A NEW VARIABLES SELECTION AND DIMENSIONALITY REDUCTION TECHNIQUE COUPLED WITH SIMCA METHOD FOR THE CLASSIFICATION OF TEXT DOCUMENTS

Ahmed Abdelfattah Saleh
University of Brasilia, Brasil
ahmdsalh@yahoo.com

Li Weigang
University of Brasilia, Brasil

Abstract:

Classification of text documents is of significant importance in the field of data mining and machine learning. However, the vector representation of documents, in classification problems, results in a highly sparse data with immense number of variables. This necessitates applying an efficient variables selection and dimensionality reduction technique that ensures model's selectivity, accuracy and robustness with fewer variables. This paper introduces a new coefficient, the *Variables Strength Coefficient* (*VSC*), which permits retaining variables with strong *Modeling* and *Discriminatory* powers. A variable with *VSC* greater than a predefined threshold is considered to have strong power in both modeling data and discriminating classes and thus retained, while weaker variables are discarded. This straightforward technique results in maximizing the differences between classes while preserving the modeling power of variables. This paper also proposes applying a classification technique that is widely used in chemical analysis domain; the supervised learning algorithm *SIMCA.* The soft and independent nature of *SIMCA* allows multi-labeling of text documents, in addition to, the ability to include new classes later on without affecting the created model. *VSC-SIMCA* was applied on the data set 'CNAE-9' and the results obtained were compared to classification and dimensionality reduction work done on the same data set in the literature. *VSC-SIMCA* technique shows superior performance over other techniques, both in the amount of dimensionality reduction, as well as, the classification performance. The improved classification precision, with substantial fewer variables, demonstrates the contribution of the proposed approach of this research.

Keywords: VSC, SIMCA, text classification, variables selection, supervised learning

1. INTRODUCTION

Text classification is one of the major topics in machine learning and data mining field with wide range of applications. Literature includes many learning algorithms aiming at text classification and similarity measurements, including Logistic Regression, Ridge Regression, Support Vector Machine (SVM), string kernels, LDA and others (Zhao, He, & Liu, 2014), (Kemp & Tenenbaum, 2008) and (Lin, Jiang, & Lee, 2014). The vector (bag-of-words) representation of documents (Kim, Howland, & Park, 2005) results in large number of variables that should be reduced to improve the computation performance of the model. In addition, noisy data (e.g. data resulted from OCR of printed documents) has significant negative impact on the robustness and accuracy of classification models. As such, several techniques should be applied to achieve data denoising and selection of relevant variables in order to ensure a robust, accurate and selective model.

For the purpose of variables selection this paper introduces a new coefficient, the *Variables Strength Coefficient* (*VSC*), which permits retaining variables with strong *Modeling* and *Discriminatory* powers. In addition to the VSC, this paper proposes the use of a classification algorithm that is widely used in chemometrics, an algorithm known as *Soft Independent Modeling of Class Analogy* or *SIMCA* (Wold, 1976). Soft implies the ability of the model to classify one sample to one or more classes (multi-labeling); this is due to the fact that two classes can overlap (and hence are 'soft'). Independent modeling on the other hand refers to creating separate models for individual classes; such feature is of significant importance when additional classes are to be added to the model without affecting the pre-created model. In contrast, other techniques, as classical discriminant analysis, the entire modeling procedure must be repeated if extra numbers of groups are to be added, this is due to the fact that the pooled variance–covariance matrix must be recalculated.

Principal Component Analysis (PCA) lies at the core of SIMCA modeling. PCA is considered as a denoising technique, since principal components (PC's) that span noise are discarded while those spanning actual variance in the data are retained. SIMCA coupled with VSC achieves both denoising of data, as well as, reduction of dimensionality, while building a robust and selective classification model.

2. CLASSICAL SIMCA

The first SIMCA method introduced by Wold in (Wold, 1976) and (Wold & Sjostrom, 1987), is based on creating a separate PCA sub-model for each class and calculating its boundary. A new sample is assigned to a particular class if it lies within those class boundaries. The sample's overall distance from each class boundary is computed through the linear combination of the sample's Orthogonal Distance (OD) and Score Distance (SD) from that class PCA sub-model. F-Test is used to compare sample's overall distance to all class boundaries, thus indicating whether a sample belongs to a specific class or not.

Suppose there is a training set $\boldsymbol{X}$ composed of *H* text documents belonging to *J* classes. The training set $\boldsymbol{X}$ is subdivided into *J* class matrices. A class matrix $\boldsymbol{X}^j$ is of size (H^j × *N)*, where H^j is the number of documents belonging to that class and *N* is the total number of words (variables) in all *H* documents. Vector representation approach is used to represent documents in all class matrices, where each row of all $\boldsymbol{X}^j$ matrices represents a document with *N*-dimensional vector of words. Principal Component Analysis (PCA) is then performed on all *J* class matrices creating *J* sub-models. A class *j* sub-model consists of 2 matrices; the scores matrix $\boldsymbol{T}^j$ of size H^jx K^j, with K^j is the number of retained principal components for model *j* and the loadings matrix $\boldsymbol{P}^j$ of size *N* x K^j (Branden & Hubert, 2005). Principal components (PC's) that span actual variance in the data are retained while those spanning noise are discarded. Several methods are used to determine the number of components K^j to be retained for each class (Peres-Neto, Jackson, & Somers, 2005). In this study, principal components that span 98% of the variance in the data are retained.

For a class *j*, the scores matrix $\boldsymbol{T}^j$ and loadings matrix $\boldsymbol{P}^j$ are used to compute the class boundaries in terms of *Orthogonal Distance* (*OD*) and *Score Distance* (*SD*). Using class boundaries, a new document *x* is assigned to a particular class *j* through calculating its *Reduced Distance* from the boundaries of all classes in the model as shown in the following steps:

i. $\boldsymbol{OD}_x^j$ is computed by projecting document *x* on the PCA model of class *j*, and then calculating the sum of squared residuals ε_{nx}^2 of that projection as seen in equation 1.

$$\boldsymbol{OD}_x^j = \sqrt{\frac{\sum_{n=1}^{N} \varepsilon_{nx}^2}{(N - K^j)}} \tag{1}$$

ii. Then, the document's score distance $\boldsymbol{SD}_x^j$ is computed using its distance from the score space boundaries of class *j* sub-model as shown in equation 2. Where, $\boldsymbol{t}_k$ is the score of document *x* on principal component *k* and $\vartheta_{k,lim}^j$ is the scores limit of the principal component *k* (i.e. the maximum and minimum scores for that PC); while, $S_{\vartheta_k}^j$ is the standard deviation of scores of the training documents H^j used to build class *j* sub-model on that principal component.

$$\boldsymbol{SD}_x^j = \sum_{k=1}^{K^j} \Phi_k^2 \left(\boldsymbol{t}_k - \vartheta_{k,lim}^j\right)^2 \quad , where$$

$$\Phi_k = \frac{OD_x^j}{S_{\vartheta_k}^j} \tag{2}$$

iii. The overall distance $\boldsymbol{d}_x^j$, known as the *Reduced Distance*, of the document is computed by linearly combining its $\boldsymbol{OD}_x^j$ and $\boldsymbol{SD}_x^j$ as seen in equation 3 (Daszykowski, Kaczmarek, Stanimirova, Heyden, & Walczak, 2006).

$$\boldsymbol{d}_x^j = \sqrt{OD_x^{j^2} + SD_x^{j^2}} \tag{3}$$

iv. So, the document *x* is assigned to class *j* if its *F* statistic, calculated using equation 4, is less than the *F*-value calculated using $(H^j - K^j - 1), (N - K^j)$ degrees of freedom. In equation 4, $\boldsymbol{S}^2$ is computed using the training documents H^j used to build class *j* sub-model.

$$\boldsymbol{F} = \frac{\boldsymbol{d}_x^{j^2}}{\boldsymbol{S}^{j^2}} \quad where,$$

$$\boldsymbol{S}^2 = \sqrt{\frac{\sum_{h=1}^{H^j} \sum_{n}^{N} \boldsymbol{\varepsilon}_{hn}^2}{(H^j - k^j - 1)\,(N - k^j)}} \tag{4}$$

The above four steps are repeated for all the *J* classes of the model. As such, document *x* can be assigned to multiple classes if its *Reduced Distance* satisfies the F-test classification rule for those classes. On the other hand, if the classification rule was not satisfied for any of the classes in the model, then the document will be considered as "*Unidentified*".

3. ALTERNATIVE SIMCA

One of the most efficient modifications that has been introduced to classical SIMCA method was that introduced by the PLS Toolbox (Eigenvector Research, 2014). This modification can be referred to as *Alternative-SIMCA*. Despite the fact that *Alternative-SIMCA* is still depending on the linear combination of the Score Distance (*SD*) and the Orthogonal Distance (*OD*), but the way these distances are computed differs from that of classical SIMCA. In *Alternative-SIMCA, Mahalanobis distance* is used to compute SD as seen in equation 5 below.

$$SD_x^j = \sqrt{\boldsymbol{t}_x \, \Lambda^{-1} \, \boldsymbol{t}_x} \tag{5}$$

Where $\boldsymbol{t}_x$ is a vector that represents the scores of sample $\boldsymbol{x}$ on the K^j PC's of class *j* sub-model and Λ^{-1} is a diagonal matrix containing the inverse of the eigenvalues associated with the K^j PC's retained in the sub-model. On the other hand, the *OD* is computed using the sum of squared residuals as discussed before.

The class boundaries for the score distance $\boldsymbol{SD^j}$ of class *j* are computed using the *Hotellings* $\boldsymbol{T^2}$, as shown in equation 6. Where $\boldsymbol{T^{2j}_{lim}}$ is the boundary of class *j* PCA model and F_{K^j,H^j-K^j} is the *Mlinowski F-test* using $K^j, H^j - K^j$ degrees of freedom (Forina, Casale, Oliveri, & Lanteri, 2009).

$$\boldsymbol{T^{2j}_{lim}} = F_{K^j,H^j-K^j}\,\frac{K^j\,({H^j}^2 - 1)}{M^j\,(H^j - K^j)} \qquad (6)$$

While class limits for the orthogonal distance OD^j are computed using $\boldsymbol{\chi^2}$ distribution and referred to as $\boldsymbol{Q^j_{lim}}$ (Pomerantsev, 2008).

For classifying a new document *x*, ***SD*** and ***OD*** will be computed for that document against each class sub-modal. Then document *x* will be assigned to that class if its *Reduced Distance* is less than or equal to the square root of 2 as stated in equation 7.

$$\sqrt{\left(\frac{Q^j}{Q^j_{lim}}\right)^2 + \left(\frac{T^{2j}}{T^{2j}_{lim}}\right)^2} \quad \leq \quad \sqrt{2} \qquad (7)$$

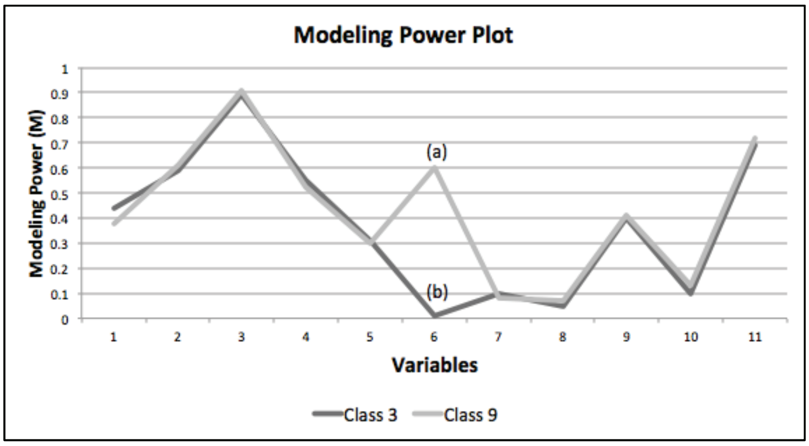
Despite the fact that SIMCA is a soft modeling technique where a single document can be assigned to multiple classes, but *Alternative-SIMCA* can be used for hard modeling as well. In hard modeling case, *Reduced Distance* of the document is computed against all the *j* classes of the model, and then the document will be assigned to the class that achieves the lowest *Reduced Distance* (Durante, Bro, & Cocchi, 2011).

4. VARIABLES STRENGTH COEFFICIENT (VSC)

One of the main challenges in text classification is the massive number of variables. In this paper, we propose a new variables selection technique that achieves substantial dimensionality reduction while maximizing the differences between different model classes, as well as preserving the modeling power of model's variables. This technique proposes a new coefficient called *Variable Strength Coefficient (VSC)*. *VSC* coefficient integrates the discriminatory power of a variable together with its ability to model the data in decomposition models (PCA models in case of two-way SIMCA) into a single term that assesses the overall modeling and discriminatory strength of a variable.

The variables selection and dimensionality reduction routines using this technique are based on calculating the *VSC* for all variables in the model. Then, variables with *VSC* exceeding a predefined threshold are retained, while other variables with lower *VSC* are discarded. This technique results in substantial reduction in the number of variables in a model, leading to a faster and easier implementation of decomposition algorithms (PCA, PARAFAC, etc.) that further reduce model dimensions. In addition the discriminatory and fitting powers of the model increase substantially.

Figure 1: Modeling Power of 11 variables for Classes 3 and 9. (a) Variable 6 has a Modeling Power of 0.6 for class 3. (b) The same variable has modeling power of 0 for class 9.



In classification analysis, Variable Strength Coefficient (VSC) is an overall measure of the strength of a variable in fitting the data, as well as discriminating classes in a model. In this paper we proposes modification to the original Modeling and Discriminatory powers of variables to create two new terms used to build VSC coefficient. As such, VSC linearly combines two terms; a *modified Modeling Power* (weighted power) of a variable for all classes; and a *modified Discriminatory Power* (weighted and normalized power) of the same variable for all classes as well. Different multi-way decomposition methods can be used to compute VSC depending on the nature of the problem. In case of SIMCA, *VSC* involves applying 2-way decomposition *Principal Components Analys*is (*PCA*) for computing the modeling and discriminatory power terms of the coefficient.

4.1. Modified modeling power

The original *Modeling Power* is used to assess the ability of a variable to model the data in a specific class in *PCA* based models. *Modeling Power* $\boldsymbol{M}_n^j$ for a variable *n* in class *j* is computed as shown in equation 8 (Brereton, 2003).

$$\boldsymbol{M}_n^j = 1 - \frac{\boldsymbol{S}_{nraw}^j}{\boldsymbol{S}_{nresid}^j} \qquad (8)$$

Where, $\boldsymbol{S}_{nraw}^j$ is the standard deviation of raw data along variable *n* in class *j*. While, $\boldsymbol{S}_{nresid}^j$ is the standard deviation of model residuals along the same variable in class *j*. The model residuals are the difference between the original data and the fitted data with a *Principal Component Analysis* (*PCA*) model of that class. Equation 9 shows how to calculate the residuals matrix $\boldsymbol{E^j}$ of size (H^j × *N*) of a PCA model. $\boldsymbol{T^j}$ is the scores matrix (H^j × K^j) of documents belonging to class *j*; while $\boldsymbol{P^j}$ is the loadings matrix (*N* × K^j) of class *j*.

$$\boldsymbol{E^j} = \boldsymbol{X^j} - \boldsymbol{T^j}.\boldsymbol{P^j} \qquad (9)$$

The modeling power computed in equation 8 above varies between 1 (excellent modeling power) and 0 (no modeling power). One of the drawbacks of applying this modeling power is that, one has to compute the modeling power of the same variable in each class and then find a way to decide which variables to be retained and which variables to be discarded, knowing that a specific variable may have strong modeling power for one class but weaker modeling power for another (see Figure 1).

To handle this dilemma, this paper introduces the *modified-Modeling Power* term of *VSC*, through weighing the variable's modeling powers in all classes according to the magnitude of contribution of each class to the model. *VSC* then linearly combines those powers into a single descriptive modeling term.

The *modified-Modeling Power* for a variable *n* is computed by applying a weighted sum of all *Modeling Powers* of that variable for all the *J* classes of the model. The applied weights are the proportion of the training documents used to model each class H^j to the total number of training documents *H* (Equation 10).

$$\boldsymbol{M}_n = \sum_{j=1}^{J} \boldsymbol{M}_n^j \cdot \frac{H^j}{H} \qquad (10)$$

4.2. Modified discriminatory power

The original *Discriminatory Power* computes the standard deviation of residuals for all samples (documents) of a specific class after being fitted to the model of another class, and then compares this standard deviation with those calculated for samples fitted to their own class model. Equation 11 computes the *discriminatory power* $\boldsymbol{D}_n^{j,j+1}$ for a variable *n* to discriminate two classes *j* and *j+1*. The higher the value of the *discriminatory power* the more efficient the variable is in discriminating between two specific classes (Brereton, 2003).

$$\boldsymbol{D}_{\boldsymbol{n}}^{\boldsymbol{j,j+1}} = \sqrt{\frac{j\ model\ (j+1)_{s_{nresid}}{}^2 + \ (j+1)\ model\ j_{s_{nresid}}{}^2}{j\ model\ j_{s_{nresid}}{}^2 + \ (j+1)\ model\ (j+1)_{s_{nresid}}{}^2}} \qquad (11)$$

As such, *Discriminatory Power* is the ratio of the sum of standard deviations of residuals computed when a model of class *j* is trying to fit samples belonging to class *j+1* and vice versa; to the sum of standard deviations of residuals computed for every class model fitting its own samples. Thus, in order to get a concrete idea on the discriminating ability of a variable, it is necessarily to compute the residuals for each class fitting the members of other classes and again trying to find a way to determine which variables to be retained or removed based on their different powers in discriminating between class pairs.

Again *VSC* solves these two problems through introducing its *modified-Discriminatory Power* term; this term acts by assessing, weighing, normalizing and combining all discriminatory powers of a specific variable for all class pairs. As such, *VSC* make it possible to determine the variables with sufficient discriminatory power in a straightforward manner. The *modified-Discriminatory Power* term of VSC is calculated in two steps:

i. The *overall Discriminatory Power* of a variable *n* for all classes *J* is calculated as seen in equation 12. It is the square root of the ratio of; the sum of standard deviation of residuals, along variable *n*, of all classes fitting members of other classes; to the sum of residuals, along the same variable *n*, of all classes fitting their own member samples.

$$\boldsymbol{D}_n = \sqrt{\frac{\sum_j^J \sum_c^J j\ model\ c\ {s_{nresid}}^2}{\sum_j^J j\ model\ j\ {s_{nresid}}^2}} \qquad (12)$$

ii. Then, normalize the overall *Discriminatory Power* of a variable *n* as seen in equation 13. Where, $min\{\boldsymbol{D}\}$ and $max\{\boldsymbol{D}\}$are the minimum and maximum discriminatory powers of all variables respectively. This transforms the *overall Discriminatory Power* to values between 0 and 1.

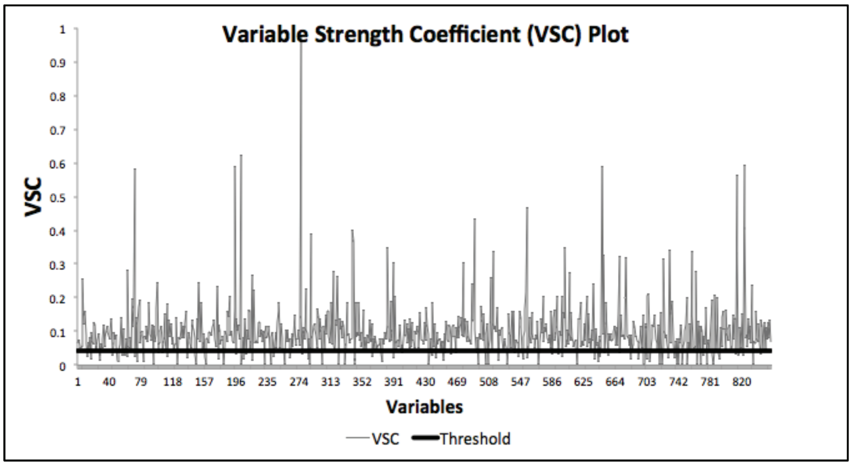
$$\boldsymbol{D}_n = \frac{\boldsymbol{D}_n - \min\{\boldsymbol{D}\}}{\max\{\boldsymbol{D}\} - \min\{\boldsymbol{D}\}} \qquad (13)$$

4.3. The coefficient

Then, *VSC* linearly combine the previously computed *modified-modeling power* $\boldsymbol{M}_n$ and *modified-discriminatory power* $\boldsymbol{D}_n$ terms as shown in equation 14, where the weight ***w*** should have a value in the range [0,2].

$$\boldsymbol{VSC}_n = \frac{\boldsymbol{w}\,.\boldsymbol{M}_n \ + \ (2-\boldsymbol{w})\,\boldsymbol{D}_n}{2} \qquad (14)$$

Figure 2: VSC calculated for 856 variables of 9-class model using weight w = 0.9 and strength threshold set at 0.041. Variables with VSC lower than the threshold line are discarded. Thus, a total of 122 variables were discarded achieving 14.25% dimensionality reduction while maximizing both modeling and discriminatory power of the whole classification model.



The weight ***w*** is adjusted to grant more influence to one of the terms of *VSC* depending on a priori knowledge of the data to be modeled. If no prior knowledge is available, then the best practice is to set the weight ***w*** to 1 to grant equal weights to both terms and then use cross-validation to set the final weights. In cases where large number of outliers is expected (highly noisy data), then one can lower the modeling power weight in favor of strengthening the discriminatory power and avoid overfitting. The resulted *VSC* ranges between 0 and 1.

The strength threshold is set at *half the standard deviation of VSC values* for all variables. As such, a variable with *VSC* greater than the strength threshold is considered to have high power in modeling data and discriminating classes. This straightforward manner enables a data analyst to simply decide the amount of strength a variable should have to be included in the model. Figure 2 shows the *VSC* calculated for 856 variables of a 9-class model using *strength threshold* set at 0.041 and *w* = 0.9).

5. EXPERIMENT

The aim of the experiment is to create a classification model for text documents that minimizes the number of variables while maximizing the model's classification performance. SIMCA model coupled with VSC technique is used to achieve that aim. VSC-SIMCA will result in substantial dimensionality reduction in number of variables that may reach 3-6% of the original numbers. Such reduction leads to significant reduction in the complexity of the model and alleviating the computational burden for classifying future samples.

A data set 'CNAE-9', provided by (Bache & Lichman, 2013), was used to create and test the model. The dataset contains 1080 documents of free text business descriptions of Brazilian companies categorized into a subset of 9 categories cataloged in a table called National Classification of Economic Activities (CNAE). The original texts were pre-processed to obtain the current data set as follows: i) only letters were kept; ii) then, prepositions were removed from the texts; iii) Next, the words were transformed to their canonical form; and iv) finally, each document was represented as a vector, where the weight of each word is its frequency in the document.

The data set composed of 856 variables (words) and 1080 documents of nine classes (120 document in each class). This data set is highly sparse (99.22% of the matrix is filled with zeros), so applying VSC is essential for improving the model's classification performance. The documents were divided into two data sets; *Training set* with 801 documents and *Testing set* with 379 documents.

The proposed variables selection technique VSC was applied on the training set. A deterministic optimization algorithm was performed to optimize VSC parameters. The optimal weight parameter (***w***) that achieved the best cross-validation results was found to be 0.9. A weight parameter less than 1 gives more weight to the ability of a variable to discriminate between classes over its ability to fit the data. As such, the discriminatory power term of VSC algorithm received a weight of 1.1, while its modeling power term got a weight of 0.9. The standard deviation of VSC values of the 856 variables was 0.0812, and thus, the strength threshold was set at 0.0406. Using strength threshold of 0.0406, 122 variables fell below the threshold and therefore they were discarded while the remaining 734 variables were retained (see figure 2). As such, the variables dimension was reduced by 14.25%.

After pre-processing the training set matrices by discarding variables with weak VSC and as a part of SIMCA modeling, *Principal Component Analysis* was performed for the nine classes. And due to the intrinsic characteristics of *Alternative SIMCA* method of being an independent modeling technique, 9 separate PCA sub-models were created for the nine classes; each class had different number of retained principal components. PCA further reduced the variables dimension form 734 to a range of 32 to 55 (each class sub-model had different number of retained PC's). Therefore, the PCA step resulted in further dimensionality reduction, where the final sub-models had number of variables equivalent to only 3.7 – 6.4% of the original variables. This means that VSC-PCA resulted in approximately 93.6-96.3% dimensionality reduction.

Then, the critical values, Hoteling (**T2**) and **Q**, for each class sub-model were computed; with confidence level of 0.95 and *F-test* level set to 0.95 as stated before. The resulted nine VSC-SIMCA sub-models were tested using the 379 documents of the testing set. And the results were compared to the results published in (Ciarelli & Oliveira, 2009) and (Ciarelli, Oliveira, & Salles, 2010). VSC-SIMCA achieved classification rate of 95.34%, i.e. 361 documents out of the 379 test documents were successfully assigned to their appropriate classes. And as seen in table 1, VSC-SIMCA shows superior performance over other techniques both in the magnitude of dimensionality reduction and in the high classification rate. Table 1 shows as well that SIMCA coupled with VSC (VSC-SIMCA) shows better performance over SIMCA alone, which indicates that the variables selection technique based on VSC coefficient was important to improve the model's selectivity, precision and robustness.

Table 1: Comparative study of the classification rates obtained after applying several variables selection, dimensionality reduction and classification techniques on CNAE dataset.

Technique	Number of Variables	Classification Rate (%)
VSC - SIMCA	*32 - 55*	*95.34*
SIMCA	*38 - 61*	*94.62*
KNN & aIB	250	91.11
KNN & sIB	200	92.22
KNN & MI-1	200 or 250	92.78
KNN & MI-2	250	20.00
KNN & LSI	100	92.78
KNN & IDF	250	27.78
KNN & IAE	250	91.10
IPNN		81.00
eMLP		84.26
IPNN-EM		10.16
ePNN$_1$		88.71
ePNN$_2$		84.45

6. CONCLUSION

This high dimensionality of text classification problem necessitates applying an efficient variables selection and dimensionality reduction technique that ensures model's selectivity, accuracy and robustness with fewer variables. This paper introduces a new method with variables selection coefficient that permits retention of variables with strong *Modeling* and *Discriminatory* abilities. The new coefficient, the *Variables Strength Coefficient* (*VSC*), linearly integrates two new terms. The first term, the modified-Discriminatory Power, evaluates the ability of a variable to discriminate between all classes in a model. The other term, the modified-Modeling Power, assesses the ability of that variable to fit the data of the model. *VSC* ranges between 0 and 1, where a variable with *VSC* greater than a predetermined threshold, is considered to have strong power in modeling data and discriminating classes. As such; strong variables, with *VSC* exceeding the strength threshold, are retained; while weaker variables are discarded. This straightforward technique results in maximizing the differences between different model classes, as well as preserving the modeling power of variables, in addition to a substantial reduction in the number of variables. The VSC technique was coupled with SIMCA supervised learning classification algorithm. *VSC-SIMCA* resulted in a robust and accurate model capable of classifying, successfully, text documents with substantial overall reduction in dimensionality, down to 3-6% of the original size.

VSC-SIMCA was applied on the data set ‘CNAE-9’, provided by “UCI Machine Learning Repository”, for classification of text documents belonging to 9 different classes. The results obtained are compared to classification and dimensionality reduction work done on the same data set in literature. *VSC-SIMCA* approach shows superior performance over other techniques, both in the amount of dimensionality reduction, as well as, the classification performance. *VSC-SIMCA* achieved 60% fewer variables than the best technique stated in literature applied on the same data set, with higher classification rate. This superior performance accounted to the efficient dimensionality reduction and variables selection techniques applied.

REFERENCE LIST

1. Bache, K., & Lichman, M. (2013). *UCI Machine Learning Repository*. (University of California, School of Information and Computer Science.) From http://archive.ics.uci.edu/ml
2. Branden, K., & Hubert, M. (2005). Robust classification in high dimensions based on the SIMCA method. *Chemom Intell Lab Syst , 79*, 10-21.
3. Brereton, R. (2003). *CHEMOMETRICS. DATA ANALYSIS FOR THE LABORATORY AND CHEMICAL PLANT.* Chichester: Wiley.
4. Ciarelli, P. M., & Oliveira, E. (2009). Agglomeration and Elimination of Terms for Dimensionality Reduction. *Ninth International Conference on Intelligent Systems Design and Applications*, 547-552.
5. Ciarelli, P. M., Oliveira, E., & Salles, E. O. (2010). An Evolving System Based on Probabilistic Neural Network. *Eleventh Brazilian Symposium on Neural Networks*, 182 - 187.
6. Daszykowski, M., Kaczmarek, K., Stanimirova, I., Heyden, Y. V., & Walczak, B. (2006). Robust SIMCA-bounding influence of outliers . *Chemometrics and Intelligent Laboratory Systems , 47*, 65–77 .
7. Durante, C., Bro, R., & Cocchi, M. (2011). A classification tool for N-way array based on SIMCA methodology. *Chemometrics and Intelligent Laboratory Systems , 106* (1), 73-85.
8. Eigenvector Research. (2014). PLS-Toolbox Manual 4.0 for MATLAB©.
9. Forina, M., Casale, M., Oliveri, P., & Lanteri, S. (2009). CAIMAN brothers: a family of powerful classification and class modeling techniques. *Chemometrics and Intelligent Laboratory Systems , 96*, 239–245.
10. Kemp, C., & Tenenbaum, J. B. (2008). The Discovery of Structural Form. *National Academy of Sciences*, *104*, 10687-10692.
11. Kim, H., Howland, P., & Park, H. (2005). Dimension Reduction in Text Classification with Support Vector Machines. *Machine Learning Research , 6*, 37-53.
12. Lin, Y., Jiang, J., & Lee, S. (2014). A Similarity Measure for Text Classification and Clustering. *IEEE Transactions on Knowledge and Data Engineering , 26* (7), 1575-1590.
13. Peres-Neto, P., Jackson, D., & Somers, K. (2005). How many principal components? stopping rules for determining the number of non-trivial axes revisited. *Computational Statistics & Data Analysis , 49* (4), 974–997.
14. Pomerantsev, A. (2008). Acceptance areas for multivariate classification derived by projection methods. *Chemometrics , 22*, 601–609.
15. Wold, S. (1976). Pattern recognition by means of disjoint principal components models. *Pattern Recognition , 8* (3), 127–139.
16. Wold, S., & Sjostrom, M. (1987). Letter to the editor — Comments on a recent evaluation of the SIMCA method. *Journal of Chemometrics , 1* (4), 243–245.
17. Zhao, D., He, J., & Liu, J. (2014). An improved LDA algorithm for text classification. *Institute of Electrical and Electronics Engineers* , 217-221.