INFLUENCE OF DATA GEOMETRY IN RANDOM SUBSET FEATURE SELECTION

D. Lakshmi Padmaja[1], and Dr. B. Vishnuvardhan[2]

[1]Department of Information Technology, Anurag Group of Institutions (CVSR),
Hyderabad, India.

[2]Department of Computer Science and Engineering, JNTUH, Hyderabad, India.

ABSTRACT

The geometry of data, also known as probability distribution, is an important consideration for accurate computation of data mining tasks, such as pre-processing, classification and interpretation. The data geometry influences outcome and accuracy of the statistical analysis to a large extent. The current paper focuses on, understanding the influence of data geometry in the feature subset selection process using random forest algorithm. In practice, it is assumed that the data follows normal distribution and most of the time, it may not be true. The dimensionality reduction varies, due to change in the distribution of the data. A comparison is made using three standard distributions such as Triangular, Uniform and Normal Distribution. The results are discussed in this paper.

KEYWORDS

Data Geometry, Gaussian Distribution, Uniform Distribution, Triangular Distribution, Dimensionality Reduction, Random Forest, Random Subset Feature Selection.

1. INTRODUCTION

Dimensionality reduction is used as a pre-processing technique for selecting relevant, important features for further data mining task. This is highly useful in the field of speech processing, scientific data mining, cancer classification. Statistical analysis is widely used for interpreting the results. For accurate interpretation, data geometry plays an important role. For the last several years, the nature of the data sets is changed, and necessitated for novel approaches are in demand. The data sets resulted from scientific experiments, is often complex and high dimensional, meaning each object has many attributes or co-ordinates. While more data is beneficial in statistical analysis, this is often leading to curse of dimensionality. However various techniques are used to avoid the "curse of dimensionality" [1], including parametric and non-parametric models. Any wrong assumption about the data geometry results in poor accuracy.

In scientific data mining, the data sets are large and pose a challenge for selection of relevant features. In general, a normal (Gaussian) distribution is assumed for statistical analysis but, in reality datasets may not follow normal distribution. It leads to imprecise interpretation. In this paper, we have presented the performance variations due to change in data geometry, by using feature selection technique Random Subset Feature Selection (RSFS), which applied on 10 data sets and with 3 distributions namely normal, triangular and uniform. We present our approach in four sections; (1) Introduction (this section) (2) Data Geometry (3) Influence of data geometry in RSFS and (4) Conclusion.

DOI: 10.5121/ijdkp.2017.7403

2. Data Geometry

In this section, the selected distributions normal, triangular and uniform are described in detail.

2.1. Normal Distribution

The normal distribution is commonly used distribution, in the statistical analysis. This distribution is used in Natural Science, Social Science and Bioinformatics to represent real valued random variables, whose distributions are not known. Central Limit Theorem states that, when the samples are sufficiently large, the data samples average, of random variables, which are drawn from independent distributions, converges in to normal distribution.

$$f(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(\mathrm{x}-\mu)^2}{2\sigma^2}}$$

Where:
μ is mean or expectation of the distribution (and also its median and mode).
σ is standard deviation
σ^2 is variance

2.1.1. Properties of Normal Distribution:

The normal distribution is a real valued distribution whose mean, median and mode are equal [2]. This is symmetrical about mean; and not suitable for the variables which are inherently positive or negatively skewed. The variables such as weight of a person, share price may be better described using log normal or pareto distributions.

The normal distribution values approaches zero, when they are more than few standard deviations away from the mean. This distribution is not suitable for the process, when more outliers are expected. In such a scenario, the statistical inferences, such as standard error, least squares are unreliable. Hence heavily tailed distribution may be better suited for the analysis.

2.1.2. Tests for Normality:

Before assuming that, the data is normally distributed, the following tests are suggested for checking the normality of data. The normality test examines that, the likely hood of given data set (X_1...... X_n) confirms to normal distribution properties or not. For this Anderson, Darling Test is performed to check the normality of data (refer Figure (1)).

Also, Goodness of Fit Test is used for checking normality of the data. This test summarizes the difference between collected data and expected data from normal distribution. This test can also be used, whether the observed values follow specific distribution or not.

The above two tests are useful, if the data is real valued and continuous in nature. When the values are discrete, then the same tests are not relevant. For discrete data Pearson's Chi-square test [3], [4] need to be performed to assess the goodness of fit.

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

Where,
O_i= an observed frequency for bin I,
E_i= an expected frequency for bin i,
The expected frequency is calculated as

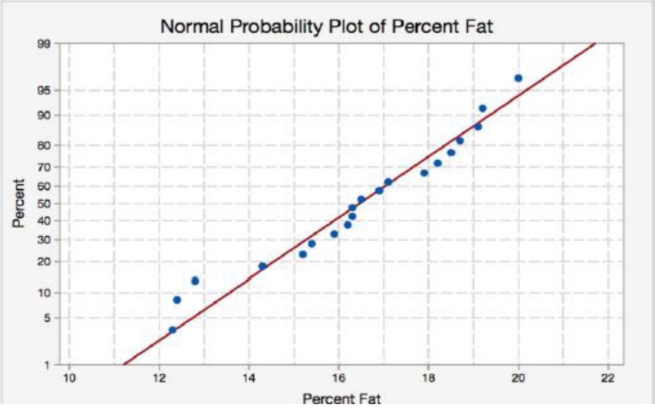
$$E_i = (F(Y_u) - F(Y_l))N$$

Where,
F = the cumulative distribution function.
Y_u = the upper limit for class i,
Y_l= the lower limit for class i,
N = the sample size
The resulting value can be used for assessing the goodness of fit.

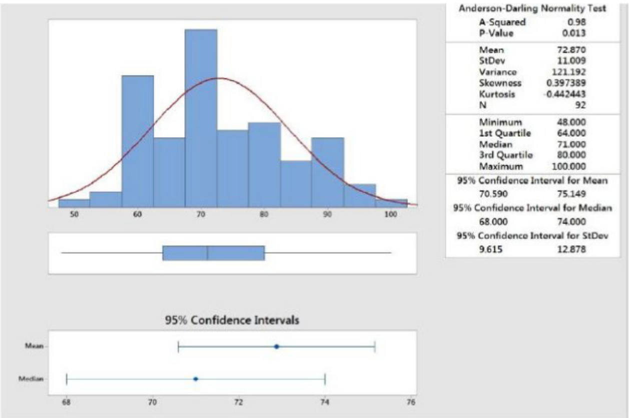
2.1.3 Examples

As an example, the normality tests and results on a data set, using Minitab® Figure(1a) and Figure(1b).

Figure 1: Normally distributed Data Property



(1a) Normal Probability Plot



```
Anderson-Darling Normality Test
    A-Squared           0.98
    P-Value            0.013

    Mean              72.870
    StDev             11.009
    Variance         121.192
    Skewness        0.397389
    Kurtosis       -0.442443
    N                     92

    Minimum           48.000
    1st Quartile      64.000
    Median            71.000
    3rd Quartile      80.000
    Maximum          100.000
95% Confidence Interval for Mean
    70.590            75.149
95% Confidence Interval for Median
    68.000            74.000
95% Confidence Interval for StDev
    9.615             12.878
```

(1b) Data is not Normal as P value is 0.013

2.2 TRIANGULAR DISTRIBUTION

The triangular distribution is used, when the data follows the pattern of optimistic, pessimistic and more likely values in estimation of outcomes. It is mainly used in business simulations, project management and audio dithering etc. The distribution is described, as a continuous probability distribution with lower limit a, upper limit b and mode c, where a ≤c ≤ b.
This distribution, as shown in Figure 2, is highly useful when the data availability is limited, due to high cost of data collection, and relationship between variables is known upfront. This distribution is dependent on *Minimum and Maximum* values of dataset. This is also known as *Lack of Knowledge* distribution.

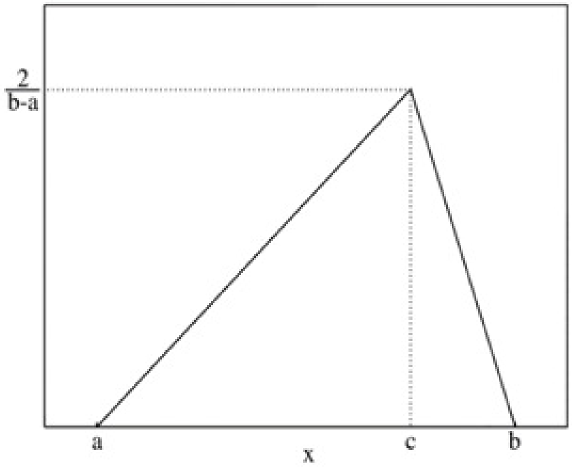


Figure 2: Triangular distribution with a, b and c , and c is Mode

2.3. UNIFORM DISTRIBUTION

In the uniform distribution [5] the events or data is evenly spaced. Due to nature of the data, the distribution is divided into discrete uniform distribution and continuous uniform distribution.

2.3.1 Discrete Uniform Distribution

This is also known as, Symmetric Distribution Function. The function is useful, when the events are known and can be observed equally likely as shown in the below Figure 3.

It is non-parametric and convenient to represent its values by integers in the interval
[a, b] Parameters:

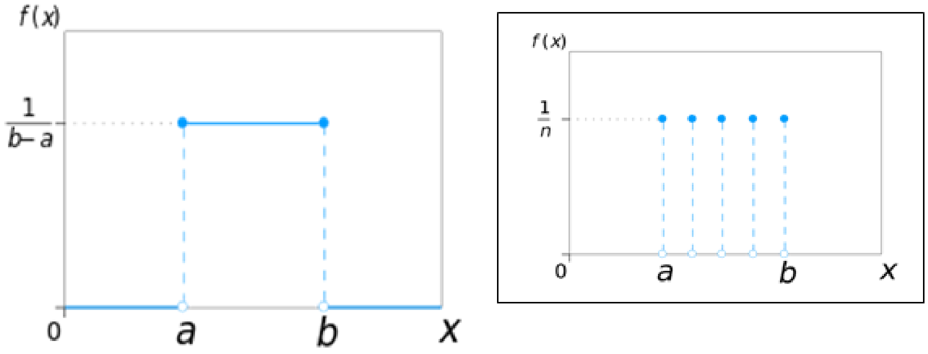
$a \in \{\ldots, -2, -1, 0, 1, 2, \ldots\}$
$b \in \{\ldots, -2, -1, 0, 1, 2, \ldots\}, b \geq a$
n=b-a+1 where n is a value

2.3.2 Example:

A well-known, Fair Die throwing is an example for discrete uniform distribution. The possible values are 1,2,3,4,5 and 6, and probability of a given score is 1/6. However, if more than one die is used, then the distribution is not uniform as all sums will not have equal probability.

2.3.3 Continuous Uniform Distribution Function:

Continuous Uniform Distribution is a symmetric probability distribution, where each member of the family has same length on the distribution support and equi-probable. Typical continuous and discrete uniform distribution are as shown in the following Figure(3a) and Figure(3b).



(a) Continuous Uniform Distribution

(b) Discrete Uniform Distribution

Figure 3: Uniform Distribution Types

3. RANDOM SUBSET FEATURE SELECTION (RSFS)

Feature selection methods are required for transforming high dimensional data set into low dimensional space without compromising the intrinsic properties of the data. Methods are divided into two categories known as subset feature selection and scoring methods [6]. RSFS is a feature subset selection method based on random forest. To understand the influence of the data geometry in RSFS, the data is subjected to pre-processing.

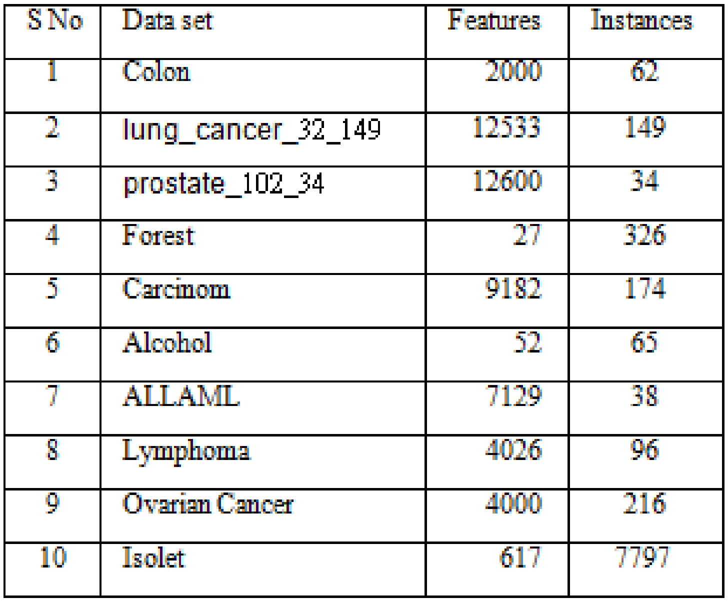
3.1 ABOUT RSFS:

Random Subset Feature Selection(RSFS), is an algorithm that is used for selecting relevant features from a large data set which is based on the Random Forest algorithm [7]. The relevant feature set is a reduced data set, which helps to improve the performance of classification task [8]. In RSFS, the features are selected by choosing a random subset from a feature set and then classifying with a kNN classifier. In each iteration, the relevance of each feature is computed and updated based on its performance. With more iterations, the quality of relevant feature selection improves gradually. In this algorithm, a set of dummy features are selected to create a random walk process and shape parameter. In each iteration, the relevancies are computed using the same process. A *normcdf ()* function is used as a transfer function to convert relevance matrix to probability matrix with threshold value greater than 99%. Most of the time, the distribution is assumed as normal distribution, but in reality, they are not. This results an erroneous classification. A detailed experiment was carried out, using 10 data sets (refer Table(I) with RSFS algorithm to demonstrate the influence of data geometry.

3.2 Random Subset Feature Selection Comparison With Traditional Methods:

In my previous paper [9] gives a comparative study of traditional feature subset selection techniques like Sequential Forward Selection (SFS), and Sequential Floating Forward Selection (SFFS) with RSFS. The SFS follows the procedure to find next best feature compared to the existing feature set as in forward selection(top down approach)[10], but it suffers with nesting problem, means; once a feature is retained, it cannot be discarded; Sequential Backward Selection(SBS), removing the worst feature as in backward selection(bottom up approach) suffers with more computation problem than SFS and nesting problem; SFFS is to avoid the problem of nesting of features by doing both(SFS & SBS) in a recursive manner [11]. Based on the experimental results, it was proved that the performance of the RSFS is better than the traditional methods.

Table I: Data set Description



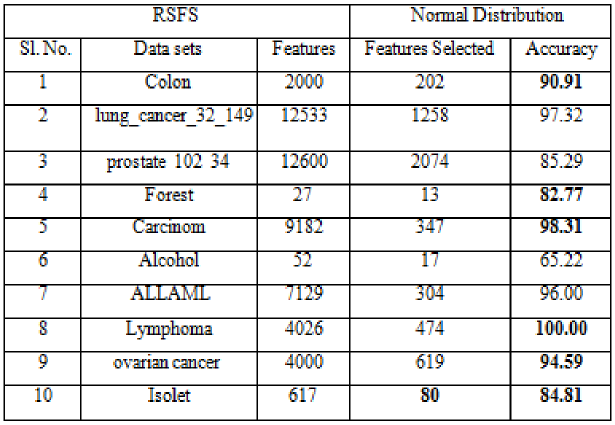
S No	Data set	Features	Instances
1	Colon	2000	62
2	lung_cancer_32_149	12533	149
3	prostate_102_34	12600	34
4	Forest	27	326
5	Carcinom	9182	174
6	Alcohol	52	65
7	ALLAML	7129	38
8	Lymphoma	4026	96
9	Ovarian Cancer	4000	216
10	Isolet	617	7797

3.2.1 Experimental Results:

The motivation for the experiment is, during the study of RSFS algorithm the on various data sets, the accuracy is varied widely, and the influence of data geometry on classification accuracy and data reduction is observed. In the research work carried out by Jouni Pohjalainena et al, [12], [8] the data geometry of features is assumed as uniform distribution. Whereas, we have observed that the change in the accuracy, due to the change in the distribution. We have carried out, the experiment to understand the influence of the same and results are discussed below.

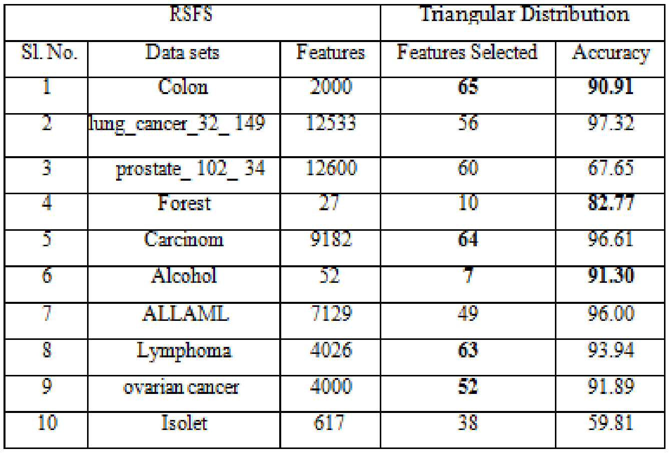
The data geometry is selected as uniform or triangular or normal distribution and results are tabulated, when applied on ten data sets. The results are shown below Tables II, III and IV:

Table II: RSFS using Normal Distribution



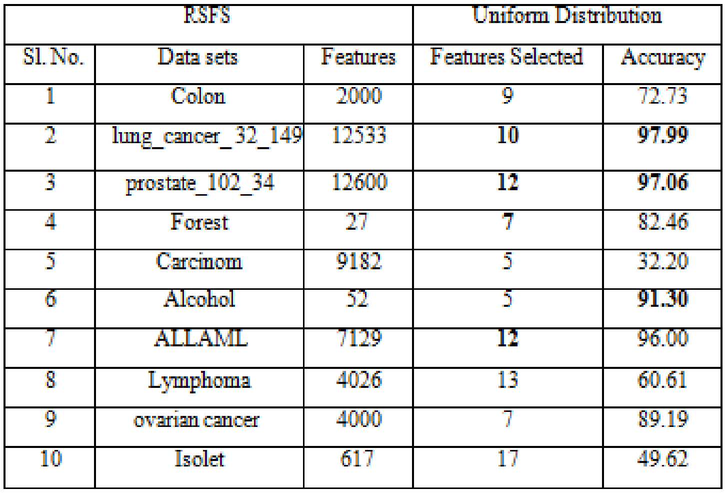
RSFS			Normal Distribution	
Sl. No.	Data sets	Features	Features Selected	Accuracy
1	Colon	2000	202	**90.91**
2	lung_cancer_32_149	12533	1258	97.32
3	prostate 102 34	12600	2074	85.29
4	Forest	27	13	**82.77**
5	Carcinom	9182	347	**98.31**
6	Alcohol	52	17	65.22
7	ALLAML	7129	304	96.00
8	Lymphoma	4026	474	**100.00**
9	ovarian cancer	4000	619	**94.59**
10	Isolet	617	**80**	**84.81**

Table III : RSFS using Triangular Distribution



RSFS			Triangular Distribution	
Sl. No.	Data sets	Features	Features Selected	Accuracy
1	Colon	2000	**65**	**90.91**
2	lung_cancer_32_ 149	12533	56	97.32
3	prostate_ 102_ 34	12600	60	67.65
4	Forest	27	10	**82.77**
5	Carcinom	9182	**64**	96.61
6	Alcohol	52	**7**	**91.30**
7	ALLAML	7129	49	96.00
8	Lymphoma	4026	**63**	93.94
9	ovarian cancer	4000	**52**	91.89
10	Isolet	617	38	59.81

Table IV: RSFS using Uniform Distribution



RSFS			Uniform Distribution	
Sl. No.	Data sets	Features	Features Selected	Accuracy
1	Colon	2000	9	72.73
2	lung_cancer_32_149	12533	**10**	**97.99**
3	prostate_102_34	12600	**12**	**97.06**
4	Forest	27	**7**	82.46
5	Carcinom	9182	5	32.20
6	Alcohol	52	5	**91.30**
7	ALLAML	7129	**12**	96.00
8	Lymphoma	4026	13	60.61
9	ovarian cancer	4000	7	89.19
10	Isolet	617	17	49.62

4. CONCLUSIONS

The scientific data sets are large, diverse, high dimensional and complex. In order to select relevant and useful features, novel algorithms are required and need to assume relevant geometry of the features under analysis. The algorithm to data is similar to the instruments to the physical world.

From the above experiments, it is understood that necessary caution must be taken about the geometry of the features. If we assume the geometry is normal distribution, relevant tests must be performed, as mentioned above, before proceeding analysis and interpretation of results. From the above tables, II, III and IV, the classification performance of Colon Forest, Carcinom Lymphoma and Isolet data sets accuracy is good, when normal distribution is selected; colon, Forest, and alcohol data sets accuracy is good, when triangular distribution is selected; and similarly, for lung-cancer-32-149, prostate-102-34, and alcohol data sets, classification performance is good, when uniform distribution is selected. Also, an observation of features selected in distributions, the total number of features are reduced, widely varied from one distribution to another. For example, Carcinom data set accuracy in normal and triangular distributions are more or less same, but more number of features are reduced in triangular distribution compared to normal distribution. Finally, it is evident that the data geometry influences in RSFS algorithm. It is established that, the influence of data geometry plays an important role, for better accuracy and dimensionality reduction. However, other parameters such as location and spread are also required in improving the performance. In future, a detailed study is required, to map the distribution vis-a-vis the data types are required for optimum performance of data mining tasks such as pre-processing and classification.

REFERENCES

[1] D. L. Donoho et al., "High-dimensional data analysis: The curses and blessings of dimensionality," AMS Math Challenges Lecture, vol. 1, p. 32, 2000.

[2] N. Altemose, K. H. Miga, M. Maggioni, and H. F. Willard, "Genomic characterization of large heterochromatic gaps in the human genome assembly," PLOS Comput Biol, vol. 10, no. 5, p. e1003628, 2014.

[3] H. Chernoff, E. Lehmann et al., "The use of maximum likelihood estimates in backslash chi square tests for goodness of fit," The Annals of Mathematical Statistics, vol. 25, no. 3, pp. 579–586, 1954.

[4] P. E. Greenwood and M. S. Nikulin, A guide to chi-squared testing. John Wiley & Sons, 1996, vol. 280. [5] G. Casella and R. L. Berger, Statistical inference. Duxbury Pacific Grove, CA, 2002, vol. 2.

[6] D. L. Padmaja and B. Vishnuvardhan, "Survey of dimensionality reduction and mining techniques on scientific data," International Journal of Computer Science & Engineering Technology, vol. 1, no. 5, pp. 1062–1066, 2014.

[7] Breiman, L, "Random forests," vol. 3, p. 5 32, 2001.

[8] J. Pohjalainen, O. Ra¨sa¨nen, and S. Kadioglu, "Feature selection methods and their combinations in high-dimensional classification of speaker likability, intelligibility and personality traits," Computer Speech & Language, vol. 29, no. 1, pp. 145–171, 2015.

[9] D. L. Padmaja and B. Vishnuvardhan, "Comparative study of feature subset selection methods for dimensionality reduction on scientific data," in Advanced Computing (IACC), 2016 IEEE 6th International Conference on. IEEE, 2016, pp. 31–34.

[10] A. W. Whitney, “A direct method of nonparametric measurement selection,” IEEE Transactions on Computers, vol. 100, no. 9, pp. 1100–1103, 1971. [11] P.Pudil, J.Novovicova, and J.Kittler, “Floating Search methods in feature selection,” vol. 1994, no. 15, pp. 1119 – 1125, Nov. 1994.

[12] O. Ra¨sa¨nen and J. Pohjalainen, “Random subset feature selection in automatic recognition of developmental disorders, affective states, and level of conflict from speech.” in INTERSPEECH, 2013, pp. 210–214.

AUTHORS

D. Lakshmi Padmaja is working as an Associate Professor, in the Department of Information Technology, Anurag Group of Institutions(CVSR), Hyderabad, India.Currently she is pursuing her Ph.D. in the field of Computer Science and Engineering from JNTU Hyderabad. She has completed her M. Tech. in the year 2000 from JNTUH.Her areas of interest are Data Mining, Data Analytics. She published her paper in IEEE.

Dr B. Vishnuvardhan is Headed Department of IT of JNTUH CE Nachupally between 2010 to 2014. He has completed his M. Tech from Birla Institute of Technology, Mesra, Ranchi in the year 2001 and completed his Ph. D from JNTUH in the year 2008. Having 19 years of teaching experience. His areas of interest are Linguistic processing, Data mining, Natural language processing, Information security and other elite fields of Engineering. He has Published papers in reputed Journals such as ACM Transactions on Asian Language and Information Processing (TALIP),Taylor and Francis based Journal of Information & Optimization Sciences WSEAS Transactions on Computers and Published 5 Inder Science Journals papers and other papers were published in IEEE, Springer, Elsevier and Scopus Indexed.