International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2014): 5.611

Bootstrapping in Text Mining Applications

C. K. Chandrasekhar[1], M. R. Srinivasan[2], B. Ramesh Babu[3]

[1]Department of Library & Information Science, University of Madras, Chennai, India

[2]Department of Statistics, University of Madras, Chennai, India

[3]Department of Library & Information Science, University of Madras, Chennai, India

Abstract: ***Text mining involves analyzing large corpora of documents with thousands of words with a high level of noise content. Dimensionality reduction, noise mitigation, accurate and stable cluster formation are principal challenges of upstream analytics. This paper proposes a methodology for dimensionality as well as noise reduction using k-fold rotation estimation. Principal Component Analysis enables selecting a reduced set of dimensions (words). The resulting noise-reduced dataset is the input to clustering algorithms. Experiments using benchmark datasets from the Brown corpus [5] and real life feedback data of a service provider show that our approach delivers improved results using the well-known performance measures: recall, precision, and F-measure [14]. We used combination of projective transforms known as principal component analysis (PCA) and visual scree plot techniques [8,6,12] for dimensionality reduction and a k-Fold rotation sampling technique [1] for noise elimination and formation of stable clusters. Experimental results with corpora of different sizes demonstrate that the approach delivers improved clustering accuracy than standard k-means clustering algorithm [2].***

Keywords: k-Fold Rotation Estimation, Clustering, k-Means, Principal Component Analysis, Dimensionality Reduction, Precision, Recall, F-Score, Scree Plot

1. Introduction

The prototypical predictive text mining application is text categorization, email spam filtering being popular instance of this. Classification is a well understood problem. The solution to the problem is a mathematical function that maps examples to labels f: w→ L, where w is vector of attributes and L is a label. In the context of text categorization, examples are drawn from a heterogeneous set of text documents called a corpus, attributes are words and labels are broad topic areas of the document. The input to the classification function is fed in the form of a term-document matrix derived after several stages of cleansing and transformation of data. The classifier is an algorithm that separates unlabeled N number of input documents into L labeled output folders where L ≪ N[13]. Term document matrices tend to be huge but sparse, having thousands of attributes and hundreds of examples. In order to achieve the best clustering performance, finding optimal set of dimensions (words) is critical. Towards that end, additional data pre-processing that involves sample datasets generated by k-fold rotation estimation is proposed. Once the noise-smoothed dataset with reduced number of dimensions is prepared, it is input to a clustering algorithm. The clustering method assigns a cluster id to each document, the required number of cluster being the choice of the analyst. Using well known measures of classification accuracy like precision, recall and F-score [6] we compared the results obtained by plain k-means clustering to that of our approach. The results indicate significant improvement in clustering accuracy.

The design of this paper is as follows:

1) Description of Proposed Approach
2) Model Assessment using Simulated Data
3) Model Validation using Real Life Data

1) Description of the Proposed Model

In a typical clustering approach a document corpus is submitted to a clustering algorithm hoping for the algorithm to categorize the incoming documents into c clusters. If the documents come from k topic areas that are well differentiated, one may expect the algorithm to classify the documents into same areas provided the c equals k, the number of topic areas. Taking this view we may assess the performance of clustering method by defining measurements that capture the accuracy of the classifier.

True Positives (TP): Relevant documents identified correctly as belonging to topic area.
True Negatives (TN): Irrelevant documents correctly identified as not belonging to topic area.
False Positives (FP): Irrelevant documents incorrectly identified as belonging to topic area.
False Negatives (FP): Relevant documents incorrectly identified as not belonging to topic area.
Accuracy measures the percentage of input documents that have been correctly classified. If the corpus had n documents and the classifier has correctly classified p of them and incorrectly classified n-p of them, then the accuracy is $(\frac{p}{n} \times 100)\%$.

Precision is the proportion of relevant documents to total number of documents identified by the classifier as relevant. It is a measure of purity of the class formed by the classifier and is given by $(\frac{TP}{TP+FP})$.

Recall is the proportion of the relevant documents identified as relevant by the classifier to all the relevant documents in the corpus. It is the ability of the classifier to correctly classify all relevant documents belonging to the class and is given$(\frac{TP}{TP+FN})$.

F-Score combines precision and recall to give a single score, is defined as harmonic mean of precision and recall i.e. $\frac{2\times Precision \times Recall}{Precision + Recall}$[4].

k-Folding is simply a resampling technique with replacement. Given a random sample of *n* vectors, $X_B = [x_1, x_2, \ldots, x_n]$, k-Folding involves generating a sample by deleting the i^{th}vector$x_i$and repeating this procedure for all X_B. The benefit is that the multiplicity of samples is leveraged to estimate statistics [1]. In this paper, a term-document matrix is used to determine the optimal set of principal components. This is achieved by computing the Eigen values of the covariance of the term-by-document matrix *X*. Optimally reduced set of dimensions is determined based on the average of the Eigen values calculated from the repeated k-Fold samples. In other words, we use multiple versions of the corpora to obtain a reliable dataset. A reliable dataset is one that is noise-free, with fewer dimensions, and retains properties that best characterize the original dataset [3]. Each Eigen value of the covariance of a dataset is associated with a principle component and all Eigen values together capture the total information in the dataset. Eigen values with larger magnitudes carry more information. Eliminating very small Eigen values is tantamount to deleting corresponding principal components, which in turn are composites of the distinctive terms in the document corpus!

Operationally, dimensionality reduction is achieved via the following procedure using *scree plots*. The scree plot is a visual graph that displays the ordered (largest to smallest) cumulative Eigen values $\sum \lambda_i$(*Y*-Axis) versus $\{1,2,\cdots,k\}$(*X*-axis), where *k* is the number of Eigen values. By construction, as Eigen values get added, the low magnitude Eigen values contribute smaller and smaller amounts such that at some point in the Y-axis, the change in the cumulative sum becomes negligible. The number of retained dimensions is that point on the X-axis where the corresponding cumulative sum on the Y-axis plateaus (forms an elbow) [15]. Since we are using SAS® code to implement the algorithm an automated process is preferable. The scree cutoff programmed by us incrementally accumulates the ordered set of Eigen values $(\lambda_i s)$ until a preset cutoff is reached. Only these many dimensions are retained in the reduced representation.

Once the noise-smoothed dataset with reduced number of dimensions is prepared, it is input to a clustering algorithm. In this paper, the popular k-Means algorithm is selected [6]. The practitioner may choose any other method such as E-M, hierarchical clustering among others.

Statistical clustering involves methods to assign an observation to one of a finite number of groups. More formally, clustering attempts to partition a heterogeneous set of observations into natural groups such that within group homogeneity is small and between-group heterogeneity is large, i.e., the signal-to-noise ratio (SNR) is high. k-Means iteratively assigns observations into one of k clusters using a distance metric under a minimization criterion. If x_i is the i^{th} document vector, and $c_j \in \{1, \ldots, k\}$ be its corresponding cluster index and m_{cj} is the centroid of the "j" clusterc_j. The objective function for obtaining the best SNR is given by; $\sum_{j=1}^{k} \sum_{i=1}^{n_j} \left(x_{ij} - m_{c_j}\right)^2$. [2]

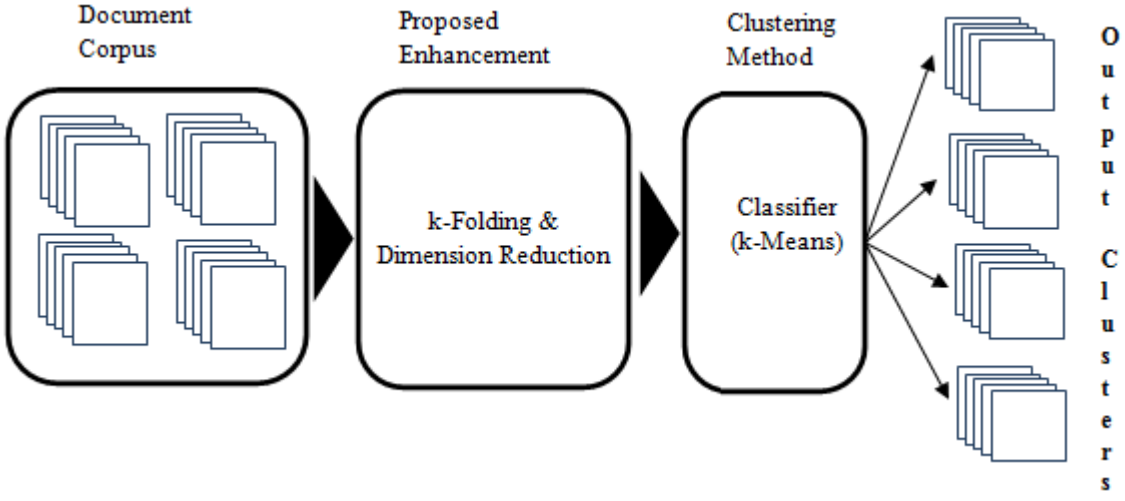

Figure 1: Proposed Approach

The block diagram in figure 1 captures the essence of our approach.
Algorithmically, our end-to-end application involves;

1) Take the corpus and preprocess to remove punctuations, redundant data and data with low information content i.e. stop words.
2) Build document by term matrix.
3) k-Fold the sample corpus.
4) For each sample, compute the principal components, Eigen values and vectors.
5) Compute the average of all samples processed
6) Decide scree cutoff and build the reduced dimension matrix.
7) Form the k-means clusters.
8) Compare against clusters formed with original document term matrix (Ground Truth), i.e. before k-folding.

2) Model Assessment using BrownCorpus data

For a preliminary assessment of the model, we needed a set of documents of manageable size, and from a corpus wherein document structures are fairly homogeneous within a topic area and well separated between topic areas. As the real life documents tend to be large and unwieldly, we used a simulated document-term structure. Brown corpus is the first serious initiative to digitize text documents and create a library of English Text. It consists of 500 samples of 2000 words each of English text of various genres. From the Brown corpus [7] created by Brown University, we chose 3 topic areas namely Sports, Politics and Religion. The Brown corpus has a well categorized structure as shown in figure 2 below:

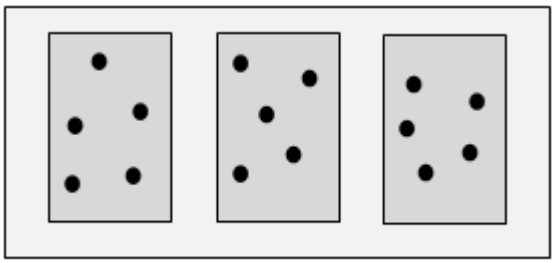

Figure 2: Structure of Brown Corpus

In the simulation design, the frequency distributions of terms from all the documents from the chosen topic areas were analyzed. Using respective frequency distributions of individual topic areas 60 simulated documents were created with each having approximately 200 terms. A part of this matrix is shown in figure 3 below:

Doc No	Category	AMERICA	ASSEMBLY	BICKER	BLABBER	BLAISE	CHALISE	CHOCOLATE	CRICKET	CUBICLE	DEMOCRACY	DOMINATE	DROP	ECHEW	ENGINE
Doc1	Sports	0	0	0	0	0	0	1	1	0	0	0	0	0	1
Doc2	Politics	1	1	0	0	0	0	0	0	0	1	0	0	0	0
Doc3	Religion	1	0	0	0	0	0	0	0	0	0	0	0	0	0
Doc4	Sports	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Doc5	Politics	1	0	1	0	0	0	0	0	0	0	1	0	1	1
Doc6	Religion	1	1	0	0	1	0	0	0	0	0	0	0	0	0
Doc7	Sports	1	0	0	1	0	0	0	1	0	0	0	0	0	1
Doc8	Politics	0	0	1	1	1	0	0	0	1	0	0	1	1	1
Doc9	Religion	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Doc10	Sports	1	0	0	1	0	0	0	1	0	0	0	1	0	1
Doc11	Politics	1	1	0	1	0	1	0	0	0	1	1	0	1	0
Doc12	Religion	1	0	0	1	1	1	1	0	0	0	0	0	0	0
Doc13	Sports	1	0	0	0	0	1	0	2	0	0	1	1	0	0
Doc14	Politics	0	1	0	1	1	0	1	0	0	1	1	1	0	0
Doc15	Religion	0	1	0	1	0	1	0	0	0	0	0	0	0	0
Doc16	Sports	1	0	0	0	0	0	1	0	0	0	1	0	0	0
Doc17	Politics	1	1	0	1	1	0	0	1	0	0	0	1	0	0
Doc18	Religion	1	0	0	1	0	1	0	0	0	0	0	0	0	0
Doc19	Sports	1	0	0	1	1	0	0	1	0	0	0	1	0	0

Figure 3: Partial Document-Term (DT) Matrix of simulated dataset

SAS® FASTCLUS [11] procedure performs a disjoint cluster analysis on the basis of distances computed from one or more quantitative variables. The observations are divided into clusters such that every observation belongs to one and only one cluster. FASTCLUS procedure uses Euclidean distances, so the cluster centers are based on least squares estimation. This kind of clustering method is called k-Means model, since the cluster centers are the means of the observations assigned to each cluster when the algorithm is run to complete convergence. Each iteration reduces the least squares criterion until convergence is achieved.

The DT matrix was submitted to FASTCLUS procedure using the following code

```
PROC FASTCLUS DATA=boot.finmat80 MAXC=3 MAXITER=10 OUT=clus;
 VAR _NUMERIC_;
RUN;
data boot.result80 (keep=category cluster assign);
setclus;
assign="Bad ";
if category = "Sports" and cluster = 1 then assign = "Good";
else if
category = "Politics" and cluster = 2 then assign = "Good";
else if
category = "Religion" and cluster = 3 then assign = "Good";
else assign = "Bad";
run;
```

Figure 4: SAS® Code for k-Means Clustering

The resulting clusters from the FASTCLUS procedure were identified and appropriately labelled. These results are shown below:

Correctly Classified	38
Incorrectly Classified	22

		Actual categories		
		Sports	**Politics**	**Religion**
Assigned Categories	**Sports**	11	7	2
	Politics	7	10	3
	Religion	1	2	17

	Accuracy	**Recall**	**Precision**	**F-Score**
Sports	0.717	0.550	0.579	0.564
Politics	0.683	0.500	0.526	0.513
Religion	0.867	0.850	0.773	0.810

Figure 5: Results of Simulated Run without k-Fold Rotation Sampling

From the results we see that the k-means classifier has done a poor job achieving only a little more than 70% accuracy. The other results are also on similar lines. In the next experiment the same document term matrix was submitted to a pre-processing technique comprising of k-Fold Rotation Estimator followed by PCA dimension reduction using a cutoff criterion of 0.90 i.e. only Eigen values(ordered by value) having a cumulative magnitude of 0.9 are retained. The numbers of dimensions were reduced to 42 from original 217.

Principal Component Analysis (PCA) is technique that is used extensively for dimensionality reduction while retaining intrinsic information in the original data. PCA serves the dual need of dimensionality reduction and eliminating collinearity amongst features. Collinearity is a problem that degrades performance of classification methods. Consider a corpus having n documents and p distinct terms (features) from all the documents in the corpus. Principal components are particular linear combination of the p-features $x_1, x_2,\ldots x_p$ in the input pattern vector. These linear combinations represent a new coordinate systems with $x_1, x_2,\ldots x_p$as directions with maximum variability[9]. Although p principal components are required to account for the total variability, majority of variation is captured by a small number of m components. If X is n x p matrix having frequencies from n documents and p terms then X`X is a p x p variance-covariance matrix.

X`X matrix may be decomposed as
$\mathbf{X`X} = \mathbf{U\Lambda U`}$
Where $\mathbf{\Lambda_{pxp}}$ is diagonal matrix with diagonal elements as the eigen values λ_j (j= 1,2, …p) of **X`X**.
and $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$
$\mathbf{U_{pxp}}$ is an orthogonal matrix whose columns are eigen vectors associated with above eigen values.
if y_i is the i^{th} eigen vector of **X`X** then variance of y_i is λ_i
covariance $(y_i, y_j) = 0$

Our approach is to iteratively obtain a k-fold sample of documents, form a document-term matrix and decompose it to extract Eigen values and Eigen vectors. This process is repeated n times where n is the number of documents in the corpus. The averages of n λ_is and n y_is obtained from n samples is computed i.e.

$$\bar{\lambda}_i = \sum_{j=1}^{n} \lambda_{ij}$$

$$\bar{e}_{ij} = \sum_{k=1}^{n} e_{ijk}$$

Where λ_i is i^{th} Eigen value and e_{ij} is coefficient of j^{th} component i^{th} Eigen vector.

A screen plot is used decide the cutoff value c, the number of Eigen vectors to be included in the reduced DT matrix. The value of c is chosen based on Eigen values. It is chosen in such a way that cumulative value $\sum_{i=1}^{c} \lambda_i$ up to i captures majority variation in the system. A scree plot is drawn to decide the cutoff. The scree cutoff programmed by us incrementally accumulates the ordered set of eigen values $(\lambda_i s)$ until a preset cutoff is reached. In this case the cutoff is chosen such that 90% of variability is explained. Only these many dimensions are retained in the reduced representation.

The scree plot for the simulated data is drawn in figure below. If we set cutoff at 90%), then the cumulative value of λ_i seen from the figure is 195 and corresponding i is 42. Therefore in the reduced DT matrix 42 terms are retained.

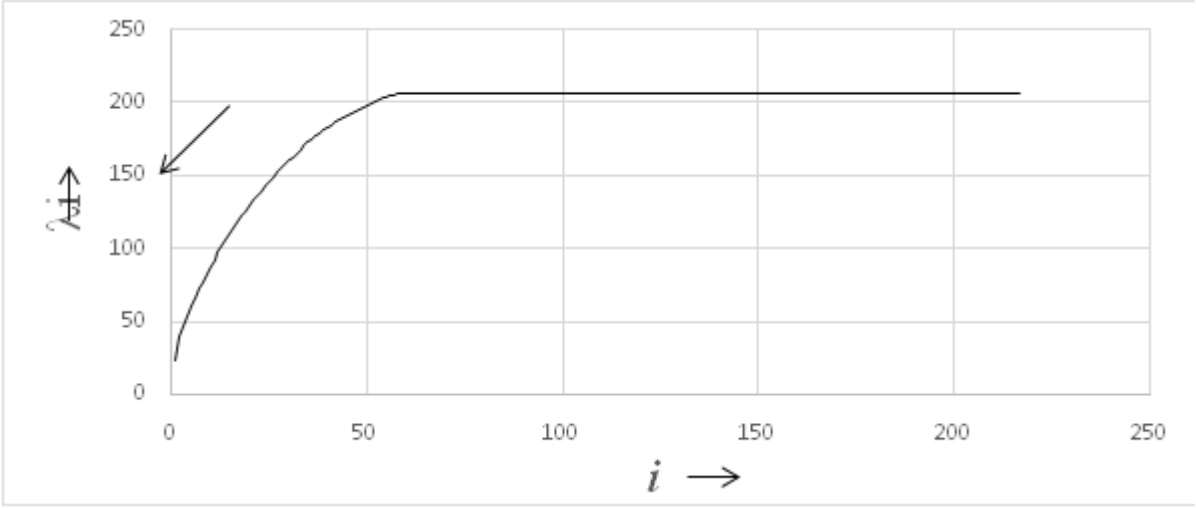

Figure 6: Scree plot to decide number of retained dimensions

The same graphical cutoff criterion can also be obtained mathematically. Let the number of terms be p. So corresponding to these p terms there are p eigen values λ_1 to λ_p. If the cutoff is 90% then i is the smallest value satisfying the condition

$$i = x: \sum_{j=1}^{x} \lambda_{(j)} \leq 0.9 \sum_{j=1}^{p} \lambda_{(j)}$$

The dimensionally reduced document term matrix was submitted to FASTCLUS with same parameters as before and the results are produced below:

Method: K-Fold, Reduced Dimension k-Means Clustering
Number of documents in corpus = 60
Scree Cutoff = 90%;
No of retained dimensions = 42/217

Correctly Classified	60
Incorrectly Classified	0

		Actual Categories		
		Sports	Politics	Religion
Assigned Categories	Sports	20	0	0
	Politics	0	20	0
	Religion	0	0	20

	Accuracy	Recall	Precision	F-Score
Sports	1	1	1	1
Politics	1	1	1	1
Religion	1	1	1	1

Figure 7: Results of Simulated Run with k-Fold Rotation Sampling

From the results we observe that the proposed method is extremely robust against noise performing at 100% accuracy dramatically improving the clustering accuracy of the plain k-Means method.

The plain k-Means could achieve only 71% accuracy and about 55% on precision, recall and F-Score statistics, whereas when augmented with k-Folding and dimensionality reduction it has perfectly segmented the corpus in to clusters with 100% accuracy. This is a dramatic improvement over the direct use of clustering method.

3) Model validation using Real life case

To evaluate the performance on real life data we used free-form customer feedback data from the website of an IT products and services company. This data is collected from the Service Provider's website as a part of freeform feedback customers provided after availing a service. The services provided belonged to three categories namely, (i) downloading software and device drivers. (ii) Issues relating to printers supplied by the service provider. (iii) Feedback about the layout, ease of use and information available on the service provider's website. In the text box provided the user is requested to provide free-form descriptive comments up to 500 words. The A sample of the data set is produced in Figure 8 below.

As of any raw data, the raw feedback received from respondent has to be cleaned and transformed before a document-term matrix is structured. (i) Tokenization: the free form feedback text was split into terms which may be words, numerals, proper nouns, punctuation marks etc. The token is an indivisible element in the document term matrix. (ii). Stop words are the words which have very less information content or discriminating power. Articles, prepositions, punctuation marks belong to this category and are removed from the list of words in documents. (iii) Rare words which occur only once in one document and never in any other document also have little discriminating power. These are generally errors, typos etc. These words are also removed from the list of words in the document.(iv) Synonyms are words which mean the same but clustering algorithms see them as different. For example the terms URL and Webpage mean the same. All the synonyms are replaced by single representative word in the documents wordlist. (v) Verbs and nouns expressed in different tense or

form mean the same. Work, worked, working all mean the same in the context of clustering and therefore are reduced to their root form "work". All numbers are also removed from the document's wordlist [10]. After performing all the above pruning operations the residual terms in the document are used to form document term matrix. The term document term matrix from customer feedback data is shown in the figure.

Feedback Id	Feedback Comment
D013025	inadequate information available on laserjet printers prior to making purchase decisions needed information on toner cartridge compatible printer selected apn your laserjet printer only to find out after the purchase that the printer would not work then after extensive researching on this website finding an obscure document indicating that the model printer usb parallel port is not compatible with terminal services printing an that the toner cartridge the next model up is compatible this just cost my client over
D031669	we are having problems with a blinking print cartridge light on our inkjet usb printer disappointed that we cannot obtain any online or telephone support of any kind because the warranty has expired unlike epson printer who were quite happy to talk thorough my personal printer problems on the telephone despite warranty expiring eventually found a deskjet question and parallel port answer section that had the same problem answered cannot believe that i have to press together.
D036728	your driver sections are very incomplete i download a your model xxxx drivers with winos and had unknown devices according to the hardware manager you d think i could go to the your website select my computer and os or winos then download compatible version of software drivers as you can with dell instead the only driver that was present was for a chameleon modem and the megabyte binary file that i downloaded was corrupt extreme waste of time
D051059	i came to website to find some information on my aged model xxxx last time i looked for information compaq url link was still compaq and i found navigation through the website to be much more better than now i did not find the information i needed a simple button click and no registration and online survey so i am giving up as i was about to leave this site. Survey pops up appalls me yet again.
D055727	i have a color laserjet model xxxx well i think that it is a laserjet xxxx i looked everywhere on the box and the printer with no set description of whether i had a xxxx xxxxi or xxxxi love the printer scared of buying cartridges and toner. I do have some difficulty networking and utilizing the usb 2 other than those love the printer. The website has more info on deskjet inkjet printers.
D062704	once i got to this site it was very easy but i did not find the compatible driver version it rather went to a driver site that had a link for your drivers for winos as they did not have the driver i was after why cannot you make the software now fully plug and play by connecting your printer to a computer it should say found device do not have software but device has given me its web site address and os code detail to get latest from the driver manufacture open you browser
D086557	the driver downloads site needs to be clearer with respect to what is needed to be downloaded to get the driver in my case i was looking for a driver compatible to your model xxxx latest software version. driver for winos server i downloaded your dss workflow 128 megabyte because it was the only thing there seems a bit large for a driver file or did i corrupt it in download what i was supposed to click on to only get the driver
D091178	this site does not offer what we need. they want you to register and take survey. The site does not has no information too much navigation, many pages require many clicks, hidden buttons and 0 information. no support, link urls not working, confusion you want all this useless information before getting to my problem all you printer models should be the same there is no need to have the serial number when trying to contact support.
D096630	over the past couple of days i have been trying to get to the support web page for my model xxxx some days i can find the link with no problem some days i cannot seem to find the url at all what is more distressing is when i type in my device model number in the online search engine sometimes it tells me that it cannot find anything on your web page about the model xxxx a month or so ago all i had to do is type in model xxxx on the service and support page.

Figure 8: Customer freeform feedback data set

DOC_NO	access	adapter	alert	audio	bit	blink	board	browse	button	cable	cartridge	click	color	compatible	corrupt	deskjet	download	driver
D013025	0	0	0	0	0	0	0	0	0	0	2	0	0	3	0	0	0	0
D031669	0	0	0	0	0	1	0	0	0	0	2	0	0	0	0	1	0	0
D036728	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	3	4
D051059	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0
D055727	0	0	0	0	0	0	0	0	0	0	2	0	1	0	0	1	0	0
D062704	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	3
D086557	0	0	0	0	1	0	0	0	0	0	0	1	0	1	1	0	4	6
D091178	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0
D096630	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
D096742	0	0	0	2	0	0	2	0	0	0	0	0	0	1	0	0	1	2
D096907	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0
D102724	0	2	0	0	0	0	0	0	0	1	1	0	0	1	0	0	0	0

Figure 9: Partial Document-Term (DT) Matrix of customer feedback dataset

The same procedure outlined earlier for simulated data was followed for this data set too. And the results obtained are shown in table in figure 10.

Correctly Classified	42
Incorrectly Classified	24

		Actual Category		
		Driver	Printer	Website
Assigned Category	Driver	14	4	4
	Printer	5	12	3
	Website	4	4	16

	Accuracy	Recall	Precision	F-Score
Driver	0.74	0.64	0.61	0.62
Printer	0.76	0.60	0.60	0.60
Website	0.77	0.70	0.67	0.68

Figure 10: Results from feedback data set without k-Fold Rotation

Method: k-Fold Reduced Dimension K-Means Clustering
Number of documents in corpus = 66
Scree Cutoff = 90%;
No of retained dimensions = 71/320

Correctly Classified	64
Incorrectly Classified	2

		Actual Category		
		Driver	Printer	Website
Assigned Category	Driver	22	1	0
	Printer	1	19	0
	Website	0	0	23

	Accuracy	Recall	Precision	F-Score
Driver	0.97	0.96	0.96	0.96
Printer	0.97	0.95	0.95	0.95
Website	1.00	1.00	1.00	1.00

Figure 11: Results from feedback data set with k-Fold Rotation

From the results it is clear that our enhancement has improved the robustness and accuracy of clustering capability of k-Means cluster method significantly and performed exceedingly well on all statistics (accuracy, precision, recall, F-Score that we have set up to measure its performance.

The plain k-Means could achieve only 75% accuracy and about 64% on precision, recall and F-Score statistics, whereas when augmented with k-Folding and dimensionality reduction the performance was boosted achieving an accuracy in excess of 95% on all statistics. This is a dramatic improvement over the direct use of clustering method.

2. Conclusion, Limitations and Application Areas

From the foregoing discussion we are gratified to discover that our approach has significantly improved the performance of a simple k-Means method on all aspects. There is a slight tradeoff between the computational resource requirements and improvement that is achieved. If the document corpus contains N documents then our approach requires n matrix decompositions, which is a computationally intensive task. More so when the input document-term matrix is very large. However this demand for additional computational resource is offset by the fact that the clustering procedure will be working with much less number of dimensions. In our case the dimension reduction was about 80%.

The k-folding technique, PCA decomposition techniques are computationally intensive and this may be a potential deterrent in use of our technique. On our laboratory computer with core duo processor and 2 GB RAM running SAS 9.1.3 the time taken was less than a second and hardly noticeable for the sizes of the documents under consideration. For larger corpora and large documents the performance may become an issue. However in view of improvement in clustering accuracy we suggest implementation of our approach in areas where accuracy is important and cost of miss classification is severe. Applications like E-mail sorting, customer feedback response are good areas. The short document size in these applications does not require heavy computational resources.

If the average document size is small, the extra demand for computational resource for matrix decomposition and iteration thereof is negligible. In such cases the significant improvement achieved by our approach is well worth it. The following are some of the applications where our approach is worth considering.

1) E-Mail Filtering: E-Mail messages tend to be short averaging about 150 to 200 words. The benefit of providing only relevant e-mails is considerable because this removes all spam and improves the end-user productivity by providing him with only important messages. The cost of missing a business opportunity is also controlled because the misclassification is reduced considerably.
2) Article abstracts: Article abstracts are short documents averaging less than 500 words and are therefore well-suited for categorization using our approach. The benefit is highly accurate classification so that an end-user who wishes retrieve documents from a particular topic area is precisely provided the relevant documents with confidence that he is not missing out any document of interest.
3) Customer Feedback Comments: The practice of collecting free form comments from customers on websites other locations after a service is rendered is quite common. These comments are critical because the structured questions that precede the free form comment might not have captured all the feelings of the customers about the service or the deficiency thereof. Such comments tend to be short averaging about 300 to 400 words and the lexicon of words is also small. Proper

classification is very important to ensure quick response to customer issues and control of misclassification ensures minimal customer dissatisfaction.

References

[1] Bradley. Effron, Robert. Tibshirani. An Introduction to the Bootstrap, Chapman & Hall, 1993
[2] Cutting D, Karger K [1992]: A cluster based approach to browsing large document collection. Proceedings of SIGIR-92, Pages 1-12, ACM Press
[3] Dhillon I and Modha D [2001]: Concept decompositions for large sparce text data using clustering. Machine Learning, 42(1): 143-175
[4] Forman G. [2003]: An Extensive empirical study of feature selection metrics for text classification, Journal of Machine Learning Research 3:1289-1305, 2003
[5] Francis, WN, Kucera H [1979] THE BROWN CORPUS: A STANDARD CORPUS OF PRESENT-DAY EDITED AMERICAN ENGLISH (computer file), Providence, RI: Department of Linguistics, Brown University [producer and distributor], 1979.
[6] Hastie T, Tibshirani R, Friedman J [2001]: "The Elements of Statistical Learning" Springer Series in Statistics Springer New York Inc., New York, NY, USA, (2001)
[7] http://www.clu.uni.no/icame/brown/bcm.html
[8] Johnson RA, Wichern DW[1992]: Applied Multivariate Statistical Analysis, 3rd edition, Englewood Cliffs, New Jersey: Prentice Hall, (1992)
[9] Jolliffe, I. [1986]: Principal Component Analysis. Springer-Verlag. MR0841268
[10] Lakshminarayan C, Yu Q, Benson A.[2005]: Improving customer experience via text mining. In DNIS. 288–299.
[11] SAS® SAS Institute Inc. Cary, NC, USA.
[12] Shalizi CR [2013]:, Advanced Data Analysis from an Elementary Point of View, www.stat.cmu.edu/~cshalizi, (2013
[13] Weiss MW, Indurkhya N, Zhang T [2004]: Text Mining: Predictive Methods for Analyzing Unstructured Information
[14] Wikipedia -- http://en.wikipedia.org/wiki/Precision_and_recall
[15] Zhu M., Ghodsi A. [2006]: Automatic dimensionality selection from the scree plot via the use of profile likelihood", Computational Statistics & Data Analysis 51 (2006) 918 – 930