On the Use of Supervised Learning Method for Authorship Attribution

Dr. Walaa M. Khalaf
Engineering College , University of Almustansiriya/ Baghdad
Email :walaakhalaf @yahoo.com

Received on: 4/4 /2011 & Accepted on: 3 /11 /2011

ABSTRACT

In this paper we investigate the use of a supervised learning method for the authorship attribution that is for the identification of the author of a text. We suggest a new, simple and efficient method, which is merely based on counting the number of repetitions of each alphabetic letter in the text, instead of using the traditional classification properties; such as the contents of the text and style of the author; which falls into four feature categories: lexical, syntactic, structural, and content-specific. Furthermore, we apply a spherical classification method.

We apply the proposed technique to the work of two Italian writers, Dante Alighieri and Brunetto Latini. With almost high reliability, the spherical classifier proved its ability to discriminate between the selected authors.

Finally the results are compared with those obtained by means of a standard Support Vector Machine classifier.

Keywords: Authorship Attribution, Spherical Classification, Support Vector Machine.

استخدام طريقة التعليم تحت الاشراف لأسناد التأليف

الخلاصه

في هذا البحث نقوم بالتحقيق في استخدام احدى طرق التعليم تحت الاشراف لتحديد كاتب نص معين. نحن نقترح طريقة جديدة، بسيطه، و كفوءة، تعتمد على احصاء عدد تكرار كل حرف ابجدي في النص بدلا من استخدام الخواص التقليدية المتبعه في التصنيف مثل محتويات النص و اسلوب الكاتب و التي تقع ضمن اربعة انواع للتصنيف وهي : المعجميه، النحويه، الهيكلية، ومحتويات محددة. علاوة على ذلك، نقوم بتطبيق طريقة التصنيف الكروي .

قمنا بتطبيق التقنيه المقترحة على اعمال كاتبين ايطاليين هما دانته و برونيتتو. مع وثوقيه عاليه اثبت المصنف الكروي المستخدم في هذا البحث قدرته على التمييز بين الكاتبين. اخيرا تمت مقارنة النتائج مع تلك المستحصلة من طريقة استخدام المصنف ذات المتجهات المدعمة.

INTRODUCTION

The process of identifying the authorship similarity of a given document, where a collection of documents with known authorship are given, is called Authorship Attribution [1]. In recent years, practical applications for authorship attribution have grown in areas such as intelligence (linking intercepted messages to each other and to known terrorists) [2], plagiarism detection (e.g. articles and essays) [3], criminal law (e.g., identifying writers of harassing messages, verifying the authenticity of suicide

https://doi.org/10.30684/etj.30.2.8
2412-0758/University of Technology-Iraq, Baghdad, Iraq

notes) [4], intrusion detection systems [5], as well as resolving historical questions of unclear or disputed authorship [6].

Authorship attribution can be considered as a method of finding the author of a text when there is a doubt about the original author. It is useful when two or more people claim to have written something or when no one is able to state that he wrote the piece.

Authorship attribution is a kind of classification or categorization problem. It is not a text classification, because in text classification the only used factor is the text content while both style of writing and text content are important in authorship attribution. The recent researches focused on different properties of texts. The most important properties used in classification are the content of the text and the style of the author [7]. Style of the author falls into four feature categories: lexical, syntactic, structural, and content-specific [2].

Lexical features can be either word or character-based. Word-based lexical features include such characteristics as total number of words, words per sentence, and word length distribution. Character-based lexical features include total number of characters, characters per sentence, characters per word, and the usage frequency of individual letters. Syntax refers to the patterns used to form sentences; this category of features consists of the tools used to structure sentences, such as punctuation and function words. Structural features deal with text organization and layout. Content-specific features are words that are important within a specific topic domain [2].

Scientific investigation into measuring style and authorship of texts goes back to the late nineteenth century, with the pioneering studies of Mendenhall [8] and Mascol [9, 10] on distributions of sentence and word lengths in works of literature and the gospels of the New Testament. By the mid-twentieth century, this line of research had grown into what became known as “stylometrics”, and a variety of textual statistics had been proposed to quantify textual style. Mosteller and Wallace [11] counted the use of words like ‘while’ and ‘upon’ to discriminate between possible authors. Binongo and Smith [12] used the frequency of occurrence of 25 prepositions to distinguish between Oscar Wilde’s plays and essays.

The usefulness of function words in Authorship attribution was examined by Argamon and Levitan [13]. The authors’ experiments with support vector machine classifiers (SVM) in twenty novels demonstrated that a classification success rates above 90% is achieved. They concluded that, using function words is a valid and good approach in authorship attribution. In 2001, Stamatatos, et al. [14] presented a fully-automated approach to the identification of the authorship of unrestricted text that excludes any lexical measure. They adapted a set of style markers to the analysis of the text performed by an already existing natural language processing tool using three stylometric levels, i.e., token-level, phrase-level, and analysis-level measures. Their experiments, presented on a Modern Greek newspaper corpus, showed that the proposed set of style markers is able to distinguish reliably the authors of a randomly-chosen group and performs better than a lexically-based approach. However, the combination of these two approaches provided the most accurate solution (i.e., 87% accuracy).

Koppel et. al. [15] presented convincing evidence of a difference in male and female writing styles in modern English books and articles, such a difference is sufficiently pronounced that it can be exploited for automated text classification; they showed that automated text categorization techniques can exploit combinations of simple lexical and syntactic features to infer the gender of the author of an unseen formal written document with approximately 80% accuracy.

In 2003, Joachim Diederich et. al. explored the use of text-mining methods for the identification of the author of a text. They applied the SVM to this problem, as it is able to cope with half a million of inputs since it requires no feature selection and can process the frequency vector of all words of a text. They performed a number of experiments with texts from a German newspaper. With nearly perfect reliability the SVM was able to reject other authors and detect the target author in 60–80% of the cases. In a second experiment, they ignored nouns, verbs and adjectives and replaced them by grammatical tags and bigrams; this resulted in slightly reduced performance. Author detection with SVMs on full word forms was remarkably robust even if the author wrote about different topics [16]. In [17], the effect of word sequences in authorship attribution is considered, taking into account both stylistic and topic features of texts, set of word sequences that combine functional and content words are used to identify the documents. The experiments are done on a dataset consisting of poems using naive Bayes classifier.

In this paper a new spherical separation algorithm with kernel transformations has been used for authorship attribution. Classification experiments have been applied on the works of the two famous Italian authors, Dante Alighieri and Brunetto Latini. Works of each author are considered as a different class type, this means that a binary classification problem must be solved. Spherical separation algorithm [18] deals with discrimination of two datasets by means of a sphere, once the center of the sphere is given. The Authorship Attribution method suggested in this paper does not depend on the lexical, syntactic, structural, or content-specific categories, but depends only on the frequency (repetition) of each alphabetic letter in each canto.

This paper is organized as follows; in section 2 the classification method that has been used in the experiments is explained. In section 3 explanations of the steps of the authorship attribution process is given. The classification results are shown in section 4, while some conclusions are given in section 5.

SPHERICAL CLASSIFICATION

The spherical separation algorithm [18] has been applied in this work. It considers a special case of the optimal separation, via a sphere, of two discrete point sets in a finite dimensional Euclidean space. In fact the center of the sphere is assumed fixed, in this case the problem is reduced to the minimization of a convex and nonsmooth function of just one variable, which can be solved by means of an “ad hoc” method in $O(p \log p)$ time, where p is the dataset size, as explained in the sequel.

A possible spherical separation of a set $\boldsymbol{A}$ from a set $\boldsymbol{B}$ consists in finding a minimal volume sphere enclosing all points of $\boldsymbol{A}$ and no points of $\boldsymbol{B}$ as shown in Figure 1,
where $\boldsymbol{A} = \{a_i, i = 1, \ldots, m\}$
and $\boldsymbol{B} = \{b_i, i = 1, \ldots, k\}$.

A sphere centered in x_0 with radius R is defined as

$$S(x_0,R)=\left\{\begin{matrix} x\in\Re^n \mid (x-x_0)^T(x-x_0) \\ \le R^2 \end{matrix}\right\} \qquad \ldots(1)$$

The sets ***A*** and ***B*** are defined to be spherically separated if there exists a sphere S(x_0,R) such that:
$\mathbf{(a_i - x_0)^T(a_i - x_0) \le R^2}$
for all points $\boldsymbol{a_i} \in \boldsymbol{A}$ **(*i* =1, . . . ,m)** and $\mathbf{(b_l - x_0)^T(b_l - x_0) \ge R^2}$
for all points $\boldsymbol{b_l} \in \boldsymbol{B}$ **(*l* =1, . . . , k)**.
The problem of minimizing both the volume of the sphere and the classification error is defined as follows

$$\begin{aligned}&\min_{x_0,R} R^2 + \\ &C\sum_{i=1}^{m}\max\left\{0,(a_i-x_0)^T(a_i-x_0)-R^2\right\}+ \\ &C\sum_{l=1}^{k}\max\left\{0,R^2-(b_l-x_0)^T(b_l-x_0)\right\}\end{aligned} \qquad \ldots(2)$$

where the positive constant *C* states the relative importance of the two objectives.

Once the center x_0 of the sphere is assumed known, by introducing the change of variable

$$z = R^2,\ z \ge 0 \qquad \ldots(3)$$

and by defining:

$$c_i = (a_i - x_0)^T(a_i - x_0) \ge 0\ \forall\ i = 1, \ldots ,m$$
$$d_l = (b_l - x_0)^T(b_l - x_0) \ge 0\ \forall\ l = 1, \ldots , k \qquad \ldots(4)$$

problem (2) becomes:

$$\min_{z\ge 0} z + C\begin{pmatrix}\sum_{i=1}^{m}\max\{0,c_i - z\}+ \\ \sum_{l=1}^{k}\max\{0,z-d_l\}\end{pmatrix} \qquad \ldots(5)$$

which is a convex, piecewise affine minimization problem in the scalar (nonnegative) variable *z*. Problem (5) can be restated in the form of a linear program as follows:

$$
\begin{aligned}
& f_p = \min_{z,x,m} z + C\left(\sum_{i=1}^{m} x_i + \sum_{l=1}^{k} m_l\right) \\
& \text{s.t.} \\
& z - c_i + x_i \geq 0 && \forall i = 1,\ldots,m \\
& d_l - z + m_l \geq 0 && \forall l = 1,\ldots,k && \ldots (6) \\
& z \geq 0 \\
& x_i \geq 0 && \forall i = 1,\ldots,m \\
& m_l \geq 0 && \forall l = 1,\ldots,k
\end{aligned}
$$

It can be shown, that problem (6) can be solved in O (p log p) time, where p= max(m, k). By using a simple cutoff algorithm once the points of the data sets have been sorted in terms of their distances from the given center. The algorithm description, together with the proof of its ability of finding the optimal solution are given in [14].

Once the optimal solution $(\bar{z}, \bar{x}, \bar{m})$ for (6) has been calculated, the optimal solution of problem (5) is also available. Such sphere can be utilized for classification purposes, in the sense that any new sample point $x \in \Re^n$ is classified according to the following rule:

x is attributed to the set A if $(x - x_0)^T (x - x_0) < \bar{z}$

x is attributed to the set B if $(x - x_0)^T (x - x_0) > \bar{z}$.

The point x remains unclassified whenever it is $(x - x_0)^T (x - x_0) = \bar{z}$. Kernel transformation of the type used in SVM can be easily embedded into the spherical separation approach. The dataset was mapped into a higher dimensional space (the feature space) and the two transformed sets were separated by means of one sphere in such space.

AUTHORSHIP ATTRIBUTION PROCESS

The present work is concerned with discriminating between two authors writing in a similar style. The source of the text used in this paper was the *Divine comedy* (La Divina Commedia) which is an epic poem written by Dante Alighieri between 1308 and 1321, and the *Little Treasure* (Il Tesoretto) by Brunetto Latini written in 1260-1266.

Divine Comedy is often considered the greatest literary work composed in the Italian language and a masterpiece of world literature. The *Divine Comedy* is composed of three canticas, Hell (Inferno), Purgatory (Purgatorio), and Paradise (Paradiso), each consisting of 33 cantos. An initial canto serves as an introduction to the poem and is generally considered to be part of the first cantica, bringing the total number of cantos to 100. In the other side, *Little Treasure* consists of 22 cantos.

LETTERS FREQUENCY STATISTICS

Text samples are described by the frequency of letters in each canto. First, all non-alphabetic characters, including spaces, are discarded, and capital letters are converted to small letters. The 100 cantos of Dante (*La Divina Commedia*) are considered as a dataset of class type ***A***, while the 22 cantos of Latini (*Il Tesoretto*) are considered as a dataset of class type ***B***.

The total number of letters in Class ***A*** is as follows: part one (*Paradiso*) consists of 132,329 alphabetic letters, part two (*Purgatorio*) consists of 133,369 alphabetic letters, and part three (*Inferno*) consists of 132,488 alphabetic letters, while class ***B*** (*Il Tesoretto*) consists of 53,569 alphabetic letters. The average length of each chapter in class ***A*** is 4000 letters, while the maximum length of chapters of class ***B*** is 6029 letters and the minimum is 448 letters. This means, there is a big difference between the maximum and minimum number of alphabetic letters in each cantos of Latini work (*Il Tesoretto*).

To apply any classification method we must have at least an equal number of samples (chapters) in each dataset group (class type) and each sample combines an approximately equal number of alphabetic letters. Our data belongs to 21 vector space (R^{21}) each space represents one alphabetic letter (Italian language consists of 21 alphabetic letters), each position in the resulted matrix contains a number represents the occurrence of that alphabetic letter in that chapter. This means, that we have a matrix of 100 rows (number of chapters) and 21 columns (number of alphabetic letters), the number in each column represents the occurrence of that letter in the desired chapter. For example, if there is a number '380' in row number '6' and column number '1', this means, the occurrence of letter 'A' (column 1) in chapter 6 is 380 times. The following procedure (as shown in Figure 2) has been used on Latini work (*Il Tesoretto*) to solve the problem mentioned above, that the two classes should have approximately equal number of samples and every sample contains approximately equal number of alphabetic letters.

At the end of this procedure will be resulted 100 samples each sample contains 4000 alphabetic letters generated randomly from the original file (*Tesoretto* file), we will calculate the occurrence of every alphabetic letter in each chapter.

Now we have two classes (datasets) with 100 samples for each dataset, and each one contains approximately 4000 alphabetic letters.

EXPERIMENTAL RESULTS

For all our experiments we used the method in [18], and we compared our results with Radial Basis Function (RBF) kernel with different widths (γ) which defines as follows:

$$K(x_i, x_j) = \exp(-\frac{\left\| x_i - x_j \right\|^2}{2g^2}) \qquad \ldots(7)$$

All experiments are performed using 10-fold cross-validation. This allows us to get a reliable indication of how well the learner will work when it is asked to make new predictions on the held-out test set. The data set is divided into ten subsets containing two fragments of equal size per author. Each one of the subsets is used as test set and the remaining subsets as a training set.

Table 1 shows the percentage of the classifier predictions that are actually correct as measured against the known classes of the test examples for different values of *C* parameter. Accuracy is measured using ten-fold cross-validation which is a standard technique used in Machine learning to evaluate classifiers. The best correctness value was 99.5% using spherical classification in comparison with 95.9% when SVM approach has been applied.

Table 2 shows the percentage of the classifier predictions using RBF kernel for different values of *C* parameter and different widths γ. Accuracy is measured also using ten-fold cross-validation. The best classification correctness value was 99.5% using spherical classification in comparison with 95.082% when SVM approach has been applied. The two Spherical Classification versions (polynomial and Radial Basis Function “RBF”) performed substantially better than the currently SVM conventional method.

CONCLUSIONS

In this paper an experimental test of information contained in the number of repetitions of each alphabetic letter in the text was presented. Such simple information has allowed discriminating, with high successful rate, between two Italian authors writing in a similar style: Dante Alighieri and Brunetto Latini. The *Divine comedy* (La Divina Commedia) and the *Little Treasure* (Il Tesoretto) have been chosen as the source of the used texts. The *Little Treasure* contains less number of cantos and each canto includes less number of letters compared with the *Divine Comedy* of Dante, for this reason we proposed a new procedure explained previously in this paper to make the *Little Treasure* cantos equals the *Divine Comedy* cantos as well as the number of letters in each canto.

Using this new proposed method which based only on counting the number of repetitions of each alphabetic letter in the text, has been reduced the complexities of finding the style of the author and also has been cancelled the dependency on the contents of the text. By calculating the repetitions of each alphabetic letter in the text as shown in the experimental results both in case of using the support vector machine (SVM) method or the proposed spherical classifier method gives better results than the methods based on classification properties; such as the content of the text and the style of the author; which falls into four main feature categories: lexical, syntactic, structural, and content-specific. Using spherical classifier gives better results than the traditional Support Vector Machine classifier, as well as the spherical classifier reduces the number of mathematical operations required to find the optimal classifier.

In this paper we propose a new method to give each writer its own finger print for his works without depending on the traditional procedures or methods to find the author style and searching the contents of his texts and then applying one of the classification or discrimination methods, for example, building neural network classification system or using SVM classifier or other classification method as mentioned earlier, our proposed method depends only on calculating the repetitions of each alphabetic letter inside author’s text, and then we apply a classifier, that proposed by [18], which has a very low computational cost, $O(p \log p)$ time, where *p* is the dataset size, these results appear comparable with those obtained by SVM method, where SVM solutions are obtained from solving quadratic programs (QP), so that the computational cost of an SVM approach is at least square of the number of training dataset size *p*. This fact would suggest considering this approach as one of the election tools to deal with very large datasets.

At the end of training phase we will get the optimal classifier, which has the ability to discriminate later among all the works of Dante and all the works of Brunetto. This method can be generalized to multiclass classification for more than two writers.

REFERENCES

[1]Zhao, Y. and Zobel, J., “Searching with Style: Authorship Attribution in Classic Literature”, The Thirtieth Australasian Computer Science Conference, Ballarat, Jan. 2007, pp.59-68.

[2]Abbasi, A. and Chen, H., “Applying Authorship Analysis to Extremist-Group Web Forum Messages”, IEEE Computer Society, 2005, pp. 67-75.

[3]Stein, B. and Eissen, S., “Intrinsic Plagiarism Analysis with Meta Learning”, In Proceedings of the SIGIR Workshop on Plagiarism Analysis,Authorship Attribution, and Near-Duplicate Detection, 2007. pp. 45-50.

[4]Chaski, C.E., “Who’s at the Keyboard? Authorship Attribution in Digital Evidence Investigations”, International Journal of Digital Evidence, 4(1), 2005.

[5]Orebaugh, A. and Allnutt, J.,”Classification of Instant Messaging Communications for Forensics Analysis”, International Journal of Forensic Computer Science, 1, 2009, pp. 22-28.

[6]Burrows, J.F., “Delta: A Measure of Stylistic Difference and a Guide to Likely Authorship”, Literary and Linguistic Computing, 17(3), 2002, pp. 267-287.

[7]Madigan, D. Genkin, A. Lewis, D. Argamon, S. Fradkin, D. and Ye, L., “Author Identification on the Large Scale”, *Proc. of the Meeting of the Classification Society of North America, 2005.*

[8]Mendenhall, T., “The Characteristic Curves of Composition”, *Science*, 1887, 11 March, pp. 237-246.

[9]Mascol, C., “Curves of Pauline and Pseudo-Pauline Style I”, *Unitarian Review,* 1888, 30:452460.

[10]Mascol, C., “Curves of Pauline and Pseudo-Pauline Style II”, *Unitarian Review*, 1888, 30:539546.

[11]Mosteller, F. and Wallace, D.L., “*Inference and Disputed Authorship: The Federalist*”, Addison-Wesley: Reading, MA, 1964.

[12]Binongo, J. and Smith, M., “A Bridge between Statistics and Literature: The Graphs of Oscar Wilde’s Literary Genres”, *J. Applied Statistics*, 1999, vol. 26, pp. 781–787.

[13]Argamon, S. Shlomo, L., “Measuring the Usefulness of Function Words for Authorship Attribution”, *Proceedings of ACH/ALLC Conference 2005* in Victoria, BC, Canada.

[14]Stamatatos, L., Fakotakis, N. and Kokkinakis, G., “Computer-Based Authorship Attribution without Lexical Measures”, Computers and Humanities, 2001, pp.193-214.

[15]Koppel, M., Argamon, S. and Shimoni, A., "Automatically Categorizing Written Texts by Author Gender", Literary and Linguistic Computing, 17(4), 2002, pp. 401-412.

[16]Joachim, D., Jörg, K., Edda, L. and Gerhard, P., "Authorship Attribution with Support Vector Machines", Applied Intelligence. 2003 pp.109-123.

[17]Coyotl-Morales, R., Villasenor-Pineda, L., Montesy-Gomez, M. and Rosso, P., "Authorship Attribution Using Word Sequences", In Proceedings of the Iberoamerican Congress on Pattern Recognition (CIARP), 2006, pp. 844–853.

[18]Astorino, A. and Gaudioso, M., "A Fixed-Center Spherical Separation Algorithm with Kernel Transformations for Classification Problems", Computer Management Science, 2009, Vol 6, pp. 357–372.

Table (1) Classification Correctness Rate Using Liner Kernel

C Value	%Classification Rate (Spherical Classification)	%Classification Rate (**SVM**)
10	99.5	95.9
5	99.5	**95.9**
2	**99.5**	95.1
1	98.5	95.1
0.5	97.5	92.6
0.1	92.5	81.9
0.05	85.0	81.9

Table (2) Classification Correctness Rate Using RBF Kernel

C Value	Gamma =0.01		Gamma = 0.1		Gamma =1.0	
	Spherical	SVM	Spherical	SVM	Spherical	SVM
0.1	83.0	81.967	92.5	81.967	92.5	81.967
0.5	88.0	81.967	97.5	81.967	97.5	90.984
1	90.0	81.967	98.5	79.508	98.5	94.262
10	**93.5**	79.508	**99.5**	95.082	**99.5**	95.082

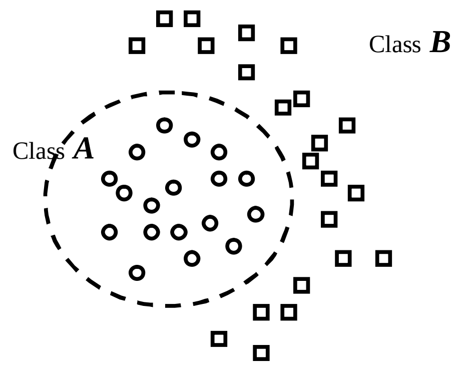

Figure (1). Spherical Classification.

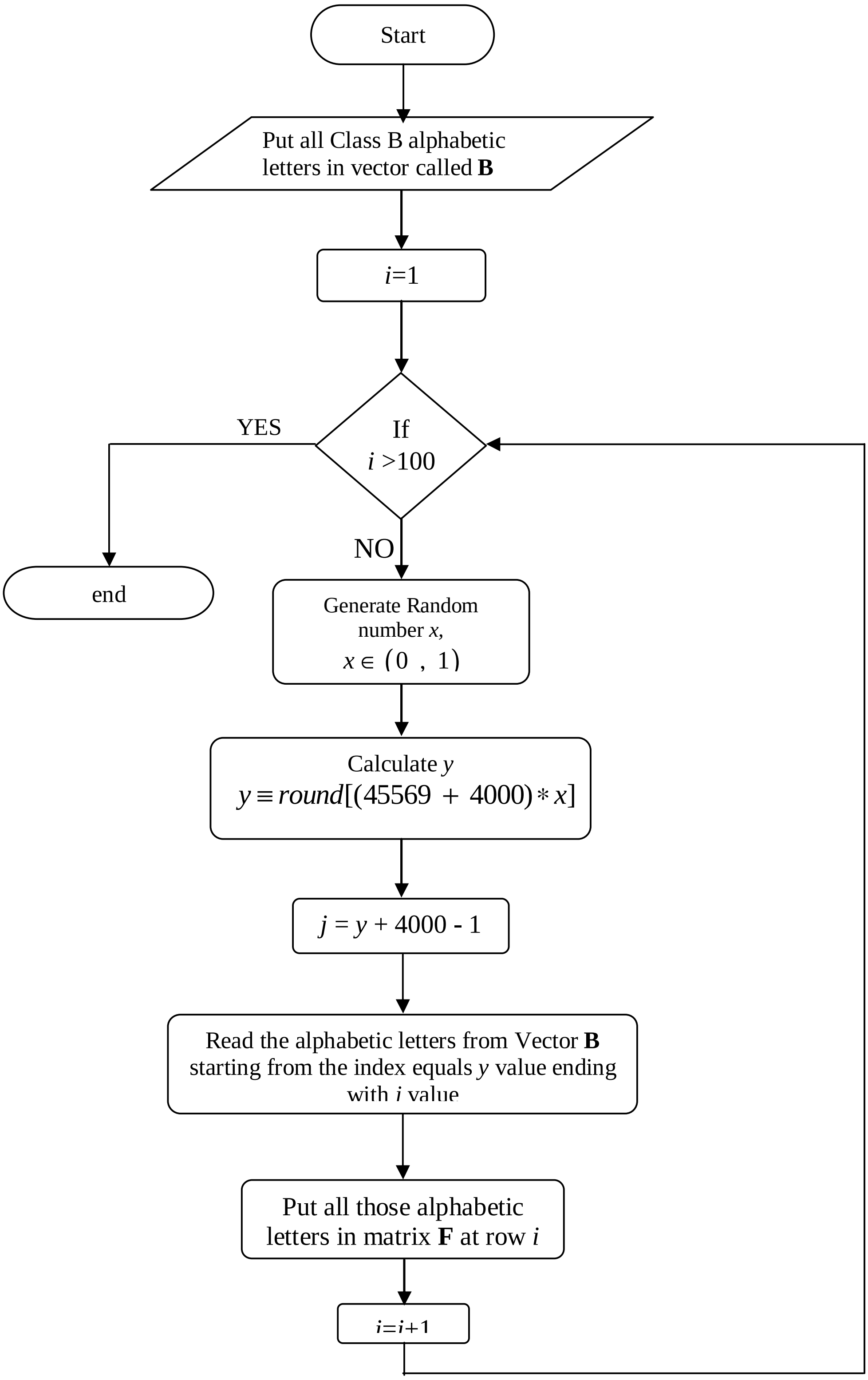
Start
Put all Class B alphabetic letters in vector called **B**
i=1
If *i* >100
YES
end
NO
Generate Random number *x*, $x \in (0\ ,\ 1)$
Calculate *y* $y \equiv round[(45569\ +\ 4000) * x]$
$j = y + 4000 - 1$
Read the alphabetic letters from Vector **B** starting from the index equals *y* value ending with *j* value
Put all those alphabetic letters in matrix **F** at row *i*
i=*i*+1

Figure (2):Procedurel flow chart