Identification of Valid Clusters for Datasets Whose Number of Clusters are Unknown

Gazala Yusufi, Smita Prava Mishra

Abstract*— The true use of clustering is not exploited properly as humans try to cluster datasets whose class labels are already known. In order to make best use of clustering, an attempt has been made in this work to find a mechanism to identify the number of clusters in the datasets whose class labels are unknown. The cluster validity techniques like Dunn's index, Davies-Bouldin index, Silhouette index, C index, Goodman-Kruskal index, etc. have been used to validate the number of clusters generated. These techniques access the clustering tendency and measure the quality of the clusters. These indexing techniques are used in conjunction with clustering algorithms like k-means, k-medoid, etc. to measure the validity of the clusters identified by the said algorithms depending on application specific data. The current work applies the above mentioned techniques to several classified datasets taken as benchmark as well as unclassified datasets so as to find the number of clusters in those datasets. Hence, suggests a better use of clustering.

***Index Terms*— Clustering, Cluster Validity Techniques, Indexing, Internal Cluster Validation, Unclassified Dataset Validation.**

I. INTRODUCTION

Many of the real life datasets do not have known class labels. Clustering is used to generate such labels after proper validation of the clusters. We require adequate domain knowledge to determine the number of clusters. The clusters identified by human beings using various clustering techniques may not always yield valid clusters because the number of clusters is to be specified by human. There is hardly any additional advantage of clustering the pre-classified datasets. Unfortunately, acceptable class labels for the data objects cannot be generated by applying clustering to an unclassified dataset. Cluster validation techniques come to the rescue by using some measures to check the validity of the clusters identified. Thus, clustering loses its significance if the number of class labels for a dataset is known.

This paper has been organized as follows: section 2 introduces some preliminaries required for discussion of the proposed technique. Section 3 discusses some literature studied in the current context. Section 4 gives description about the datasets used for experimental analysis. Section 5 proposes the model, algorithm and supporting explanation.

Manuscript received February, 2014.

Gazala Yusufi, Department of Computer Science & IT, Institute of Technical Education & Research, Siksha O Anusandhan University, Bhubaneswar, India.

Smita Prava Mishra, Department of Computer Science & IT, Institute of Technical Education & Research, Siksha O Anusandhan University, Bhubaneswar, India.

In section 6, experimental evaluation is given followed by section 7 which explains the result analysis. Section 8 concludes the paper and discusses future directions.

II. PRELIMINARIES

A. Clustering

Cluster analysis is about partitioning a given data set into groups (clusters) such that the data points in a cluster are more similar to each other than points in different clusters i.e. maximizing the intra-class similarity and minimizing the inter-class similarity [1]. There are many clustering algorithms like *k*-means, *k*-medoids, DBSCAN, BIRCH, etc. which can be used to generate the clusters. The problem with clustering is that most of the clustering techniques want the number of clusters to be given as input for performing clustering which may be unknown at times. Also the clusters identified by human beings may not always be valid enough to be acceptable for applications concerned.

B. Cluster Validation

Evaluation of clustering results is referred to as cluster validation [2, 3]. This can be done by using various techniques like *Dunn's index*, *Davies-Bouldin index*, *Silhouette's index*, *C index*, *GK index*, etc. which fall under the internal indexing techniques. There are also few external indexing techniques like *Jaccard index*, *Random index*, etc. which can be used for the same purpose but not on unclassified datasets as they do not have original class labels. The above indices measure the goodness of the clustering in comparison to other clusters created by the same or different clustering techniques using different parameters.

The main objectives of cluster validation are that it helps in finding the clustering tendency of a set of data; checks how well the results of a cluster analysis fit the data without reference to external information; compares the results of two different sets of cluster analyses and determines the '*correct*' number of clusters.

C. Internal and External Indexing

For measuring the clustering quality, the existing methods may be categorized into two types:

- **Internal cluster validation:** Measures the goodness of a clustering structure without the use of any external information. Applicable to both pre-classified and unclassified data.
- **External cluster validation**: Measures the extent to which cluster labels match externally supplied class labels. Applicable to pre-classified data.

Some internal cluster validation methods which have been implemented in this work are *Dunn's index*, *Davies-Bouldin index*, *Silhouette's index*, *C index* and *Goodman-Kruskal index*.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication Pvt. Ltd.

III. RELATED WORK

In the context of this study, there are many related works by various authors who have used the clustering and different indexing techniques for cluster validation in different ways in an application specific view.

Rendon E., *et al*. [2, 3], Bolshakova N., *et al*. [4, 5, 6], Ansari Z., *et al*. [7], Bezdek J., *et al*. [8, 9], Agarwal P., *et al*. [10], Maulik U., *et al*. [11], Vinh N., *et al*. [12], have studied various cluster validation and indexing techniques and made a comparison among those techniques. Studies have revealed that internal indexing techniques like *Dunn's index*, *Davies-Bouldin index*, *Silhouette index*, *C index*, *GK index*, etc., when applied on different clustering algorithms for different kinds of datasets, give more accurate validity measures for a given clustering structure.

Jaroszewicz S., *et al*. [13, 14], Taheri S.M., *et al*. [15], Hryniewicz O. [16], Beh E., *et al*. [17], Goodman L., *et al*. [18] have thoroughly analyzed the *GK indexing* technique for cluster validation and applied it to different areas of application. From obtaining smaller decision trees without sacrificing accuracy to being used as a measure of asymmetry for two-way contingency tables, GK index has been proved to have a wide range of applications. The fuzzy version of *GK* γ statistics has also been used for handling fuzzy ordered categorical data.

Bolshakova N., *et al*. [19], Yang C., *et al*. [20], have used *C index* and exploited its strength in various areas of applicability. *C index* has been very well used for assessing the cluster validity based on similarity knowledge and has proved to have given far better results in comparison to other techniques when applied to multi-source clustering.

Halkidi M., *et al*. [1], Jaccard P., *et al*. [21], Rand W.M., *et al*. [22], Frossyniotis D., *et al*. [23] in their papers, have presented the fundamental concepts of clustering while surveying the widely known clustering algorithms in a comparative way. Moreover, the papers address an important issue of clustering process regarding the quality assessment of the clustering results by clearly explaining few of the external indexing techniques in detail.

IV. PROPOSED MODEL

In order to find a better use of clustering, this work requires the validation of the clustering results obtained as a result of application of the clustering algorithms on unclassified data sets. In this regard, we first apply indexing techniques to classified datasets to generate standard minimum-maximum index value range for each of the indexing techniques. Then the input to the system is an un-clustered dataset without class labels. Application of appropriate clustering algorithms like *k*-means and *k*-medoid will result in a clustered dataset. Subsequently, the application of appropriate indexing techniques will generate the index values which will be compared to the standard minimum-maximum index value range and help in validating the number of clusters generated. Finally, the validity of the clusters will be established and the results are to be recorded and compared.

The problem can be better explained with the use of a proposed model designed for better understanding of the concepts and techniques that are applied in this work.

The model in figure 1 suggests the steps to be followed for validating the clustering results of an unclassified dataset.

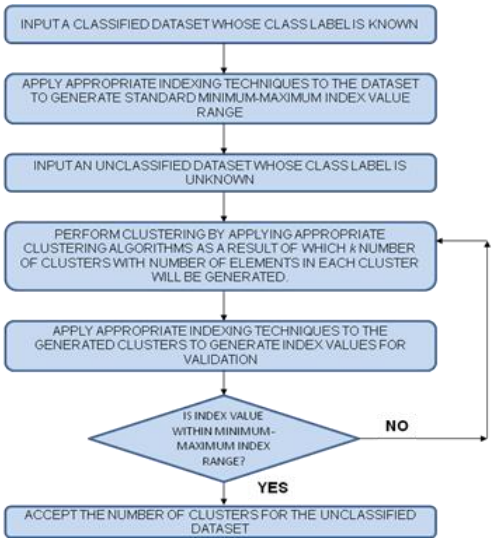

Figure 1: Proposed Model

***Step 1*: Input a set of classified datasets with known class labels to the system:**
The datasets with known class labels are taken as input to the system. Iris dataset, SPECTF Heart dataset, Wine dataset, Connectionist Bench (Sonar, Mines vs. Rocks) dataset and Teaching Assistant Evaluation dataset are five such datasets whose class labels are appropriately numbered and are taken as input to the system under study.

***Step 2*: Apply appropriate indexing techniques to the dataset to generate standard minimum-maximum index value range:**
On applying indexing techniques like *Dunn's index*, *Davies-Bouldin index*, *Silhouette index*, *C index* and *Goodman-Kruskal index* to the above mentioned datasets standard minimum-maximum index values are generated.

***Step 3*: Input a set of un-clustered datasets without class labels to the system:**
The datasets without known class labels are taken as input to the system. Concrete Compressive Strength dataset and Concrete Slump Test data set are two such unclassified datasets those are taken as input to the system under study.

***Step 4*: Apply appropriate clustering algorithms as a result of which *k* number of clusters with the number of elements in each cluster will be generated:**
In this step we apply the clustering algorithms like *k*-means and *k*-medoids to the above given unclassified datasets. Several iterations of the algorithm on each dataset will generate the number of elements in *k* number of clusters for different values of *k*.

***Step 5*: Apply appropriate indexing techniques to generate index values for validation of the generated clusters:**
Thereafter, we apply indexing techniques like *Dunn's index, Davies-Bouldin index, Silhouette index, C index* and *Goodman-Kruskal index* to the clustered dataset obtained in ***Step 4*** to generate index values so as to validate the clustering thus generated.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication Pvt. Ltd.

***Step 6*: Check whether the index values fall within minimum-maximum index range:**
In this step, we check and decide upon the cluster validity for a given dataset obtained as a result of application of clustering algorithms and indexing techniques. If the index value for a particular unclassified dataset is within the standard minimum-maximum index value range obtained in ***Step 2*** then we proceed to ***Step 7*** otherwise we repeat ***Step 4*** to ***Step 7*** by applying other appropriate clustering techniques as the present clustering technique has failed.

***Step 7*: Accept the number of clusters for the unclassified datasets:**
Finally, we accept the number of clusters generated for the unclassified datasets.

V. DATASET DESCRIPTION

In this paper, few classified and unclassified datasets from various domains have been used to evaluate and compare the results of the clustering and indexing techniques used on them. Their description is shown in the table 1 and table 2:

Table 1: Classified Dataset Description

Datasets	No. of Rows	No. of Columns	No. of Classes
Iris	150	4	3
SPECTF Heart	267	44	2
Wine	178	13	3
Connectionist Bench (Sonar, Mines vs. Rocks)	208	60	2
Teaching Assistant Evaluation	151	5	3

Table 2: Unclassified Dataset Description

Datasets	No. of Rows	No. of Columns	No. of Classes
Concrete Compressive Strength	1030	9	0
Concrete Slump Test	103	10	0

VI. EXPERIMENTATION

The experimentation involved in this work starts with data collection. Classified and unclassified datasets [28] of varied domains were first collected from different sources. Indexing techniques were applied to the classified datasets to generate standard minimum-maximum index value range on the basis of which the cluster validity of the unclassified datasets will be measured. After implementation of clustering algorithms i.e. k-means and k-medoids, the unclassified datasets, D and the number of clusters desired, n were supplied as input to the system. Each unclassified dataset, D then goes through several iterations to generate the number of elements in the n clusters. The most frequently occurring value over twenty executions is recorded as the best value for n numbers of clusters. This process is repeated by changing (increasing) the value of n for each unclassified dataset, D until any one of the clusters contains no elements. Here we stop further executions because we have to test the validity of the clusters generated for each n below this value.
This technique is then repeated and the same procedure is followed for all the unclassified datasets. Table 3 and table 4 represent the findings of k-means and k-medoid respectively on two unclassified data sets. This method has been adopted to standardize the techniques to work with unknown data.

Table 3: Findings of k-means for unclassified datasets

Indexing Techniques	Classified Datasets				
	I	SH	W	S	T
Dunn's Index	0.1636	0.7261	0.0920	0.2201	0.1709
Davies-Bouldin Index	9.0887	1.7546	6.4598	8.3136	9.6186
Silhouette Index	-0.5129	0.0658	-0.2105	-0.0393	0.0003
C Index	-Inf*	0.6894	-0.0200	0.4010	0.4539
Goodman-Kruskal Index	0.8838	-0.1767	0.5401	0.0751	0.0492

Table 4: Findings of k-medoid for unclassified datasets

Data sets	Number of Clusters	Number of elements in each cluster			
		C1	C2	C3	C4
Concrete Compressive Strength	2	601	429	NA	NA
Concrete Slump Test	3	47	13	43	NA

VII. RESULT ANALYSIS

On application of internal indexing techniques i.e. Dunn's index, Davies-Bouldin index, Silhouette index, C index and Goodman-Kruskal index to the classified datasets i.e. Iris dataset, Wine dataset, Connectionist Bench (Sonar, Mines vs. Rocks) dataset and Teaching Assistant Evaluation dataset, index values are generated for each dataset as shown in table 5, using which we can find the standard minimum-maximum index value range for each indexing technique as shown in table 6. Table 6 will help us to establish the validity of the clusters for the unclassified datasets.

Table 5: Index values for classified datasets.

Data sets	Number of Clusters	Number of elements in each cluster			
		C1	C2	C3	C4
Concrete Compressive Strength	2	468	562	NA	NA
Concrete Slump Test	3	30	48	25	NA

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication Pvt. Ltd.

(I=Iris dataset, SH=SPECTF Heart dataset, W=Wine dataset, S=Connectionist Bench (Sonar, Mines vs. Rocks) dataset, T=Teaching Assistant Evaluation dataset)

* -Inf is a very small value. As Iris dataset is the most appropriately classified dataset and small values of C index indicate a good clustering, so C index gives a very small value for Iris dataset.

Table 6: Minimum-Maximum Index value table.

Indexing Techniques	Minimum Index Value	Maximum Index Value
Dunn's Index	0.0920	0.2201
Davies-Bouldin Index	1.7546	9.6186
Silhouette Index	-0.5129	0.0658
C Index	-Inf*	0.6894
Goodman-Kruskal Index	-0.1767	0.8838

As a result of application of clustering algorithms i.e. k-means and k-medoids we obtained the number of elements in each cluster providing the number of clusters, n as input. Clusters are then validated by applying the internal indexing techniques i.e. Dunn's index, Davies-Bouldin index, Silhouette index, C index and Goodman-Kruskal index to the clusters generated for the above mentioned datasets. These techniques help in establishing the validity of the clusters obtained. Each of the internal indexing techniques will generate an index value for each dataset. These values will be compared and analyzed to find whether they fall within the minimum-maximum index value range as shown in table 6 to find the best index value for a given dataset which finally results in a better clustering. Table 7 compares the results of index values for the unclassified datasets obtained using k-means and k-medoids.

Table 7: Compared results of index values using k-means and k-medoid for unclassified datasets.

Indexing Techniques	Unclassified Datasets	
	CCS	CST
$D_{k\text{-means}}$ Index	1.0931	1.4156
$D_{k\text{-medoid}}$ Index	0.0041	0.3565
$DB_{k\text{-means}}$ Index	1.7393	1.3255
$DB_{k\text{-medoid}}$ Index	5.1043	4.5388
$SIL_{k\text{-means}}$ Index	-0.2254	-0.3290
$SIL_{k\text{-medoid}}$ Index	-0.1501	-0.1169
$C_{k\text{-means}}$ Index	0.0045	-0.5054
$C_{k\text{-medoid}}$ Index	0.1338	0.0331
$GK_{k\text{-means}}$ Index	0.4952	0.7358
$GK_{k\text{-medoid}}$ Index	0.5892	0.3902

($D_{k\text{-means}}$ Index=Dunn's index for *k*-means, $DB_{k\text{-means}}$ Index=Davies-Bouldin index for *k*-means, $SIL_{k\text{-means}}$= Index=Silhouette index for *k*-means, $C_{k\text{-means}}$ Index= C index for *k*-means, $GK_{k\text{-means}}$ Index= Goodman-Kruskal index for *k*-means. $D_{k\text{-medoid}}$ Index=*Dunn's index* for *k*-medoid, $DB_{k\text{-medoid}}$ Index=*Davies-Bouldin index* for *k*-medoid, $SIL_{k\text{-medoid}}$= Index=*Silhouette index* for *k*-medoid, $C_{k\text{-medoid}}$ Index= *C index* for *k*-medoid, $GK_{k\text{-medoid}}$ Index= *Goodman-Kruskal index* for *k*-medoid, CCS=Concrete Compressive Strength dataset, CST=Concrete Slump Test dataset)

From table 7, we can check the validity of the clustering obtained for the two unclassified datasets (Concrete Compressive Strength and Concrete Slump Test). If we compare the $D_{k\text{-means}}$ and $D_{k\text{-medoid}}$ for Concrete Compressive Strength (CCS) and Concrete Slump Test (CST) data sets, we find that the best value of *Dunn's index* for the two data sets which fall within the minimum-maximum range as shown in table 6 is 1.0931 and 1.4156 respectively (by using *k*-means). So, the unclassified data sets, CCS and CST, clustered using *k*-means generate two and three number of valid clusters respectively.

Thus, the results converge to the fact that the two unclassified datasets, CCS and CST were classified with two and three number of valid clusters respectively by using a combination of the above mentioned techniques.

VIII. CONCLUSION AND FUTURE DIRECTIONS

This work has helped us find a way to exploit the true use of clustering by applying clustering algorithms on unclassified and un-clustered datasets to generate clusters for them as well as validating the clusters generated by the application of indexing techniques. Hence, this work suggests a better use of clustering. Finally, using the above explained experimental set up the number of clusters i.e. two clusters for Concrete Compressive Strength data set and three clusters for Concrete Slump Test data set, could be successfully identified using a combination of the clustering and indexing techniques and the validity of the clusters obtained for them could be established. The clustering techniques used in this work can be replaced by better ones to generate much better results for all kinds of data sets. In future, this work may be used in medical, biomedical, physical & other areas of research and may be enhanced to get much better and valid clusters. The same facts can be cross-examined using external indexing techniques like *Jaccard index*, *Rand index*, etc.

ACKNOWLEDGMENT

Our heartfelt and sincere thanks goes to God, our beloved parents, brothers, sisters, family, friends and well-wishers.

REFERENCES

[1] Halkidi M., Batistakis Y., Vazirgiannis M. (2001). On clustering validation techniques, Journal of Intelligent Information Systems, volume 17(2), pages 107–145.

[2] Rendon, E., Abundez, I. M., Zagal, C. G., Arizmendi, A., Quroz, E.M., Arzate, E. (2011). A comparison of internal and external cluster validation indexes. AMERICAN-MATH'11/CEA'11 Proceedings of the 2011 American conference on applied mathematics and the 5th WSEAS international conference on Computer engineering and applications, pages 158-163.

[3] Rendon, E., Abundez, I. M., Arizmendi, A., Quroz, E. M. (2011). Internal versus external cluster validation indices. International Journal of Computers and Communications, volume 5 (1), pages 27-34.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication Pvt. Ltd.

[4] Bolshakova, N., Azuaje, F. (2006). Estimating the Number of Clusters in DNA Microarray Data. Medline Journal for biomedical articles, volume 45(2), pages 153-157.
[5] Bolshakova, N., Azuaje, F., Cunningham, P. (2005). An integrated tool for microarray data clustering and cluster validity assessment. Bioinformatics, volume 21(4), pages 451-455.
[6] Bolshakova, N., Azuaje, F. (2003). Cluster validation techniques for genome expression data. Signal Processing, volume 83 (4), pages 825 – 833.
[7] Ansari, Z., Babu, A. V., Azeem, M. F., Ahmed, W. (2011). Quantitative Evaluation of Performance and Validity Indices for Clustering the Web Navigational Sessions. World of Computer Science and Information Technology Journal (WCSIT), volume 1 (5), pages 217-226.
[8] Bezdek, J., Pal, N. (1995). Cluster Validation with Generalized Dunn's indices. ANNES '95 Proceedings of the 2nd New Zealand Two-Stream International Conference on Artificial Neural Networks and Expert Systems, pages 190-193.
[9] Bezdek, J., Pal, N. (1998). Some new indexes of cluster validity. IEEE Transactions on System, Man and Cybernetics-Part B, volume 28 (3), pages 301-315.
[10] Agarwal, P., Alam, M. A., Biswas, R. (2011). Issues, Challenges & Tools of Clustering Algorithms. International Journal of Computer Science Issues (IJCSI), volume 8 (3), pages 523-528.
[11] Maulik, U., Bandhopadhyay, S. (2002). Performance evaluation of some clustering algorithms and validity indices. IEEE Transactions on Pattern Analysis and Machine Intelligence, volume 24 (12), pages 1650-1654.
[12] Vinh, N., Epps, J., Bailey, J. (2009). Information Theoretic measures for cluster comparison: Is a correction for chance necessary?. Proceedings of the 26th International Conference on Machine Learning, Montreal, Canada, June 14th -18th.
[13] Jaroszewicz, S., Simovici, D., Kuo, W., Ohno, L. (2004). The Goodman–Kruskal Coefficient and Its Applications in Genetic Diagnosis of Cancer. IEEE Transactions on Biomedical Engineering, volume 51 (7), pages 1095-1102.
[14] Simovici, D., Jaroszewicz, S. (2004). A metric approach to building decision trees based on Goodman-Kruskal association index. Advances in Knowledge Discovery and Data Mining: 8th Pacific-Asia Conference, PAKDD 2004, Sydney, Australia, May 2004 Proceedings, volume 8, pages 181-191.
[15] Taheri, S.M., Hesamian, G. (2011). Goodman–Kruskal Measure of Association for Fuzzy-Categorized Variables. Kybernetika, volume 47 (1), pages 110–122.
[16] Hryniewicz, O. (2006). Goodman-Kruskal γ measure of dependence for fuzzy ordered categorical data. Journal of Computational Statistics & Data Analysis, pages 323-334.
[17] Beh, E., Simonetta, B., D'Ambra, L. (2007). Partitioning a non-symmetric measure of association for three-way contingency tables. Journal of Multivariate Analysis 98, pages 1391-1411.
[18] Goodman, L., Kruskal, W. (1954). Measures of associations for cross-validations. Journal of the American Statistical Association, volume 49, pages 732-764.
[19] Bolshakova, N., Azuaje, F., Cunningham, P. (2005). A Knowledge-driven approach to cluster validity assessment. Journal of Bioinformatics, volume 21 (10), pages 2546-2547.
[20] Yang, C., Zing, E., Li, T., Narsimhan, G. (2005). A knowledge-driven method to evaluate multi-source clustering. Proceeding ISPA'05 Proceedings of the 2005 international conference on Parallel and Distributed Processing and Applications, pages 196-202.
[21] Jaccard, P. (1912). The distribution of flora in the alpine zone. New Phytologist, volume 11, pages 37–50.
[22] Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. Journal of the American Statistical Association, volume 66 (336), pages 846-850.
[23] Frossyniotis D., Likas A., Stafylopatis A., (2004). A clustering method based on boosting, Pattern Recognition Letters 25, pages 641-654.
[24] Dunn, J.C. (1974). Well separated clusters and optimal fuzzy partitions. Journal of Cybernetics, volume 4 (1), pages 95-104.
[25] Davies, D. L., Bouldin, D. W. (1979). A cluster separation measure. IEEE Transactions on Pattern Analysis and Machine Intelligence, volume PAMI-1 (4), pages 224-227.
[26] Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, volume 20, pages 53-65.
[27] Hubert L., Schultz J. (1976). Quadratic assignment as a general data-analysis strategy. British Journal of Mathematical and Statistical Psychologie, volume 29, pages 90-241.
[28] http://archive.ics.uci.edu/ml/datasets.html

Gazala Yusufi has received her BSc. (Computer Science) degree from the Utkal University, Odisha, India in 2006. Later she had also received her MSc. (Computer Science) degree from Ravenshaw University, Odisha, India in 2008. She has also completed her M. Tech. (Computer Science & Informatics) from the Institute of Technical Education & Research, Siksha O Anusandhan University, Odisha, India in 2013. She is currently working as a faculty member in the Department of Computer Science & IT, Mazoon College, Muscat. She is very keen to be involved in some research work in data mining.

Smita Prava Mishra is an Assistant Professor in the Department of CS & IT, Institute of Technical Education & Research (ITER), SOA University. Bhubaneswar, Odisha. She is an active researcher, currently pursuing her Ph. D. in the area of Data Mining and Soft Computing. She has several publications in diverged areas of Data Mining. She has more than nine years of teaching experience and has taught several papers of Computer Science and Information Technology Domain. She has also supervised several M. Tech thesis works.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication Pvt. Ltd.