
# A Brief Report on Filter Based Novel Feature Selection Schemes

*By :*

*Devashish Deshpande (2013A7PS122G)*

*Bhargav Srinivasa (2013A7PS053G)*

*Gauri Kholkar (2013A7PS002G)*

*Mrunmayee Nasery (2013A7PS087G)*

*Abhirav Gholba (2013A7PS027G)*

Submitted in partial fulfilment towards

Information Retrieval (CS F469) Course Project.

Under the guidance of :

*Mr. Rajendra Kumar Roul*

Date :

*April 28, 2016.*

***Abstract -***

In text categorisation, we consider the contribution of each word of the text towards the categorisation of each document. Thus each word becomes a feature of the document. Since a document may have a large number of features, feature selection techniques are important for text categorisation, such that the accuracy of a classifier is not reduced and the process becomes feasible to compute. In this brief study, we have implemented, tested and compared the performance of four feature selection techniques namely, Comprehensive Measure Feature Selection (CMFS), Improved CMFS, Improved Global Feature Selection Scheme (IGFSS) using Correlation Coefficient (CC) as the local one-sided feature selection scheme. Once we have used CMFS as the global feature selection method in IGFSS and other time we have used ICMFS. The performance of the above four is compared with Chi Square (CHI) for 20NG dataset and the classifier used is Naive Bayes (NB) classifier.

## Comprehensive Measure Feature Selection (CMFS): [1]

- LOGIC :

  It comprehensively measures the significance of a term both in inter-category and intra-category. The number of categories is predefined in text classification, so we regard this vector space model as a term-to-category feature-appearance matrix where rows are the features and columns are category vector. The elements in the term-to-category matrix are the number of documents in which a feature occurs in every category. (DIA and DF)

- ALGORITHM :

  The terms used are

  **$tf(t_k, c_i)$** is the frequency of a term $t_k$ in category $c_i$ ;

  **$tf(t_k)$** is the frequency of a term $t_k$ in the entire training set ;

  **$tf(t, c_i)$** is the sum of frequency of all terms in category $c_i$ ;

  **|C|** is the number of the categories ;

  **|V|** is the total number of terms in the feature vector space.

$$CMFS(t_k, c_i) = \frac{(tf(t_k, c_i)+1)^2}{(tf(t_k)+|C|)(tf(t, c_i)+|V|)} = P(c_i \mid t_k) P(t_k \mid c_i)$$

To measure globally the goodness of a term, two alternate ways can be used to combine the category specific scores of a term. We have considered the maximum score over averaging.

$$CMFS_{max}(t_k) = \max_{i=1}^{|C|} \{CMFS(t_k, c_i)\}$$

ALGORITHM 1 :

Input: V – the term-to-category feature-appearance matrix where rows are the features and columns are categories

k – the number of the selected features

Output: $V_s$ – the feature subset

Step 1: obtains the number of categories (the number of columns in V) – |C|

Step 2: obtains the size of the feature vector space (the number of rows in V) – |V|

Step 3: for each column (each category) $c_i$

Step 4: obtains the sum of frequency of all features in category $c_i$ – tf (t, $c_i$ )

Step 5: end for

Step 6: for each row (each feature) $t_k \in V$

Step 7: obtains the frequency of the feature t k in the entire training set – tf ($t_k$)

Step 8: for each column (each category) $c_i$

Step 9: obtains the frequency of the feature $t_k$ in category $c_i$ – tf ($t_k$, $c_i$)

Step 10: calculates the significance of the feature $t_k$ against the category $c_i$ – CMFS ($t_k$ , $c_i$)

Step 11: obtains the maximum value of CMFS ($t_k$ , $c_i$) – CMFS ($t_k$)

Step 12: end for

Step 13: end for

Step 14: ranks all features in V based on CMFS ($t_k$)

Step 15: selects top k features into $V_s$

## Improved Comprehensive Measure Feature Selection (ICMFS): [2]

- DRAWBACKS OF CMFS:

  (1) *P (c | t)* doesn't consider the effect of category size on feature selection results, thus the terms which occur in the categories with small sizes are always ignored. This has been taken care in ICFMS.

  *P ($c_i$)* can reduce the chance that a term which yields small *P ($c_i$ | t)* value in a category with small size is ignored, improving the accuracy of feature selection.

$$ICMFS(t_k, c_i) = \frac{P(c_i \mid t_k)\ P(t_k \mid c_i)}{P(c_i)}$$

  (2) *P (t | c)* term doesn't consider the effect of a terms contribution towards other categories. Features which which have high frequncy accross all the categories tend to get selected.

We tried to normalise this by introducing the average frequency of a term in the ICMFS equation, but the results we got were worse compared to CHI. So we abandoned this idea.

- ALGORITHM 2

  By taking maximum, ICFMS becomes a global feature selection technique. Exactly like ***CMFS***.

## Correlation coefficient (CC) : [3]

Correlation coefficient measures the lack of independence between a term t and a category c :

$$CC(t_k, c_i) = \frac{\sqrt{N}\,[P(t, c_i)\,P(\bar{t}, \bar{c}_i) - P(t, \bar{c}_i)\,P(\bar{t}, c_i)]}{\sqrt{P(t)\,P(\bar{t})\,P(c_i)\,P(\bar{c}_i)}}$$

It's a variant of Chi-Square (CHI) and can be viewed as a one-sided CHI metric.

## Improved Global Feature Selection Scheme (IGFSS) [4]

Consequently, IGFSS aims to improve the classification performance of global feature selection methods by creating features representing all classes almost equally. For this purpose, a local feature selection method is used in IGFSS to label features according to the indiscriminative power on classes and these labels are used while producing the feature sets.

Another categorization about characteristics of filter-based feature selection methods is whether they are one-sided or two-sided. In one-sided metrics, while features indicating membership to classes have a score greater than or equal to 0, features indicating non-membership to classes have a score smaller than 0. As features are ranked in descending order and the features having highest scores are included in the feature set, the negative features are not used in case there is no candidate positive feature. However, scores of two-sided methods are greater than or equal to 0. They implicitly combine positive and negative features which indicate the membership and non-membership to any class, respectively. In this case, considering one-against-all strategy in feature selection, positive features attain higher scores than negative ones. Thus, the negative features are rarely added to the feature set in two-sided metrics.

The main idea behind this study is to consider the imbalance factor of the training sets in the globalization process of class-based feature selection scores. It is reported that this improved

scheme can significantly improve the performance of feature selection methods.

Equal number of features representing each class equally with a certain membership and non-membership degree were included in the final feature set. In the experiments, an empirically determined negative feature ratio was used to represent each class with nearly same number of negative features.

For this purpose, a one-sided local feature selection method (CC) is integrated to the feature selection process besides an existing global feature selection method.

- ALGORITHM 3

  Stage1.(Feature labeling)-

  Calculate one-sided local feature selection scores of features for each class. Create a label set *l* for features including $m*2$ class labels where *m* is the number of classes. While the first *m* class labels represent membership, the second *m* labels represent non-membership to these classes. For each feature, determine the highest local feature selection score regarding their absolute values and assign the associated class label from the label set *l* to the feature.

  Stage2. (Common global feature selection process)-

  Calculate feature selection scores for features using one of the global feature selection metrics. Sort the features in descending order according to the scores and the sorted list is named as *sl*.

  Stage3.(Construction of the new feature set)-

  Suppose that the size of the final feature set was given as *fs* and a set of negative feature ratios was determined as *nfrs*. The values in *nfrs* may change from 0 to 1 with a specified pre-determined interval such as 0.1.

  Iterate over the sorted list *sl* obtained in the previous stage and put the appropriate features in the final feature set *ffs*. Make the *ffs* equally representative for each class by using the feature labels determined in stage 1. At the end of this stage, *ffs* must contain equal number of features for each class considering a specific negative membership ratio value *nfr* inside *nfrs*.

  Stage 4. (Conditional part) -

  If the number of features in *ffs* is less than *fs*, finalize the feature selection process by adding missing amount of disregarded features having highest global feature

selection scores to *ffs*.

- WHY ONE-SIDED FEATURE SELECTION SCHEME IN IGFSS? [3]

  The impact of imbalanced data problem on the standard feature selection can be illustrated as follows, which primarily answers the question, "How sub-optimal are two-sided metrics?" :

  First, for the methods using one-sided metrics (e.g. SIG, CC, and OR), the non-relevant documents are subject to misclassification. It will be even worse for the imbalanced data problem, where non-relevant documents dominate. How to confidently reject the non-relevant documents is important in that case.

  Second, given a two-sided metric, the values of positive features are not necessarily comparable with those of negative features. Let us use CHI for example. The upper limit CHI value of a positive or negative feature is N. For the positive feature, it represents the case that the feature appears in every relevant document, but never in any non-relevant document. For the negative features, it means that the feature appears in every non-relevant document, but never in any relevant document. Due to the large amount and diversity of the non-relevant documents in imbalanced data set, it is much more difficult for a negative feature to reach the same maximum that a positive feature does. This extreme example shed light on why the CHI values of positive features are usually much larger than those of negative features. CHI and CC are very similar when the size of the feature set is small and the data set is highly imbalanced.

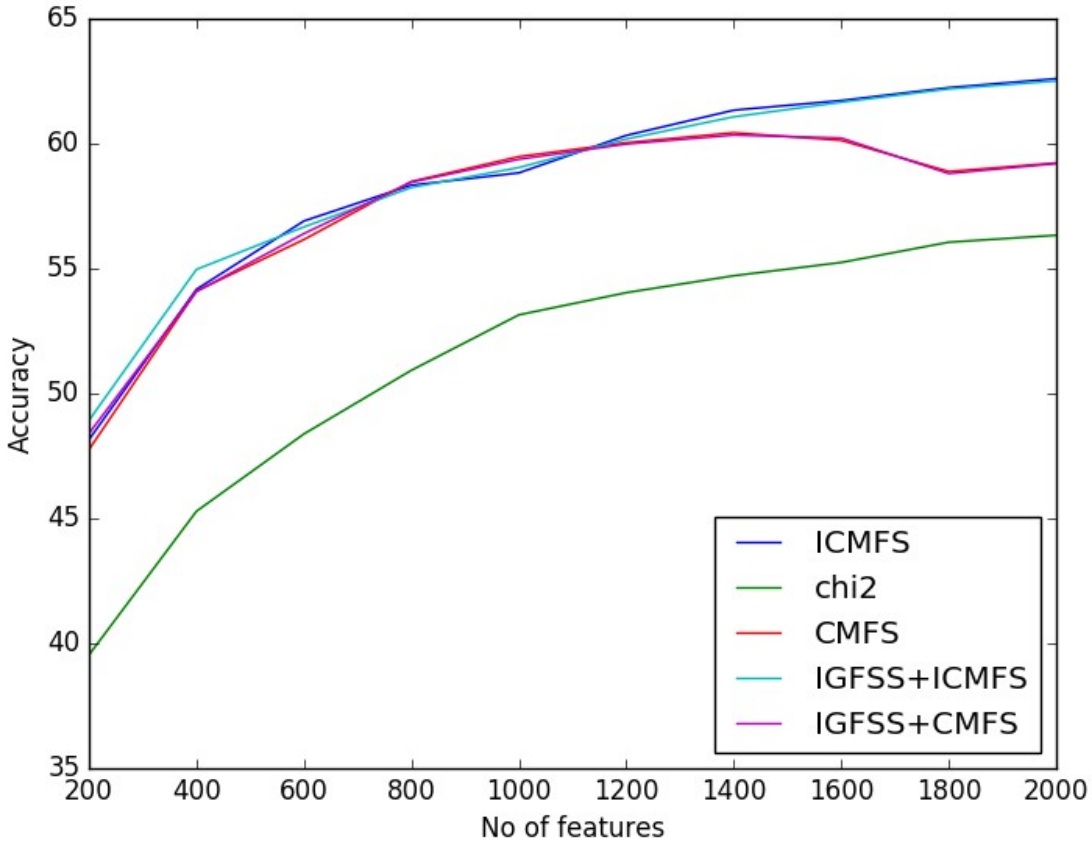
## Observations and Conclusions:

➔ CMFS gives much better performance as compared to CHI for all sizes of the feature set.

➔ ICMFS gives considerably better results as compared to CMFS at feature set sizes > 1400.

➔ IGFSS using ICMFS performs comparable to IGFSS using CMFS and beats it at lower values of feature set size.

➔ By tweaking the *nfr* values, greater accuracy can be obtained. The value of *nfr* should depend on the performance measure under consideration. [3]




Dataset – 20 NewsGroup

| | 200 | 400 | 1000 | 1600 | 2000 |
|---|---|---|---|---|---|
| CHI | 0.3953 | 0.4528 | 0.5314 | 0.5314 | 0.5633 |
| CMFS | 0.4775 | 0.5410 | 0.5947 | 0.6013 | 0.5922 |
| ICMFS | 0.4844 | 0.5407 | 0.5937 | 0.6022 | 0.5920 |
| IGFSS (CMFS + CC) | 0.4815 | 0.5416 | 0.5882 | 0.6172 | 0.6259 |
| IGFSS (ICMFS + CC) | 0.4880 | 0.5496 | 0.5904 | 0.6165 | 0.6250 |

Table 1 – relative performance on NB classifier measured on F1 (micro)

## Future Work :

We would like to test CMFS, ICMFS, IGFSS (CMFS + CC), IFGSS (ICMFS + CC) with other datasets like Classic3 and Reuters and other classifiers like SVM etc. Also optimize accuracy of IGFSS by empirically determining the best values for the *nfrs.* We would also like to improve upon the ICMFS selction techniques and by trying to rectify the second drawback.

## References :


[1] *A new feature selection based on comprehensive measurement both in inter-category and intra-category for text categorization [Yang, Y Liu, Zhu, Z Liu, Zhang] (2012.)*

[2] *Improved Comprehensive Measurement Feature Selection Method for Text Categorization [Feng, Zuo, Wang] (2015 International Conference on Network and Information Systems for Computers.)*

[3] *Feature Selection for Text Categorization on Imbalanced Data [Zheng, Wu, Srihari] (Sigkdd Explorations.)*

[4] *An improved global feature selection scheme for text classification [Alper Kursat Uysal] (Elsevier, 2015.)*