 6. In situations of same kappa, same rank is assigned to both however the next rank is skipped for purpose of assigning the next rank. Highest valued kappa is given the rank 1 because the closer the kappa statistic is to 1 the better model it is.

5.5 RANKING THE MEAN ABSOLUTE ERROR STATISTIC IN EACH DATASET

The lowest mean absolute error across each dataset is given rank 1 and lowest mean absolute error is given rank 6. In situations of same mean absolute error, same rank is assigned to both however the next rank is skipped for purpose of assigning the next rank. Lowest valued mean absolute error is given the rank 1 because the closer the mean absolute error statistic is to 0 the better model it is.

5.6 RANKING THE ROOT MEAN SQUARED ERROR STATISTIC IN EACH DATASET

The lowest root mean squared error across each dataset is given rank 1 and lowest root mean squared error is given rank 6. In situations of same root mean squared error, same rank is assigned to both however the next rank is skipped for purpose of assigning the next rank. Lowest valued RMSE is given the rank 1 because the closer is the RMSE statistic to 0 the better is the corresponding model.

Table.5. Rankings based on different parameters

A) Rankings based on retained reduced feature sets						
Data Set	**No attribute selection**	**Chi Square**	**Gain Ratio**	**Info Gain**	**Relief**	**Symmetrical**
1	6	1	1	1	5	1
2	6	1	1	1	5	1
3	6	1	1	1	5	1
4	6	1	1	1	5	1
5	6	1	1	1	5	1
B) Rankings based on accuracies obtained						
Data Set	**No attribute selection**	**Chi Square**	**Gain Ratio**	**Info Gain**	**Relief**	**Symmetrical**
1	6	1	1	1	5	1
2	5	1	1	1	6	1
3	5	1	1	1	6	1
4	6	1	1	1	5	1
5	6	1	1	1	5	1
C) Rankings based on time taken to build the models						
Data Set	**No attribute selection**	**Chi Square**	**Gain Ratio**	**Info Gain**	**Relief**	**Symmetrical**
1	6	3	1	2	5	4
2	6	1	2	3	5	4
3	6	4	1	1	5	1
4	6	2	4	2	5	1
5	6	1	2	3	5	3

Table.6. Rankings based on Kappa, Mean Absolute Error and Root mean squared error

Data Set	Error metrics	Attribute Selection Filters					
		No attribute selection	Chi Square	Gain Ratio	Info Gain	Relief	Symmetrical
1	Kappa statistic	6	3	4	1	5	2
	MAE	6	4	1	1	5	1
	RMSE	6	4	2	2	5	1
2	Kappa statistic	5	2	1	4	6	3
	MAE	5	1	3	3	6	1
	RMSE	5	1	3	4	6	1
3	Kappa statistic	5	2	2	1	6	2
	MAE	5	2	4	1	6	2
	RMSE	6	2	4	1	5	2
4	Kappa statistic	6	1	1	1	5	1
	MAE	6	1	1	1	5	1
	RMSE	6	1	1	1	5	1
5	Kappa statistic	6	1	2	2	5	2
	MAE	6	1	1	1	5	1
	RMSE	6	1	1	1	5	1

5.7 OBTAINING FINAL RANKING

With various ranks given to the filters based on various parameters, it is difficult to arrive at conclusions. Therefore, there is a strong need to come up with a single factor based on which conclusions can be derived about the filters. For this purpose, we computed final rankings from the various individual ranks by summating the ranks obtained across all data sets and all metrics but by Attribute selection filter viz., the summation was computed across No attribute selection, CS, GR, IG, Relief and SU Attribute selection filters. The lowest aggregate among the obtained aggregates is ranked 1 and the highest aggregate is ranked with lowest possible rank of 6. The filter selection algorithm that obtained the rank 1 is the best filter selection algorithm amongst the algorithms considered for this research. Fig.1 shows the rank sums and the appropriate ranks provided to the attribute evaluation filters. The values in the brackets are the ranks obtained by each filter and the rank is awarded based on the rank sums.

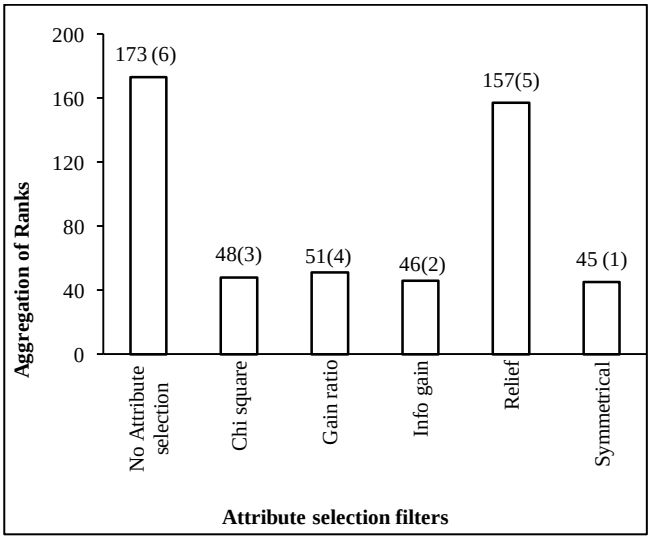

Fig.1. Rank sums and overall ranks based on rank sums (in brackets) across the various filters included in the experiment

6. CONCLUSION AND FUTURE DIRECTIONS

From the analysis of results, it is evident that the application of rank based filter selection algorithms on the data sets and then building models with the resultant reduced data sets yields faster and more accurate models than models built with no feature selection.

It is also observed that all algorithms except Relief filter based algorithm performed the same way when judged by accuracy alone with reduced number of features/ therefore all algorithms except Relief filter were ranked the same. However, slight differences were recoded in the time taken to build the models, Kappa, Mean Absolute Error and Root Mean Squared Error. These differences yielded significantly to the overall ranking of feature selection filters for comparing their overall performance.

Based on our results, we conclude the Symmetric Uncertainty attribute evaluation as the best method with an overall rank aggregation metric of 45. The rank aggregation metric of Symmetric Uncertainty attribute evaluation filter is better by 1 rank than the rank aggregation metric of Information Gain attribute evaluation filter. Symmetric Uncertainty filter's rank aggregation metric is better by 3, 6, 112 ranks respectively when compared to rank aggregation metric of Chi Squared filter, Gain Ratio filter and Relief filters. From these measurements, Symmetrical Uncertainty attribute evaluation algorithm is proved to be the best feature selection method. Information Gain attribute evaluation is the close second algorithm to Symmetrical Uncertainty algorithm.

While in this paper we were able to apply various rank based feature selection filters to data sets to identify the best filter, applying wrappers to filter attributes is another perspective to explore. Further research is required to apply dimensionality reduction techniques such as Principal Component Analysis and perform feature transformation to verify if the model's performance can be improved. For the purpose of our research,

we used SMO polynomial kernel. However as next step we want to verify if the models yield any different results if different kernels are used with SVM during training.

REFERENCES

[1] J. Brank, M. Grobelnik, N. Milic-Frayling and D. Mladenic, "Interaction of Feature Selection Methods and Linear Classification Models", *Proceedings of the 19th International Conference on Machine Learning – Workshop on Text Learning*, 2002.

[2] X. Han, G. Zu, W. Ohyama, T. Wakabayashi and F. Kimura, "Accuracy Improvement of Automatic Text Classification Based on Feature Transformation and Multi-classifier Combination", *Content Computing Lecture Notes in Computer Science*, Vol. 3309, pp. 463-468, 2004.

[3] Lindsay I Smith, "*A tutorial on Principal Components Analysis*", pp. 1-13, 2002.

[4] G. Zu, W. Ohyama, T. Wakabayashi and F. Kimura, "Accuracy improvement of automatic text classification based on feature transformation", *Proceedings of the 2003 ACM Symposium on Document Engineering*, pp. 118-120, 2003.

[5] A. L. Blum and P. Langley, "Selection of relevant features and examples in machine learning", *Artificial Intelligence*, Vol. 97, No. 1-2, pp. 245-271, 1997.

[6] M. Dash and H. Liu, "Feature selection for classification", *Intelligent Data Analysis*, Vol. 1, No. 1-4, pp. 131-156, 1997.

[7] K. Thankavel and A. Pethalakshmi, "Dimensionality reduction based on rough set theory: a review", *Applied Soft Computing*, Vol. 9, No. 1, pp. 1-12, 2009.

[8] Anirban Dasgupta, Petros Drineas, Boulos Harb, Vanja Josifovski and Michael W. Mahoney, "Feature selection methods for text classification", *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data mining*, pp. 230-239, 2007.

[9] Monica Rogati and Yiming Yang, "High-performing feature selection for text classification", *Proceedings of the Eleventh International Conference on Information and Knowledge Management*, pp. 659-661, 2002.

[10] Y. Yang and J. O. Pedersen, "A Comparative Study on Feature Selection in Text Categorization", *Proceedings of the Fourteenth International Conference on Machine Learning*, pp. 412-420, 1997.

[11] George Forman, "An extensive empirical study of feature selection metrics for text classification", *Journal of Machine Learning Research*, Vol. 3, pp. 1289-1305, 2003.

[12] Mark A. Hall and Geoffrey Holmes, "Benchmarking Attribute Selection Techniques for Discrete Class Data Mining", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 15, No. 3, pp. 1-16, 2003.

[13] Sunil Kumar and R. J. Rama Sree, "Assessment of Performances of Various Machine Learning Algorithms During Automated Evaluation of Descriptive Answers", *ICTACT Journal on Soft Computing -Special Issue on Soft Computing in System Analysis, Decision and Control*, Vol. 4, No. 4, pp 781-786, 2014.

[14] Sunil Kumar and R. J. Rama Sree, "Experiments towards determining best training sample size for automated evaluation of descriptive answers through sequential minimal optimization", *ICTACT Journal on Soft Computing*, Vol. 4, No. 2, pp. 710-714, 2014.

[15] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann and Ian H. Witten, "The WEKA Data Mining Software: An Update", *SIGKDD Explorations*, Vol. 11, No. 1, pp. 10-18, 2009.

[16] Ian H. Witten and Eibe Frank, "*Data Mining: Practical Machine Learning Tools and Techniques*", Second Edition, Morgan Kaufmann Publisher, 2005.

[17] Vishal Gupta and Gurpreet Singh Lehal, "A Survey of Common Stemming Techniques and Existing Stemmers for Indian Languages", *Journal of Emerging Technologies in Web Intelligence*, Vol. 5, No. 2, pp. 157-161, 2013.

[18] N. Abe and M. Kudo, "Entropy criterion for classifier-independent feature selection", *Lecture Notes in Computer Science*, Vol. 3684, pp. 689-695, 2005.

[19] H. Liu and R. Setiono, "Chi2: Feature selection and discretization of numeric attributes", *Proceedings of the IEEE 7th International Conference on Tools with Artificial Intelligence*, pp. 338-391, 1995.

[20] Jasmina Novakovic, "Using Information Gain Attribute Evaluation to Classify Sonar Targets", *17th Telecommunications forum TELFOR*, pp. 1351-1354, 2009.

[21] M. A. Hall and L. A. Smith, "Practical feature subset selection for machine learning", *Proceedings of the 21st Australian Computer Science Conference*, pp. 181-191, 1998.

[22] C. Gayathri, "Feature subset selection using filtering with Mutual information and Maximal information coefficient", *International Journal of Innovative Research in Computer and Communication Engineering*, Vol. 2, No. 1, pp. 1350-1354, 2014.

[23] R. S. Marko and K. Igor, "Theoretical and empirical analysis of relief and rreliefF", *Machine Learning Journal*, Vol. 53, No. 1-2, pp. 23-69, 2003.

[24] John C. Platt, "Fast Training of Support Vector Machines using Sequential Minimal Optimization", *Advances in Kernel Methods - Support Vector Learning*, pp. 185-208, 1998.

[25] Ian H. Witten and Eibe Frank, "*Data Mining: Practical Machine Learning Tools and Techniques - Chapter 5, Credibility: Evaluating What's Been Learned*", Second Edition, Morgan Kaufmann Publisher, 2005.

[26] The Hewlett Foundation, "Automated Essay Scoring", Available at: http://www.kaggle.com/c/asap-aes, 2012, accessed: 11-07-2014.

[27] The Hewlett Foundation, "Code for evaluation metric and benchmarks", Available at: https:// www.kaggle.com /c/ asap-aes/ data? Training_Materials.zip, 2012, accessed: 11-07-2014.

[28] Wikipedia, "n-gram", Available at: http://en.wikipedia.org/wiki/N-gram, accessed: 11-07-2014.

[29] Wikipedia, "Entropy (information theory)", Available at: http://en.wikipedia.org/wiki/Entropy_(information_theory), accessed: 11-07-2014.

[30] Weka Sourceforge, "Class ChiSquaredAttributeEval", Available at: http://weka.sourceforge.net/doc.stable

/weka/attributeSelection/ChiSquaredAttributeEval.html, accessed: 11-07-2014.

[31] Weka Sourceforge, "ClassInfoGainAttributeEval", Available at: http://weka.sourceforge.net/doc.dev/weka/attributeSelection/InfoGainAttributeEval.html, accessed: 11-07-2014.

[32] Weka Sourceforge, "Class GainRatioAttributeEval", Available at: http://weka.sourceforge.net/doc.dev/weka/attributeSelection/GainRatioAttributeEval.html, accessed: 11-07-2014.

[33] Weka Sourceforge, "Class SymmetricalUncertAttributeEval", Available at: http://weka.sourceforge.net/doc.dev/weka/attributeSelection/SymmetricalUncertAttributeEval.html, accessed: 11-07-2014.

[34] Weka Sourceforge, "Class ReliefFAttributeEval", Available at: http://weka.sourceforge.net/doc.dev/weka/attributeSelection/ReliefFAttributeEval.html, accessed: 11-07-2014.

[35] Kaggle, "Root Mean Squared Error (RMSE)", Available at: https://www.kaggle.com/wiki/RootMeanSquaredError, accessed: 11-07-2014.

[36] Wikipedia, "Cohen's Kappa", Available at: http://en.wikipedia.org/wiki/Cohen%27s_kappa, accessed: 11-07-2014.

[37] Kaggle, "Mean Absolute Error", Available at: https://www.kaggle.com/wiki/MeanAbsoluteError, accessed: 11-07-2014.