The Nucleus **51**, No. 2 (2014) 177-185

The Nucleus

A Quarterly Scientific Journal of Pakistan
Atomic Energy Commission

NCLEAM, ISSN 0029-5698

CHALLENGES IN DETERMINING TERM RELEVANCE FOR TEXT DATA

*A. REHMAN, K. JAVED[1] and H. A. BABRI[1]

Department of Computer Science and Engineering, University of Engineering and Technology, Lahore, Pakistan

[1]Department of Electrical Engineering, University of Engineering and Technology, Lahore, Pakistan

(Received May 13, 2014 and accepted in revised form May 26, 2014)

Although the nature of text data is different from ordinary non-text datasets in a number of ways, existing algorithms from Machine Learning domain have been borrowed for the classification of text data. Machine learning algorithms cannot be readily applied on raw text data. Text data needs to be transformed to a suitable form for the application of machine learning algorithms. The transformation produces further problems for feature selection and classification algorithms. In this paper we highlight the problems introduced by transformation of text data. We also show how different feature selection algorithms including bi-normal separation, information gain and ROC are affected by text data.

Keywords: Text classification, Term frequency, Feature selection, Document length, True positive rate, Inverse document frequency

1. Introduction

Text data is different from ordinary non-text data sets in a number of ways. Yet feature selection and classification of text data has been considered as variants of ordinary machine learning problems. Therefore, existing algorithms from machine learning domain have been applied for feature selection and classification of text data. For the application of machine learning technique, text data is represented in the form of a matrix. A collection of documents, also called corpus, is represented by an m n matrix, where m is total number of documents and n is total number of unique terms in the corpus. A value in m^{th} row and n^{th} column of the matrix shows term count of n^{th} term in m^{th} document. This representation is called "bag of words" representation.

In addition to the problems inherent to the text data, "bag of words" representation causes a few more problems. A document in a text corpus contains a smaller fraction of total number of words in the whole corpus. To represent the data in matrix form, each document is augmented with zero values for the absent terms. So, the resulting matrix becomes very high dimensional and highly sparse. High dimensionality of text data makes it practically impossible for the application of classification algorithms without applying feature selection on text data.

In non-text datasets, relevance of a term to a class is measured by its values in different classes. A feature having similar values in positive and negative classes is considered irrelevant for classification and is removed. In text datasets, term counts are the feature values. Instead of using term count of a term in different classes, document frequencies of the term are used to measure its relevance to a class. Frequent terms in different classes are considered irrelevant for classification task. These include stop words like "for", "the", "is" etc., and frequently occurring nouns for the underlying domain [1].

After removing stop words, rarely occurring and frequently occurring terms, feature set may still contain a number of irrelevant terms. Detection and removal of such terms is done using some suitable feature selection algorithm. Again, feature ranking metrics use document frequency of a term in a class to measure the relevance of a term to a class. More frequent is a term in a class, higher will be its relevance.

High class skew is another problem of text data. Text data is multi-labeled, multi-class high dimensional data where a document can belong to one or more than one classes out of M classes [2]. Conventionally, multi-class classification problem is decomposed into M binary class problems. In each binary problem, one class is considered positive class while all other classes are merged to form negative class. This is called one-against-all settings. The resulting data sets using one-against-all settings are skewed even if the original data set is balanced.

True representation of terms in a document is yet another problem. Commonly used representation of terms in a corpus is *tfidf* where *tf* is the normalized term count and *idf* is inverse document frequency. It has been indicated by Xiao-Bing and Zhi-Hua [3] that *tf* does not always truly represent term importance in a document. We investigated that new terms are introduced with increase in document length, which slows down increase in term counts of existing terms. It causes the term frequencies of longer documents to decrease as

[*] Corresponding author : a.rehman@kics.edu.pk

compared to shorter documents.

Rest of the paper is arranged as follows. Section 2 lists different challenges of text data and their brief explanation. Section 3 discusses behavior of two features ranking metrics on an artificial sample dataset. Section 4 presents conclusions.

2. Issues with Text Data

Following sub-section describe number of challenges posed by text data for machine learning algorithms.

2.1 Variable Document Length

In a text corpus document length often varies widely [4]. Shorter documents normally contain topic specific information and term counts are very small. Longer documents may contain words from other domain. Term counts for relevant terms are high while terms from other domains may appear with low term counts. Longer documents may show higher relevance to a user query due to higher term counts than shorter documents [5]. Term counts are therefore normalized to transform term counts to a uniform scale. Non-text data sets have a fixed number of features for all instances where a feature is measured on the same scale for all instances.

2.2. Frequent Terms are Less Important

Term frequencies in text documents follow Zipf distribution [6]. According to Zipf distribution, term frequency of a term is inversely proportional to its rank among all terms ranked by term frequencies, i.e., df(w) = 1=r(w)p, where df(w) is the document frequency of a word, r(w) is word rank and p is close to 1 [7]. It shows that frequent terms are ranked lower in a text corpus. A small number of words, for example, stop words such as \the", \at", \this", \and", occur very commonly and a large number of words, used in a special context, occur only in a few documents. Too frequent and too rare terms need to be removed to keep informative terms only. Removal of frequent terms may also cause some important terms to be removed.

2.3 Documents with Different Number of Features Lie in the Same Corpus

Text documents in a corpus differ widely with respect to the set of words they contain. Documents belonging to even the same category do not contain the same set of words. Word sharing is reduced in documents be-longing to different categories, which is not the case for non-text datasets where all instances have the same set of features.

2.4 Feature Augmentation

Text documents in a corpus do not contain similar set of term. Machine learning algorithms require equal number of features for all instances in a dataset. To fulfill this requirement, each document is augmented with 0s for absent terms, which gives rise to sparseness. Sparseness goes on increasing as more and more documents are added. Table 2 shows documents from R8 datasets. It can be seen that absent features are filled by 0 values.

2.5 Imbalanced Datasets Due to One Against All Settings

In multi class set-tings, the documents in a corpus belong to a number of different classes. Commonly used classifiers like SVM operate on binary classes which contain only two classes. To use binary classifiers, dataset is divided into multiple two class datasets where one class is considered positive class while remaining are merged to form negative class. It is called one-against-all settings.

One-against-all settings produces highly skewed datasets where positive class becomes minor class in majority of the cases. If the original dataset is balanced, all classes contain nearly equal number of instances and the resulting subsets will be equally skewed. If the original dataset is not balanced, resulting datasets will have different skews. The classifier can learn to generate high accuracy by simply assigning all the instances to the negative class.

2.6. Feature Presence Considered Instead of Feature Values

For non-text datasets, feature ranking metrics like information gain, mutual information and chi-square ranks features on the basis of feature values. Feature values for text data are *tfidf*s, which are real in nature. For feature ranking of text data, these metrics have different version than for non-text data, which operate on document frequencies of the terms instead of *tfidf*. Table 1 shows version of different feature ranking metrics for text and non-text datasets. It can be seen that these metrics completely ignore term frequencies while determining relevance of a term to a class.

2.7 Positive and Negative Features are Treated Equally

In a two class problem, a feature can have stronger association to one of the classes. Features having stronger association to positive class are called positive features, while those having stronger association to negative class are called negative features. Some feature ranking metrics consider absolute value of *tpr-fpr,* treat ranks of both positive and negative features equally without using any weighting factors for different type of features.

Table 1. Versions of different feature ranking metrics for text and non-text datasets.

Metric	Non-text data	Text data
Information Gain	$\sum_{n=1}^{n} P(H_i) log_2 P(H_i)$	$e(pos, neg)[P_w e(tp, fp) + \bar{P}_w e(tp, fp)]$
Chi Square	$\sum \frac{(observed\ value - expected\ value)}{expected\ value}$	$\frac{N[P(t, c_i)P(\bar{t}, \bar{c}_i)P(t, \bar{c}_i)P(\bar{t}, c_i)]^2}{P(t)P(\bar{t})P(c_i)P(\bar{c}_i)}$

Table 2. Term counts and term frequencies for two documents from R8 dataset. It shows how documents are augmented with 0 values for absent features.

category	acq	earn	acq	earn
term	term count	term count	term frequency	term frequency
div	1	0	0.1	0
divided	1	0	0.1	0
march	2	0	0.2	0
pai	1	0	0.1	0
prior	1	0	0.1	0
qtly	2	0	0.2	0
rai	1	0	0.1	0
record	1	0	0.1	0
subsidiary	0	2	0	0.2
annual	0	1	0	0.1
dai	0	1	0	0.1
manage	0	1	0	0.1
oper	0	1	0	0.1
product	0	1	0	0.1
revenue	0	1	0	0.1
set	0	1	0	0.1
unit	0	1	0	0.1

2.8 Impact of Larger Class

Document frequencies of a term in the negative class can be affected by the larger class in unbalanced data sets. If the term is associated with the larger class, it can gain a high document frequency in the negative class. Those features ranking metrics which consider document frequency as a measure of importance of a term to a class, terms in larger class will get a higher rank. If the term is associated to a smaller class, presence of larger class in the negative class further decreases ratio of documents containing the term to total number of documents in the negative class, which is the false positive rate f pr. Thus the term rank and chances of its selection in the final set of features will be low.

2.9 Documents May Get Deprived of Features

As mentioned by Forman in [11], after feature selection all features may be removed from a number of documents. Such documents cannot be used for training classifiers. It happens due to imbalanced datasets and different document lengths. Due to larger size, a major part of features is selected from larger class which may not be present in other classes. So the documents from smaller classes may be deprived of all features as a result of feature selection. It does not happen in non-text datasets where each instance contains complete set of features.

2.10 All features Range from 0 to Maximum

In text data, range of values for all features starts from 0 and reaches a maximum depending upon the document lengths and term relevance to the class. It makes the feature selection and classification tasks difficult. Document frequency in therefore combined with term frequencies to get different range of values in different classes, which is still affected by 0 values.

2.11 Analysis of Some Feature Ranking Metrics

In this section we show how different feature ranking metrics, document frequency DF, information gain IG, bi normal separation BNS and odds ratio OR, are affected by the issues discussed in section 2. We discuss their behavior using a sample dataset given in 3. The dataset contains 11 term and 3 classes.

Probability of occurrence of a term in a class shows its association to the class. Table 4 lists document frequencies of the terms in the three classes of sample dataset. Table 5 lists probabilities of occurrence of terms in different classes. The main dataset is converted to three binary data sets by applying one against all settings. Associations of terms to classes as shown in table 5 are narrated below:

1. t1 and t2 have equal document frequencies but different term counts in different classes
2. t3 is strongly associated to the larger class c3
3. t4 is strongly associated to smaller classes c1 and c2
4. t5 is strongly associated to smallest class c1
5. t6 is strongly associated to smaller class c2

6. t7 is strongly associated to only one class which is the largest class c3
7. t8 is weakly associated to only one class which is the largest class c3
8. t9 is weakly associated to only one class which is a smallest class c1
9. t10 is strongly associated to only one class which is the smallest class c1
10. t11 is strongly associated to smallest and largest classes c1 and c3

Table 3. A sample unbalanced dataset showing containing 26 documents from three classes. Class c3 is the major class containing 16 documents.

s no	class	t1	t2	t3	t4	t5	t6	t7	t8	t9	t10	t11
1	c1	2	4	0	2	2	0	0	0	1	1	2
2	c1	1	0	1	2	2	0	0	0	0	1	1
3	c1	0	3	0	4	4	1	0	0	0	0	1
4	c1	1	2	0	0	0	0	0	0	0	0	0
5	c2	1	10	0	2	0	0	0	0	0	0	0
6	c2	1	6	1	0	0	1	0	0	0	0	0
7	c2	4	0	0	2	0	2	0	0	0	0	0
8	c2	0	4	0	3	1	1	0	0	0	0	1
9	c2	1	0	0	1	0	2	0	0	0	0	0
10	c2	0	7	1	1	0	3	0	0	0	0	0
11	c3	4	1	0	0	1	1	0	0	0	0	0
12	c3	2	1	1	0	0	0	0	0	0	0	1
13	c3	6	1	0	0	0	0	1	0	0	0	0
14	c3	0	0	0	0	1	0	0	0	0	0	0
15	c3	0	1	4	0	0	0	4	1	0	0	4
16	c3	0	0	3	0	0	0	2	0	0	0	3
17	c3	0	0	8	0	0	0	0	0	0	0	8
18	c3	0	0	3	0	0	0	4	1	0	0	3
19	c3	7	1	5	1	0	1	0	1	0	0	5
20	c3	3	0	4	0	0	0	1	0	0	0	4
21	c3	5	0	2	0	0	0	2	0	0	0	2
22	c3	6	1	2	0	0	0	3	0	0	0	2
23	c3	0	1	5	0	0	0	0	0	0	0	5
24	c3	4	2	1	0	0	0	0	0	0	0	1
25	c3	0	0	1	0	0	0	0	0	0	0	1
26	c3	0	0	2	0	0	0	5	0	0	0	2

Probabilities of occurrences of a term in positive and negative classes are true positive rate *tpr* and false positive rate *fpr* respectively. Table 6 shows *tpr* and *fpr* of the terms in three binary subsets of sample dataset formed as a result of one-against-all settings. We choose accuracy ACC [12] as a raw estimate of relevance of a term to positive or negative class. Definition of accuracy as given by [12] is:

$$accuracy = |tpr - fpr| \quad (1)$$

As the terms t1 and t2 have same document frequencies in different classes, both have equal ACC values for each subset. There is no major difference of ACC for one subset from the others. It shows that these terms are not associated to any particular class. The term t3 is associated to class ACC, there is a major difference between ACC for c1 than ACC for the other two classes. Although the term t4 is strongly associated with smaller classes, its ACC values for all the three classes are high. ACC for the classes c1 and c2 becomes high due to presence of larger class in the negative class. As the larger class has very low probability for the term t4 is very low, overall probability of the term t4 in the negative class remains low. For class c3, ACC is high as negative class has very high probability of occurrence of the term t4.

ACC for the term t5 is high for the class c1 as the term is strongly associated to the class c1. Lower probability of the term in class c2 and presence of larger class in the negative class keeps the ACC for the class c2 low. ACC for the class c3 is higher relative to ACC for the class c2. It is due to presence of both the smaller classes in the negative class where probability in one class is very high.

The term t6 is strongly associated to the class c2. ACC for the class c2 is high. ACC for the class c3 is reasonably higher than the ACC for the class c1. It is due to the presence of both smaller classes in the negative class where probability of the term t6 is very high in the larger class among smaller classes.

Terms t7, t8, t9 and t10 are associated to one class only. The term t7 is associated to the largest class c3 only. ACC values for the three classes are relatively closer to each other. For the cases where the largest class is present in the negative class, it dominates the overall probability as the term t7 is absent in the other class, which is a smaller class. ACC for the class c3 is highest among the three cases.

The term t8 is weakly associated to the class c3 only. It shows similar behavior as the term t7. ACC for all the three cases is almost the same, where ACC for the class c3 is the highest.

In contrast to the terms associated to the largest class only t7 and t8, terms associated to smallest only, t9 and t10, show higher ACC values for the class having stronger association to the term than the other classes. When c1 is the positive class, ACC is the same as the probability of t1 in the class c1. For the class c2, ACC is very low due to presence of largest class in the negative class which lowers overall probability of the terms in the negative class. In case of c3, probability of the terms is zero in the positive class. As the term is present only in the smallest class, the overall probability of the term lowers in the negative class.

The term t11 is strongly associated to the smallest and the largest classes. Surprisingly, ACC for the irrelevant class c2 is much higher than ACCs for the classes c1 and c3 to which the term is strongly associated. If c1 is the positive class, positive class has higher probability of occurrence of the term t11. The negative class also has higher probability of occurrence of the term t9 due to presence of larger class in negative class, which has high probability of the term t11. ACC for the class c2 is also high as both classes in negative class have high probability of occurrence of the term t11. ACC for the class c3 is higher than ACC for the class c1 and c3. Probability of the term t11 in the negative class is not too low as the classes in the negative class are smaller in size and the term t11 is strongly associated to one of the smaller terms.

We highlight issues of text data by comparing performance of four feature ranking metrics on sample dataset as discussed in [12], document frequency DF, bi-normal separation BNS [13], information gain IG and odds ratio OR. The equations for the four feature ranking metrics are given by [12] as:

DF=tp+fp

$$IG = e(pos, neg)[P_w e(tp, fp) + \bar{P}_w e(tp, fp)] \qquad (2)$$

$$OR = \frac{tp \times tn}{fp \times fn} \qquad (3)$$

$$BNS = |F^{-1}(tpr)F^{-1}(fpr)| \qquad (4)$$

Where $e(x, y) = -\frac{x}{x+y}\log_2\frac{x}{x+y} - \frac{y}{x+y}\log_2\frac{y}{x+y}$,

$P_w = -\frac{(tp+fp)}{N}$ and $\bar{P}_w = 1 - P_w$

3. Discussion and Conclusions

Ranks assigned by the four feature ranking metrics DF, BNS, IG and OR for different cases of one-against all settings are listed in table 7. Figures 1, 2 and 3 shows scatter plots for all three cases of one-against all settings. Ranks assigned by OR, IG, BNS and DF are listed with each data point respectively. In [12], Forman says that the terms located in top left and bottom right corners are strongly predictive words as they occurrences in one class are very high as compare to the other class.

Table 4. Document frequencies of terms in example dataset given in Table 3 in different categories.

df	t1	t2	t3	t4	t5	t6	t7	t8	t9	t10	t11
c1	3	3	1	3	3	1	0	0	1	2	3
c2	4	4	2	5	1	5	0	0	0	0	1
c3	8	8	13	1	2	2	8	3	0	0	13

Table 5. Concentrations of terms in example dataset given in Table 3 in different categories.

Probability	t1	t2	t3	t4	t5	t6	t7	t8	t9	t10	t11
c1	0.75	0.75	0.25	0.75	0.75	0.25	0.00	0.00	0.25	0.50	0.75
c2	0.67	0.67	0.33	0.83	0.17	0.83	0.00	0.00	0.00	0.00	0.17
c3	0.50	0.50	0.81	0.06	0.13	0.13	0.50	0.19	0.00	0.00	0.81

Table 6. True positive and false positive rates of terms in example dataset given in Table 3 for different cases of one-against-all settings.

c1 is the positive class											
Df	t1	t2	t3	t4	t5	t6	t7	t8	t9	t10	t11
Tpr	0.75	0.75	0.25	0.75	0.75	0.25	0.00	0.00	0.25	0.50	0.75
Fpr	0.55	0.55	0.68	0.27	0.14	0.32	0.36	0.14	0.00	0.00	0.64
c2 is the positive class											
Df	t1	t2	t3	t4	t5	t6	t7	t8	t9	t10	t11
Tpr	0.67	0.67	0.33	0.83	0.17	0.83	0.00	0.00	0.00	0.00	0.17
Fpr	0.55	0.55	0.70	0.20	0.25	0.15	0.40	0.15	0.05	0.10	0.80
c3 is the positive class											
Df	t1	t2	t3	t4	t5	t6	t7	t8	t9	t10	t11
Tpr	0.50	0.50	0.81	0.06	0.13	0.13	0.50	0.19	0.00	0.00	0.81
Fpr	0.70	0.70	0.30	0.80	0.40	0.60	0.00	0.00	0.10	0.20	0.40

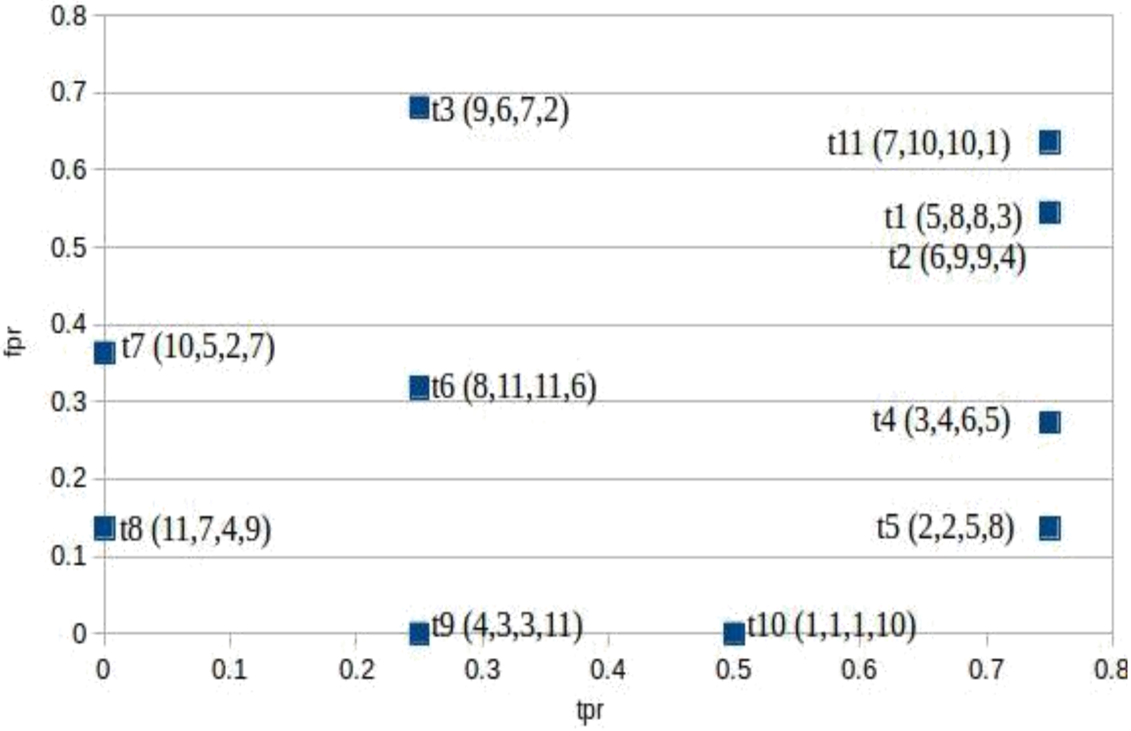

Figure 1. Scatter plot for the sample dataset where c1 is the positive class in one-against-all settings.

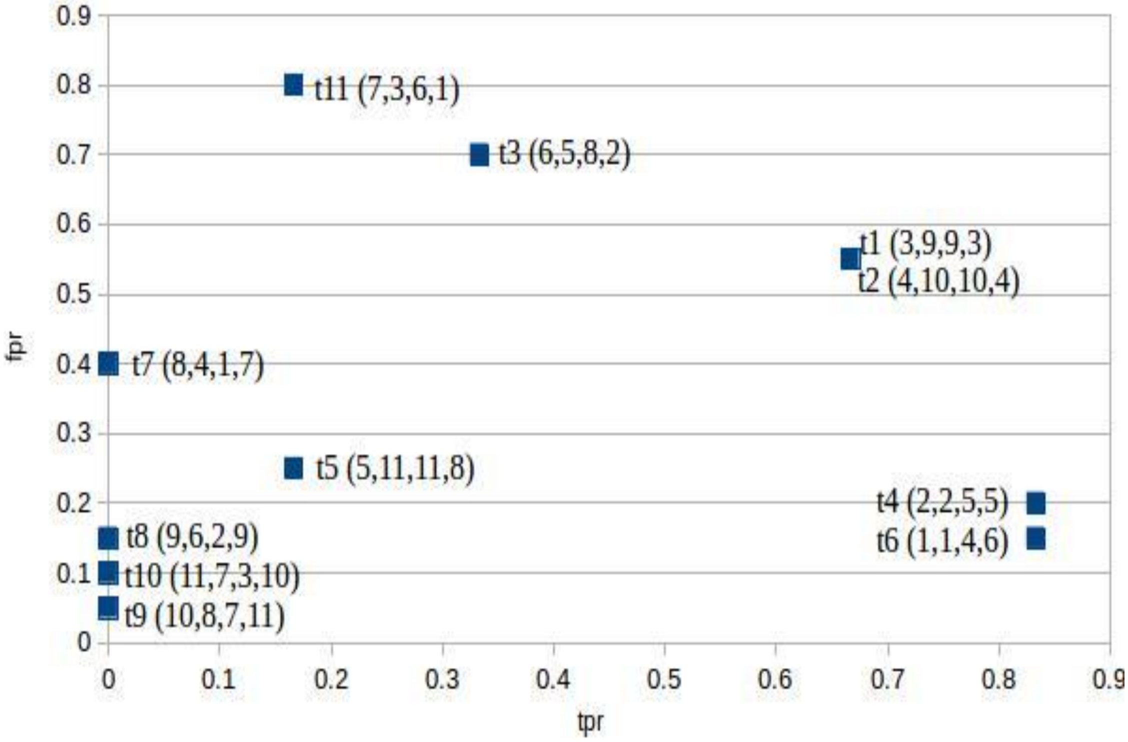

Figure 2. Scatter plot for the sample dataset where c2 is the positive class in one-against-all settings.

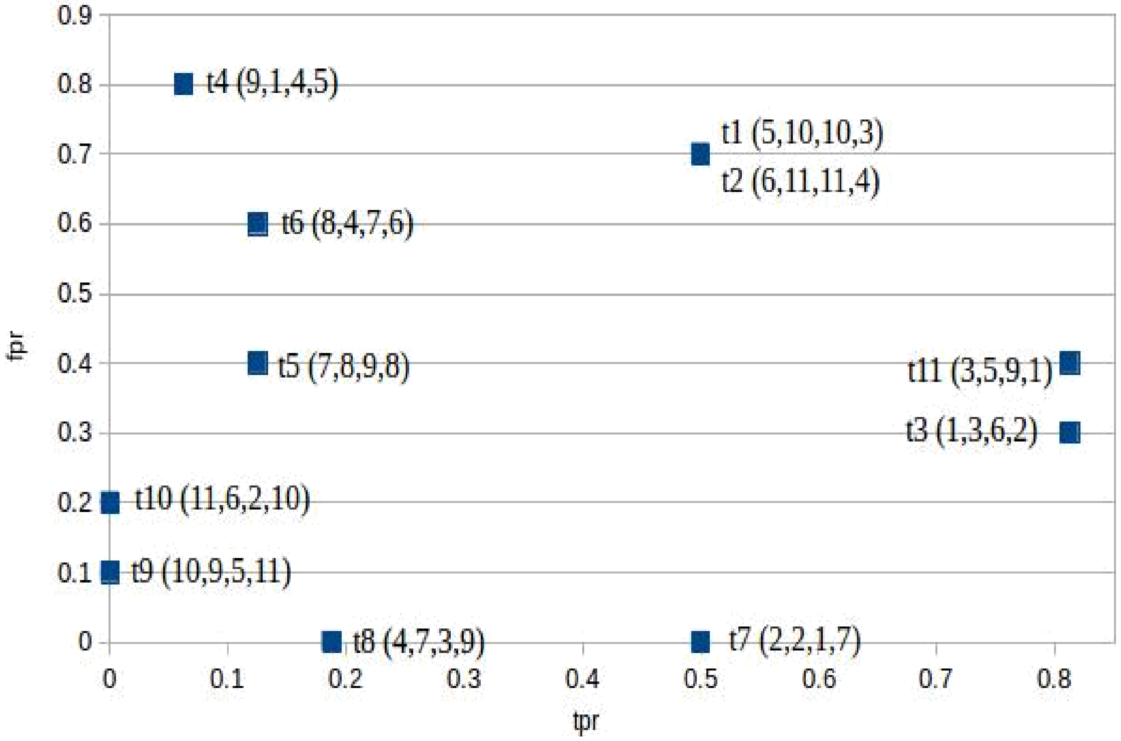

Figure 3. Scatter plot for the sample dataset where c3 is the positive class in one-against-all settings.

Analysis of the three scatter plots shows that data points move away from the diagonal line as the class skew decreases. The class skew decreases as the positive class changes from the class c1 to c3. Scatter plot for c1 has three points t1,t6 and t11 very close to the diagonal line. Scatter plot for c2 has two points t1 and t5 close to the diagonal line. Scatter plot for c3 has no data point close to the diagonal line. We analyze the four feature ranking metrics according to the criteria set by Forman. Our observations are listed below:

3.1 Analysis of Document Frequency

DF Document frequency is a raw measure to estimate the importance of a term. It ranks terms by their document frequencies in the corpus. As document frequency is a global mea-sure, rank assigned by DF is independent of corpus split into positive and negative class. Therefore, the term rank remains the same for all cases of one against all settings, see Table 7.

Terms having stronger association to larger classes get higher document frequencies. DF ranks the terms t11 and t3 the highest as they are strongly associated to the largest class. Terms t1 and t2 also get the higher ranks as they occur frequently in all classes.

Terms associated to only one class t7, t8, t9 and t10 are ranked lowest. The terms t9 and t10 are associated to the smallest class only. So, they do not get high document frequencies. The other two terms t7 and t8 are not strongly associated to the largest class. Their document frequencies also remain low.

The top four terms selected by DF show the poorest feature selection. Three out of top four terms, t1, t2 and t11 are worst terms to select as they are frequent in all classes. Bottom four terms t7, t8, t9 and t10 are also an example of poor feature selection as these terms are associated to only one class, which creates zero entropy.

3.2 Analysis of BNS

BNS shows better performance by assigning top ranks to the terms associated to only one class t7, t8, t9 and t10. In all three cases of one-against-all settings, at least three terms out of top four terms belong to this group. Also the terms t1 and t2 are present in bottom four terms for all three cases of one-against-all settings.

BNS starts selecting terms located on axis. After selecting terms on axis, terms located in top-left and bottom right corners are selected. Then comes terms on top right corner which are closer to the diagonal. Terms located near middle part of diagonal are selected at the end.

BNS unfairly favors terms associated to only one class. The term t3 is a good discriminator as it is highly associated to one class c3 than the other two classes.

Table 7. Comparison of ranks for IG and BNS for the terms in sample dataset. It is evident from ranks of t1 and t2 that both FR metrics completely ignore term frequencies.

c1 is the positive class							
OR		IG		BNS		DF	
t10	24	t10	24	t10	24	t11	9
t5	20	t5	20	t7	16	t3	7
t4	16	t9	20	t9	16	t1	1
t9	16	t4	16	t8	15	t2	1
t1	8	t7	7	t5	12	t4	0
t2	6	t3	1	t4	7	t6	0
t11	0	t8	1	t3	1	t7	0
t6	0	t1	0	t1	0	t5	0
t3	0	t2	0	t2	0	t8	0
t7	0	t11	0	t11	0	t10	0
t8	0	t6	0	t6	0	t9	0
c2 is the positive class							
OR		IG		BNS		DF	
t6	18	t6	18	t7	18	t11	9
t4	14	t4	14	t8	17	t3	7
t1	7	t11	2	t10	15	t1	1
t2	5	t7	1	t6	8	t2	1
t5	5	t3	1	t4	6	t4	0
t3	0	t8	1	t11	1	t6	0
t11	0	t10	1	t9	1	t7	0
t7	0	t9	1	t3	1	t5	0
t8	0	t1	0	t1	0	t8	0
t9	0	t2	0	t2	0	t10	0
t10	0	t5	0	t5	0	t9	0
c3 is the positive class							
OR		IG		BNS		DF	
t3	10	t4	17	t7	18	t11	9
t7	9	t7	9	t10	16	t3	7
t11	6	t3	2	t8	15	t1	1
t8	6	t6	1	t4	8	t2	1
t1	1	t11	1	t9	8	t4	0
t2	1	t10	1	t3	2	t6	0
t5	1	t8	1	t6	1	t7	0
t6	1	t5	1	t11	1	t5	0
t4	0	t9	1	t5	1	t8	0
t9	0	t1	0	t1	0	t10	0
t10	0	t2	0	t2	0	t9	0

Minimum rank assigned by BNS for t3 is 6. BNS ranks the term t9, which is present in only one document, higher than the term t3, which is present in sixteen documents.

3.3 Analysis of IG

IG favors positive features over negative features. In all three cases of one-against-all settings, the positive features are ranked higher than the negative features.

IG ranks higher the terms located in top left and bottom right corners. Term located in these two corners have high ACC value. Although terms in these two corners are equally important, IG ranks terms in bottom left corner higher than the terms located in top right corner. After selecting terms in two corners, terms located on axis are selected. Diagonal terms are ranked lowest by IG.

3.4 Analysis of OR

In all three cases of one-against-all settings, OR ranks positive features higher than negative features. It starts selecting features from right half of the feature space and moves towards left. In general, terms having higher tpr values are ranked higher than those having lower tpr values.

OR remains un-affected by association of the term to the larger class. In all three cases, top four terms selected by OR are different except t4 which is common for two cases. So each term is present in top four terms in either of the three cases. Even the worst discriminative terms t1 and t2 are present at third and fourth position for the case where c1 is the positive class.

References

[1] Li-Ping Jing, H.K. Huang and H.B. Shi, Proc. of Int. Conf. on Improved Feature Selection Approach to *tfidf* in Text Mining, Machine Learning and Cybernetics **2** (2002) pp. 944-946.

[2] P.M. Ciarelli, E. Oliveira, C. Badue and A.F. De Souza, International Journal of Computer Information Systems and Industrial Management Applications **1 (**2009) 133.

[3] X.B. Xue and Z.H. Zhou, IEEE Transactions on Knowledge and Data Engineering **21**, No. 3 (2009) 428.

[4] H. Moisl, Data Normalization for Variation in Document Length in Exploratory Multivariate Analysis of Text Corpora (2008).

[5] A. Singhal, G. Salton and C. Buckley, Length Normalization in Degraded Text Collections, Proceedings of Fifth Annual Symposium on

Document Analysis and Information Retrieval, (1995) p. 1517.

[6] T. Joachims. Proc. of 24th Annual Int. ACM SIGIR Conference on Research and Development in Information Retrieval, NY, USA (2001) p. 128.

[7] M. Baroni, 39 distributions in Text, University of Trento (2005). http://clic.cimec.unitn.it/marco/publications/hsk_39_dist_rev2.pdf

[8] M.J. Saary, Journal of Clinical Epidemiology **61**, No. 4 (2008) 311.

[9] D. Fouarge and R. Muffers, Social Exclusion in European Welfare States, urn:nbn:nl:ui:27-21326. (2002).

[10] David D. Lewis. Reuters-21578.

[11] G. Forman, Proceedings of the 21st Int. Conf. on Machine Learning, ACM (2004) p. 38.

[12] G. Forman, I. Guyon and A. Elisseeff, Journal of Machine Learning Research **3** (2003) 12891305.

[13] Y. Yang and J.O. Pedersen, Proc. of the 14th Int. Conf on Machine Learning, San Francisco, CA, USA (1997) p. 412.