iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: ETPAM-2018 **Issue:** Conference **Month of publication:** March 2018

DOI:

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

Introduction to Cluster Analysis

V. Saniya Sulthana[1] A. Kulandhai Terese[2]
[1](PG-MATHS) Department of Mathematics, St. Joseph's College of Arts and science for Women Periyar University
[2]Assistant Professor Department of Mathematics St. Joseph's college of Arts and science for Women Periyar University

Abstract: Cluster analysis is a generic name for a large set of statistical methods that all aim at the detection of groups in a sample of objects, these groups usually called clusters. Here we see the types, levels, evaluation and applications of cluster analysis.

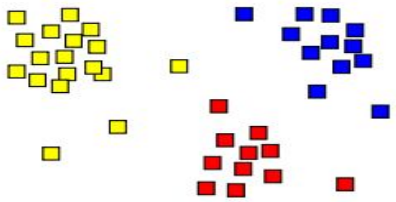
Keywords: Cluster, hierarchical, clustering, agglomerative, divisive internal evaluation by Davies Bouldin index.

I. INTRODUCTION

Cluster analysis was originated in anthropology by DRIVER and KROEBER in 1932.Cluster analysis is a generic name for a large set of statistical methods that all aim at the detection of groups in a sample of objects, these groups usually called clusters. It is a main task of exploratory data mining, and a common technique for statistical data analysis, used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, bioinformatics, data compression and computer graphics.

A. Definition

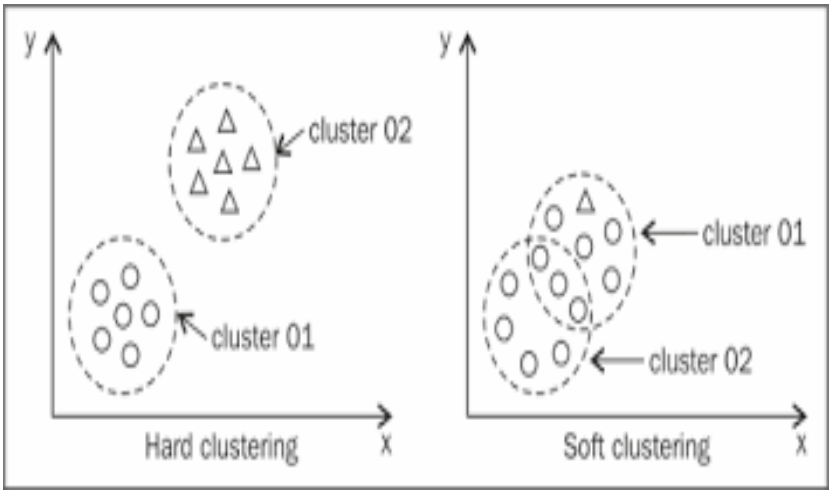
Cluster analysis or clustering is a task of grouping a set of objects in such a way that object in the same group (called a cluster) are more similar to each other than to those in other groups(clusters).
For example, a hierarchy of clusters embedded in each other.



The result of a cluster analysis shown as the coloring of the squares into three clusters.

B. Types Of Clustering

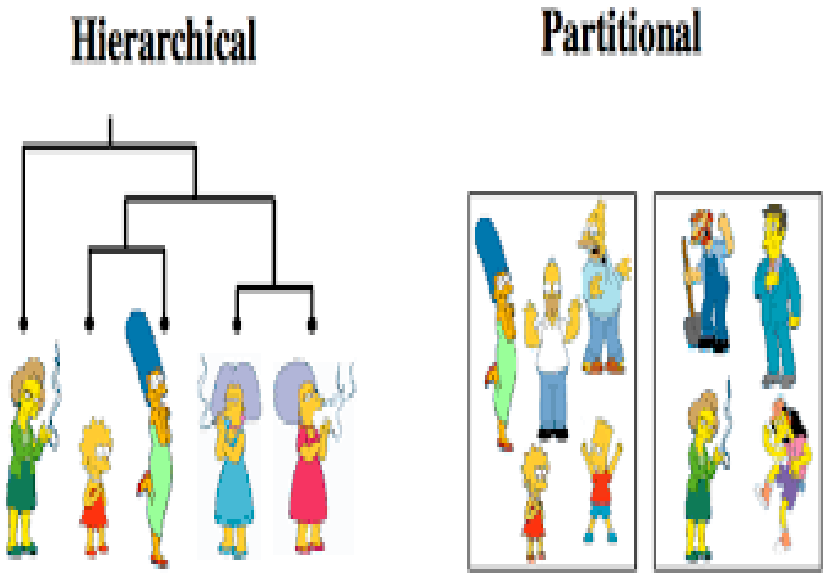
Hard clustering: each object belongs to a group or not. Soft clustering (also: fuzzy clustering): each object belongs to each cluster to a certain degree



©IJRASET (UGC Approved Journal): All Rights are Reserved

C. Hierarchical Clustering

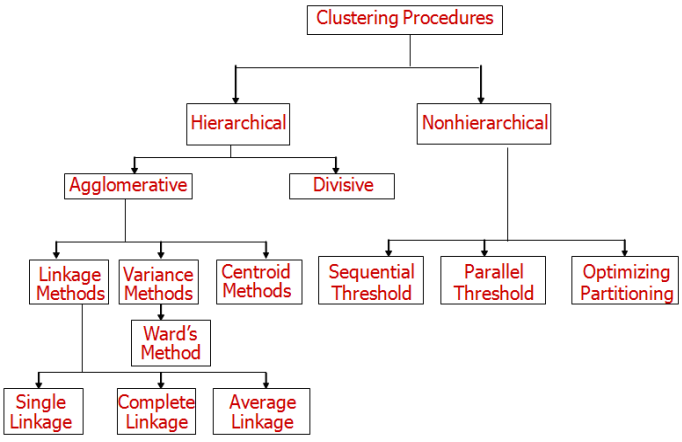
Objects that belong to a child cluster also belongs to a parent cluster. Subspace clustering: while an overlapping clustering within a uniquely defined subspace, clusters are not expected to overlap. Strict partitioning clustering: each object belongs to exactly one cluster. Overlapping clustering (also: alternative clustering, multi-view clustering): objects may belong to more than one cluster; usually involving hard cluster.



What is not cluster analysis?

Supervised classification - have class label information. Simple segmentation - dividing students into different registration groups alphabetically. Results of a query – grouping are a result of an external specification.

Classification of Clustering Procedures

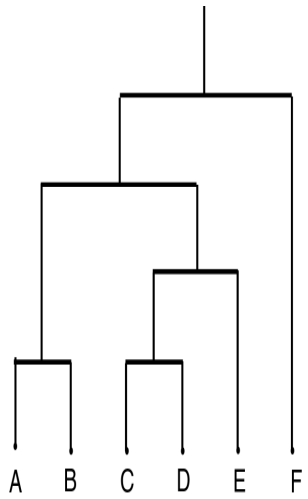


D. Hierarchical Clustering

Clusters are created in levels actually creating sets of clusters at each level.Hierarchical clusters are nested tree-like structures, and usually reflect a development sequence. It may help for "seeing the market structure" in terms of brands. For a set of 100 persons the H.C.A will start with 100 clusters, each containing 1 object and finish with 1 cluster

. tree data structure which illustrates hierarchical clustering techniques.

©IJRASET (UGC Approved Journal): All Rights are Reserved

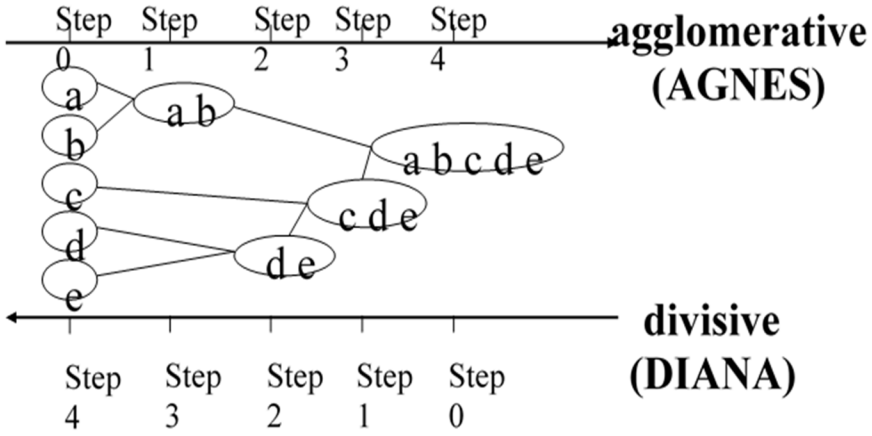


E. Agglomerative

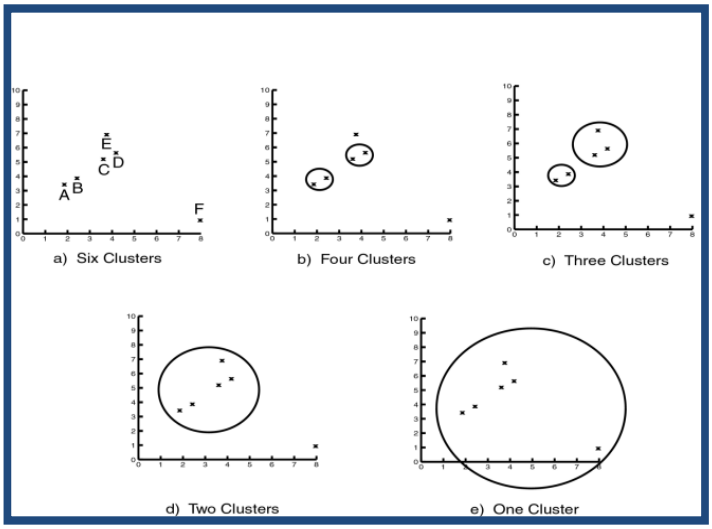
Initially each item in its own cluster iteratively clusters are merged together Bottom Up

F. Divisive

Initially all items in one cluster Large clusters are successively divided Top Down



G. Levels Of Clustering

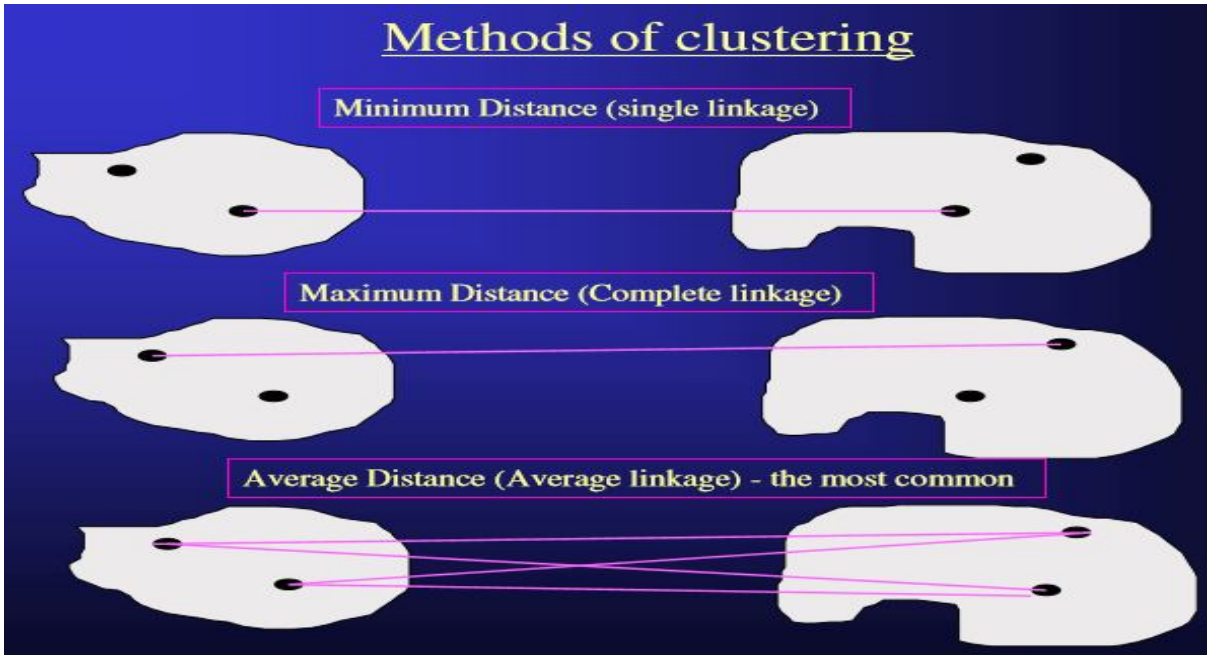


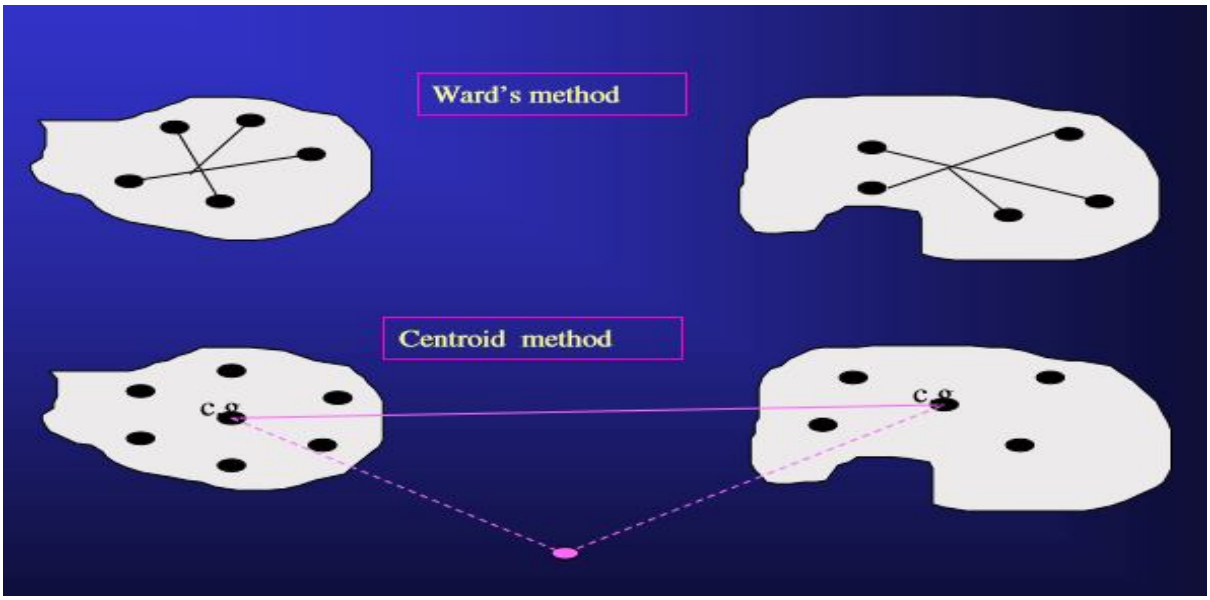
©IJRASET (UGC Approved Journal): All Rights are Reserved

H. Hierarchical Algorithm

1) Single Link: smallest distance between points.*Complete Link:* largest distance between points.

2) Average Link: average distance between points. *Centroid:* distance between Centroid.





.

I. Internal Evaluation

Davies- Bouldin index .Dunn index. Silhouette coefficient.

J. Davies-Bouldin Index

$$DB = \frac{1}{n}\sum_{i=1}^{n} \max_{j \neq i}\left(\frac{\sigma_i + \sigma_j}{d(c_i, c_j)}\right)$$

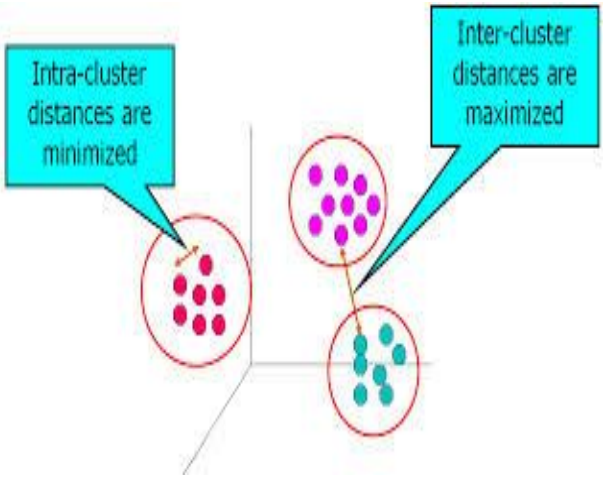
©IJRASET (UGC Approved Journal): All Rights are Reserved

Where n is the number of clusters,

C_x Is the centroid of the cluster,

d (c_i, c_j) is the distance between the Centroid.

Since, algorithm that produces clusters with low intra-cluster distance and high intra-cluster distance will have low Davies-Bouldin index. Smallest Davies-Bouldin index is the best algorithm.



II. APPLICATIONS

A. *Marketing:* Help marketers discover distinct groups in their customer bases, and then use this knowledge to develop targeted marketing programs
B. Land use: Identification of areas of similar land use in an earth observation databases.
C. *Insurance:* Identifying groups of motor insurance policy holders with a high average claim cost.
D. *City-planning:* Identifying groups of houses according to their house type, value, and geographical location.
E. *Earth-quake studies:* Observed earth quake epicenters should be clustered along continent faults.
F. *Transcriptomics:* Build group of genes in with related expression patterns. WWW: Document classification, Cluster WebloData to discover groups of similar access patterns.

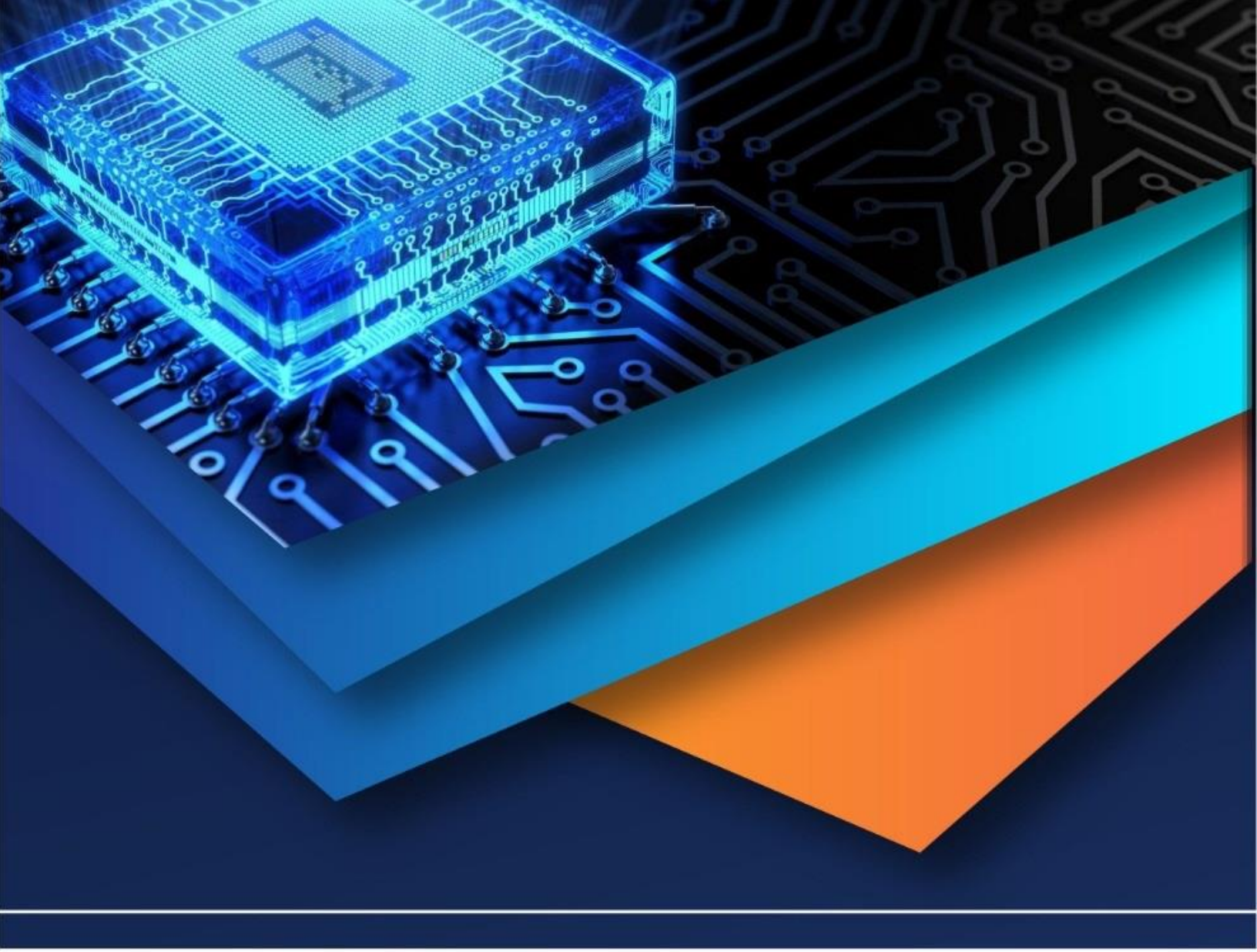
III. CONCLUSION

Recent trend of applied mathematics. It has wide applications and useful in fraud detection. There are still lot of research issues in cluster analysis

REFERENCES

[1] Bailey, ken(1940).”Numerical Taxonomy and Cluster Analysis”. Typologies and Taxonomies.p.34 ISNB 9780803952591
[2] Tryon, Robert C.(1939). Cluster Analysis: Correlation Profile and Orthometric9factor) Analysis for the Isolation of Unities ib mind and personality. Edwards Brothers
[3] Cattell,R.B (1943),”The description of personality: Basic traits resolved into clusters”. Journal of Abnormal and social psychology
[4] -Castro, Vladimir (20 June 2002).”Why so many clustering algorithms –A position paper”.ACMSIGKD
[5] Everitt,Brain(2011) Cluster analysis.Chichester,Sussex, U Wiley.

©IJRASET (UGC Approved Journal): All Rights are Reserved



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)