iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 5 **Issue:** VI **Month of publication:** June 2017

DOI:

www.ijraset.com

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

Volume 5 Issue VI, June 2017
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

Study of Clustering Approach in Data Mining and Machine Knowledge

Ms. Jasmeen Kaur[1], Mr. Saurabh Mittal[2]
[1]M.Tech. Research Scholar Computer Science and Engineering GGGI Dinarpur Ambala Haryana, India
[2]Assistant Professor Computer Science and Engineering GGGI Dinarpur Ambala, Haryana, India

Abstract: ***Data mining alludes to the investigation of the expansive amounts of information that are put away in PCs. Information mining has been called exploratory information examination, not to mention a variety of other things. Masses of information produced from money registers, from filtering, from theme particular databases all through the organization, are investigated, broke down, lessened, and reused. Quests are performed crosswise over distinctive models proposed for anticipating deals, advertising reaction, and benefit. Traditional factual methodologies ar***
principal to information mining. Computerized AI techniques are likewise utilized. Information mining obliges ID of an issue, alongside accumulation of information that can prompt better comprehension, and PC models to give factual or different method for investigation. This paper underlines the clustering techniques used for data mining and machine learning for multiple applications.
Keywords - Clustering, Data Mining, Knowledge Discovery from Databases

I. INTRODUCTION

Information comes in from numerous sources. It is coordinated and put in some regular information store. A piece of it is then taken and preprocessed into a standard arrangement. This
'arranged information' is then gone to an information mining calculation which delivers a yield as standards or some other sort of 'examples'.

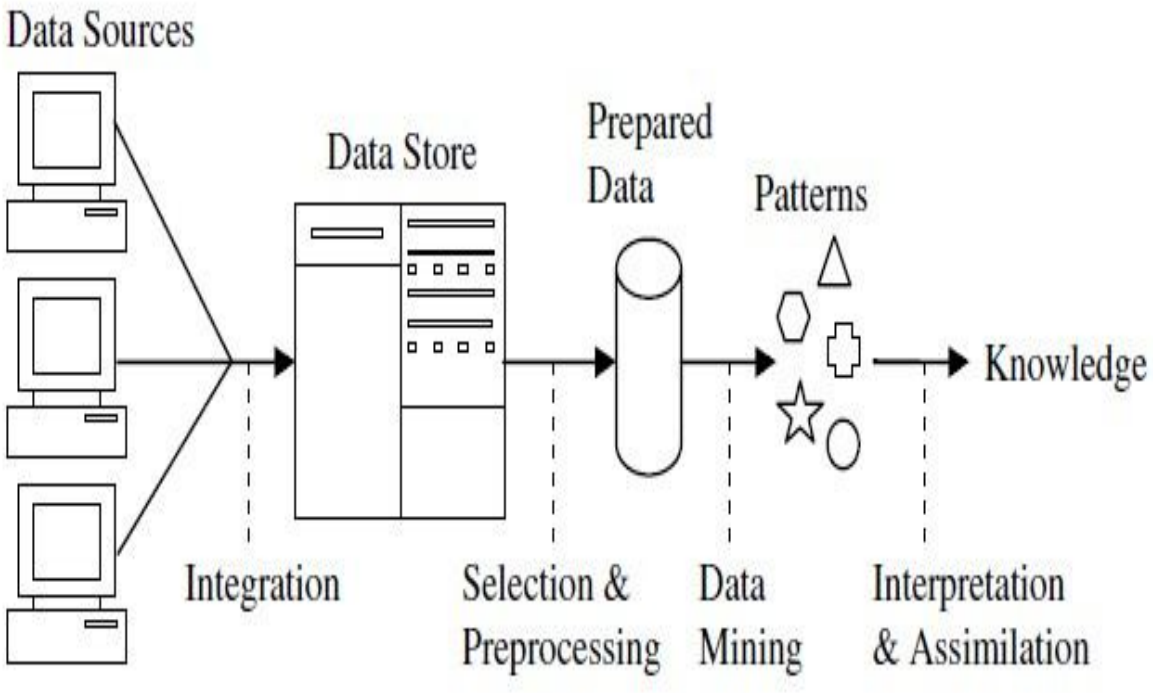


Figure 1.1: The Knowledge Discovery Process

©IJRASET: All Rights are Reserved

Volume 5 Issue VI, June 2017
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

A variety of analytic computer models have been used in data mining. The standard model types in data mining include regression (normal regression for prediction, logistic regression for classification), neural networks, and decision trees. These techniques are well known. This book focuses on less used techniques applied to specific problem types, to include association rules for initial data exploration, fuzzy data mining approaches, rough set models, support vector machines, and genetic algorithms.

II. DATA MINING REQUISITES

Information mining obliges ID of an issue, alongside accumulation of information that can prompt better comprehension, and PC models to give measurable or different method for investigation. This may be upheld by visualization apparatuses, that show information, or through essential measurable examination, for example, relationship investigation. Information mining instruments need to be adaptable, versatile, equipped for precisely foreseeing reactions in the middle of activities and results, and fit for programmed execution. Flexible alludes to the capacity of the apparatus to apply a wide assortment of models. Adaptable devices suggest that if the instruments takes a shot at a little information set, it ought to additionally chip away at bigger information sets. Computerization is helpful, however its application is relative. Some expository capacities are frequently robotized, however human setup preceding actualizing systems is needed. Indeed, expert judgment is discriminating to fruitful usage of information mining. Fitting choice of information to incorporate in inquiries is basic. Information change likewise is regularly needed. An excess of variables deliver an excessive amount of yield, while excessively few can ignore key connections in the information. Major comprehension of measurable ideas is obligatory for fruitful information mining.

III. CLUSTERING

Clustering is an important KDD technique with numerous applications, such as marketing and customer segmentation. Clustering typically groups data into sets in such a way that the intra- cluster similarity is maximized and while inter-cluster similarity is minimized. Clustering is an unsupervised learning. Clustering algorithms examines data to find groups of items that are similar. For example, an insurance company might group customers according to income, age, types of policy purchased, prior claims experience in a fault diagnosis application, electrical faults might be grouped according to the values of certain key variables.

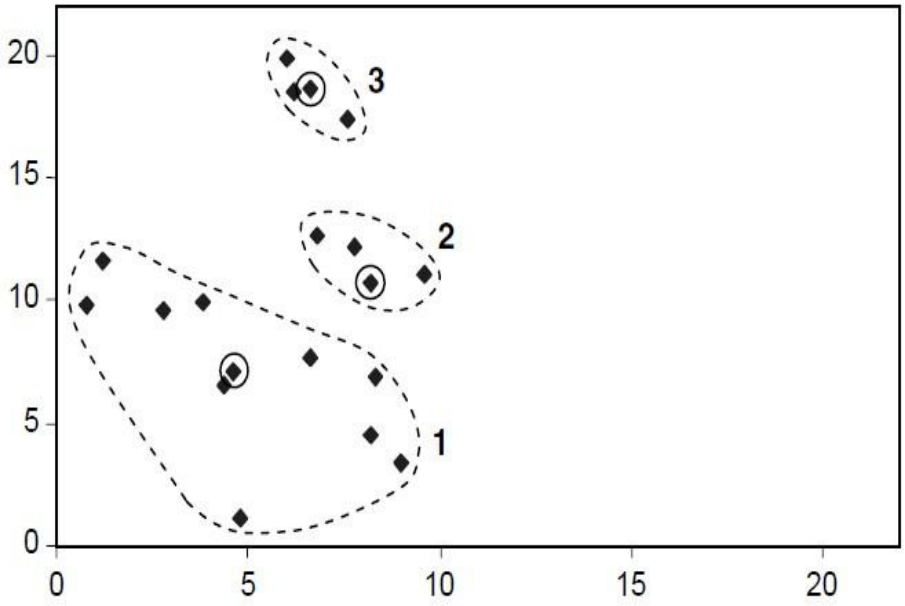


Figure 1.2: Clustering of Data

Most past clustering calculations concentrate on numerical data whose intrinsic geometric properties can be misused regularly to characterize separation capacities between data focuses. Then again, a great part of the data existed in the databases is downright, where characteristic qualities can't be characteristically requested as numerical qualities. Because of the unique properties of absolute characteristics, the clustering of all out data appears to be more muddled than that of numerical data. To defeat this issue, a few data-driven similitude measures have been proposed for absolute data. The conduct of such measures straightforwardly relies on upon the data.

©IJRASET: All Rights are Reserved

Volume 5 Issue VI, June 2017
ISSN: 2321-9653

Grouping can be viewed as the most vital unsupervised learning issue; thus, as every other issue of this kind, it manages discovering a structure in a gathering of unlabeled data. A free meaning of grouping could be "the procedure of sorting out items into gatherings whose individuals are comparative somehow". A cluster is along these lines a gathering of items which are "comparable" in the middle of them and are "divergent" to the articles fitting in with different clusteres.
We can show this with a simple graphical example:

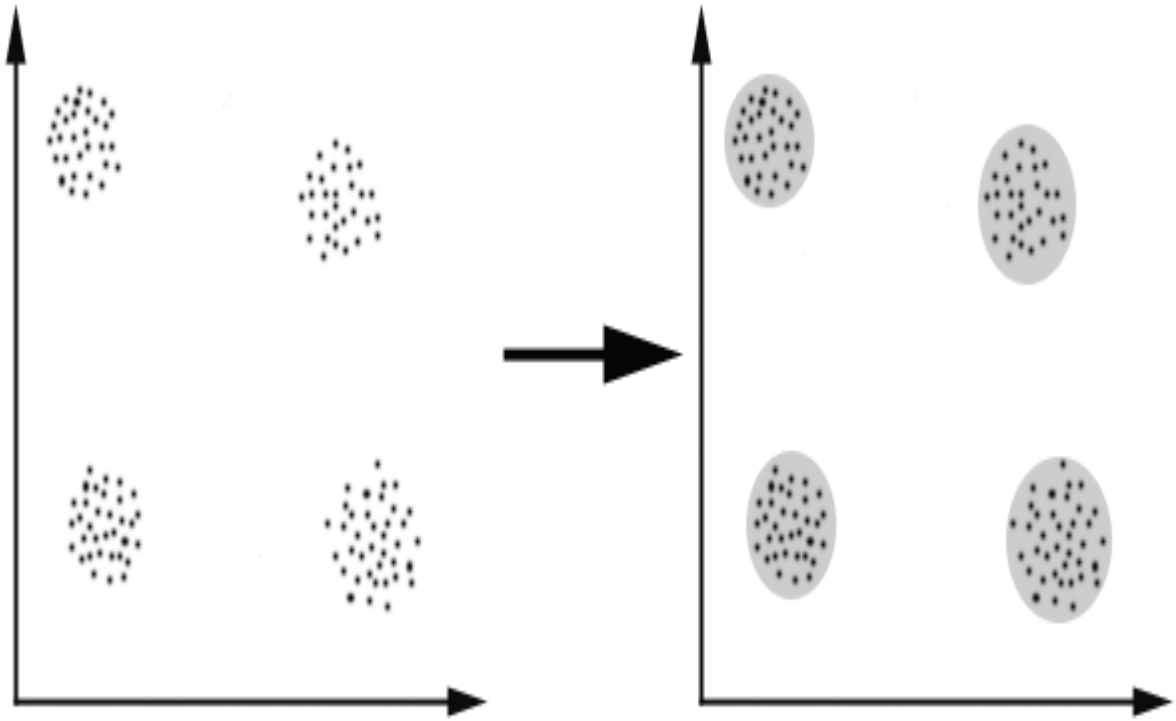


Figure 1.3: Clustering

In this case we easily identify the 4 clusters into which the data can be divided; the similarity criterion is distance: two or more objects belong to the same cluster if they are “close” according to a given distance (in this case geometrical distance).This is called distance-based clustering.
Another kind of clustering is conceptual clustering: two or more objects belong to the same cluster if this one defines a concept common to all that objects. In other words, objects are grouped according to their fit to descriptive concepts, not according to simple similarity measures.
The goal of clustering is to determine the intrinsic grouping in a set of unlabeled data. But how to decide what constitutes a good clustering? It can be shown that there is no absolute “best” criterion which would be independent of the final aim of the clustering. Consequently, it is the user which must supply this criterion, in such a way that the result of the clustering will suit their needs. For instance, we could be interested in finding representatives for homogeneous groups (data reduction), in finding “natural clusters” and describe their unknown properties (“natural” data types), in finding useful and suitable groupings (“useful” data classes) or in finding unusual data objects (outlier detection).

A. Partitioned clustering

Partition-based methods construct the clusters by creating various partitions of the dataset.So, partition gives for each data object the cluster index pi. The user provides the desired number of clusters M, and some criterion function is used in order to evaluate the proposed partition or the solution. This measure of quality could be the average distance between clusters; for instance, some well-known algorithms under this category are k-means, PAM and CLARA. One of the most popular and widely studied clustering methods for objects in Euclidean space is called k- means clustering. Given a set of N data objects xi and an integer M number of clusters.

©IJRASET: All Rights are Reserved

Volume 5 Issue VI, June 2017
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

The problem is to determine C, which is a set of M cluster representativescj, as to minimize the mean squared Euclidean distance from each data object to its nearest centroid.

B. Hierarchical clustering

Hierarchical clustering methods build a cluster hierarchy, i.e. a tree of clusters also known as dendogram. A dendrogram is a tree diagram often used to represent the results of a cluster analysis. Hierarchical clustering methods are categorized into agglomerative (bottom-up) and divisive (top-down) as shown in Figure 4. An agglomerative clustering starts with one-point clusters and recursively merges two or more most appropriate clusters. In contrast, a divisive clustering starts with one cluster of all data points and recursively splits into nonoverlapping clusters.

C. Density-based and grid-based clustering

The key idea of density-based methods is that for each object of a cluster the neighbourhood of a given radius has to contain a certain number of objects; i. e. the density in the neighborhood has to exceed some threshold. The shape of a neighborhood is determined by the choice of a distance function for two objects. These algorithms can efficiently separate noise. DBSCAN and DBCLASD are the well-known methods in the densitybased category.The basic concept of grid- based clustering algorithms is that they quantize the space into a finite number of cells that form a grid structure. And then these algorithms do all the operations on the quantized space. The main advantage of the approach is its fast processing time, which is typically independent of the number of objects, and depends only on the number of grid cells for each dimension. Famous methods in this clustering category are STING and CLIQUE.

D. Outliers

An outlying observation, or outlier[24], is one that appears to deviate markedly from other members of the sample in which it occurs. Outlier points can therefore indicate faulty data, erroneous procedures, or areas where a certain theory might not be valid. However, in large samples, a small number of outliers is to be expected.

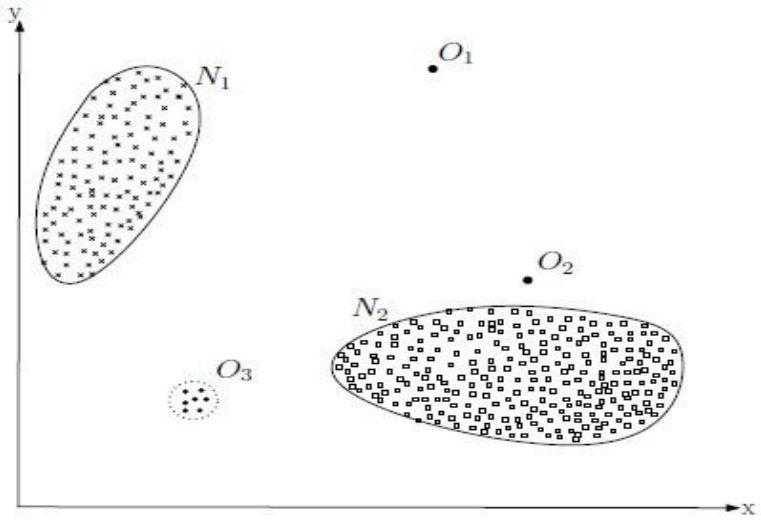


Figure1.4: Outliers in two-dimensional dataset

E. Outlier Detection

Most outlier detection techniques treat objects with K attributes as points in ℜK space

which distinguish potential outliers from others based on the number of objects in the neighborhood. Distribution-based approach deals with statisticalmethods that are based on the probabilistic data model. A probabilistic model can be either a priori given or automatically constructed using given data. If the object does not suit the probabilistic model, it is considered to be an outlier. Third, density-based approach detects local outliers based on the local density of an object's neighbourhood. These methods use different density estimation strategy. A low local density on the observation is an indication of a possible outlier.

©IJRASET: All Rights are Reserved

Volume 5 Issue VI, June 2017
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

F. Distance-based approach

In Distance-based methods outlier is defined as an object that is at least dmin distance away from k percentage of objects in the dataset. The problem is then finding appropriate dmin and k such that outliers would be correctly detected with a small number of false detections. This process usually needs domain knowledge.

Definition: A point x in a dataset is an outlier with respect to the parameters k and d, if no more than k points in the dataset are at a distance d or less from x.

To explain the definition by example we take parameter k = 3 and distance d as shown in Figure

1.5. Here are points xi and xj be defined as outliers, because of inside the circle for each point lie no more than 3 other points. And x′ is an inlier, because it has exceeded number of points inside the circle for given parameters k and d.

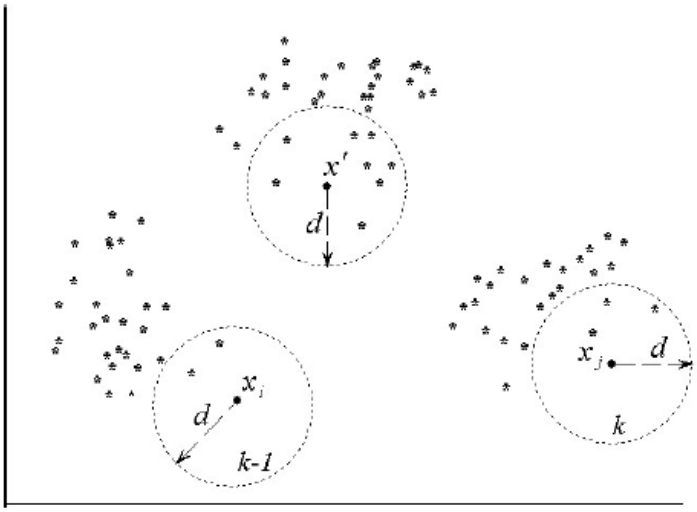


Figure 1.5: Illustration of outlier definition by Knorr and Ng.

G. Distribution-based approach

Distribution-based methods originate from statistics, where object is considered as an outlier if it deviates too much from underlying distribution. For example, in normal distribution outlier is an object whose distance from the average object is three times of the variance.

H. Density-based approach

Density-based methods have been developed for finding outliers in a spatial data. These methods can be grouped into two categories called multi-dimensional metric space-based methods and graph-based methods. In the first category, the definition of spatial neighborhood is based on Euclidean distance, while in graph-based spatial outlier detections the definition is based on graph connectivity. Whereas distribution-based methods consider just the statistical distribution of attribute values, ignoring the spatial relationships among items, density-based approach consider both attribute values and spatial relationship.

IV. FUZZY FITNESS VALUE

To apply a transformative calculation to a given issue, we require somehow of assessing applicant arrangements. This is finished with a wellness capacity. A wellness capacity is a capacity that takes as enter a representation (or genotype), makes an interpretation of this into the relating applicant arrangement (or phenotype), tests this hopeful arrangement on the given issue, and afterward gives back a number that shows how great this arrangement is, i.e., its wellness. Case in point, in the subset aggregate issue, the wellness capacity takes as enter somewhat string, from this it builds the comparing subset of numbers, and contrasts the whole of this subset with the given number M. In the TSP issue, the information is a stage, which gets interpreted into the relating visit, and the length of this visit is then figured.

Since we by and large view development as a boost issue, i.e., wellness ought to be amplified, a wellness capacity is normally built in such a route, to the point that a higher wellness worth is

appointed to better arrangements. In the cases given, we are managing minimization issues (as is frequently the case in

©IJRASET: All Rights are Reserved

certifiable issues), so better arrangements will have lower qualities, (for example, a littler contrast with the given number M, or shorter visit length). On the other hand, each minimization issue can be effectively changed over into an amplification issue, and the wellness capacity could, for instance, give back the backwards of the computed wellness qualities, or some huge number short the figured quality. In this way, here we will just consider transformative calculations as augmenting the wellness values, in relationship with science. The wellness capacity for a given issue (whether it is a minimization or amplification issue) can simply be composed in the suitable structure.

There are many ways to handle constraints in a GA. At the high conceptual level we can distinguish two cases: indirect constraint handling and direct constraint handling.

Indirect constraint handling means that we circumvent the problem of satisfying constraints by incorporating them in the fitness function *f* such that *f* optimal implies that the constraints are satisfied, and use the power of GA to find a solution.

Direct constraint handling means that we leave the constraints as they are and 'adapt' the GA to enforce them.

Notice that direct and indirect constraint handling can be applied in combination, i.e., in one application we can handle some constraints directly and others indirectly.

Formally, indirect constraint handling means transforming constraints into optimization objectives.

A fitness function is a particular type of objective function that is used to summarise, as a single figure of merit, how close a given design solution is to achieving the set aims.

In particular, in the fields of genetic programming and genetic algorithms, each design solution is represented as a string of numbers (referred to as a candidate). After each round of testing, or simulation, the idea is to delete the 'n' worst design solutions, and to breed 'n' new ones from the best design solutions. Each design solution, therefore, needs to be awarded a figure of merit, to indicate how close it came to meeting the overall specification, and this is generated by applying the fitness function to the test, or simulation, results obtained from that solution.

The reason that genetic algorithms are not a lazy way of performing design work is precisely because of the effort involved in designing a workable fitness function. Even though it is no longer the human designer, but the computer, that comes up with the final design, it is the human designer who has to design the fitness function. If this is designed wrongly, the algorithm will either converge on an inappropriate solution, or will have difficulty converging at all.

Moreover, the fitness function must not only correlate closely with the designer's goal, it must also be computed quickly. Speed of execution is very important, as a typical genetic algorithm must be iterated many times in order to produce a usable result for a non-trivial problem.

A. Fitness approximation may be appropriate, especially in the following cases:

1) Fitness computation time of a single solution is extremely high

2) Precise model for fitness computation is missing

3) The fitness function is uncertain or noisy.

Two main classes of fitness functions exist: one where the fitness function does not change, as in optimizing a fixed function or testing with a fixed set of test cases; and one where the fitness function is mutable, as in niche differentiation or co-evolving the set of test cases.

nother way of looking at fitness functions is in terms of a fitness landscape, which shows the fitness for each possible candidate.

Definition of the fitness function is not straightforward in many cases and often is performed iteratively if the fittest solutions produced by GA are not what is desired. In some cases, it is very hard or impossible to come up even with a guess of what fitness function definition might be. Interactive genetic algorithms address this difficulty by outsourcing evaluation to external agents (normally humans).

In order to understand the fitness function, you first have to understand that a genetic algorithm is one which changes over time (it evolves). In nature we have things like predators and harsh environments which eliminate unwanted specimens of animals (a slow zebra will get eaten by a lion). We need to simulate this behavior when programming genetic algorithms.

The fitness function basically determines which possible solutions get passed on to multiply and mutate into the next generation of solutions. This is usually done by analyzing the "genes," which hold some data about a particular solution to the problem you are trying to solve. The fitness function will look at the genes and make some qualitative assessment, returning a fitness value for that solution. The rest of the genetic algorithm will discard any solutions with a "poor" fitness value and

©IJRASET: All Rights are Reserved

accept any with a "good" fitness value.
In short: the goal of a fitness function is to provide a meaningful, measurable, and comparable value given a set of genes.

B. Direct constraint handling

Treating constraints directly implies that violating them is not reflected in the fitness function, thus there is no bias towards candidates satisfying them. Therefore, the population will not become less and less infeasible w.r.t. these constraints. This means that we have to create and maintains feasible candidates in the population. The basic problem in this case is that the regular operators are blind to constraints, mutating one or crossing over two feasible candidates can result in infeasible offspring.
liminating infeasible candidates is very inefficient, and therefore hardly applicable. Repairing infeasible candidates requires a repair procedure that modifies a given candidate such that it will not violate constraints. This technique is thus problem dependent.
The preserving approach amounts to designing and applying problem-specific operators that do preserve the feasibility of parent candidates. Note that the preserving approach requires the creation of a feasible initial population, which can be NP-complete.
Decoding can simplify the problem search space and allow an efficient genetic algorithm. Formally, decoding can be seen as shifting to a search space that is different from the Cartesian product of the domains of the variables in the original problem formulation.

C. Indirect Constraint Handling

In the case of indirect constraint handling the optimization objectives replacing the constraints are viewed *penalties* for constraint violation hence to be minimized. In general penalties are given for violated constraints although some GAs allocate penalties for wrongly instantiated variables or as the distance to a feasible solution.

1) Advantages of indirect constraint handling are:

a) generality
b) reduction of the problem to 'simple' optimization
c) ossibility of embedding user preferences by means of weights.

2) Disadvantages of indirect constraint handling are:
a) loss of information by packing everything in a single numberdoes not work well with sparse problems.

V. APPLICATIONS

A. Biology, computational biology and bioinformatics

Plant and animalecology: Cluster analysis is used to describe and to make spatial and temporal comparisons of communities (assemblages) of organisms in heterogeneous environments; it is also used in plant systematics to generate artificial phylogenies or clusters of organisms (individuals) at the species, genus or higher level that share a number of attributes

1) Transcriptomics

clustering is used to build groups of genes with related expression patterns (also known as coexpressed genes). Often such groups contain functionally related proteins, such as enzymes for a specific pathway, or genes that are co-regulated. High throughput experiments using expressed sequence tags (ESTs) or DNA microarrays can be a powerful tool for genome annotation, a general aspect of genomics.

2) Sequence analysis

lustering is used to group homologous sequences into gene families. This is a very important concept in bioinformatics, and evolutionary biology in general.
High-throughput genotyping platforms
clustering algorithms are used to automatically assign genotypes.
Human genetic clustering The similarity of genetic data is used in clustering to infer population structures.

3) Medical imaging

On PET scans, cluster analysis can be used to differentiate between different types of tissue and blood in a three-dimensional

©IJRASET: All Rights are Reserved

image. In this application, actual position does not matter, but the voxel intensity is considered as a vector, with a dimension for each image that was taken over time. This technique allows, for example, accurate measurement of the rate a radioactive tracer is delivered to the area of interest, without a separate sampling of arterial blood, an intrusive technique that is most common today.

4) Analysis of antimicrobial activity

Cluster analysis can be used to analyse patterns of antibiotic resistance, to classify antimicrobial compounds according to their mechanism of action, to classify antibiotics according to their antibacterial activity.

IMRT segmentation

Clustering can be used to divide a fluence map into distinct regions for conversion into deliverable fields in MLC-based Radiation Therapy.

5) Market research

Cluster analysis is widely used in market research when working with multivariate data from surveys and test panels. Market researchers use cluster analysis to partition the general population of consumers into market segments and to better understand the relationships between different groups of consumers/potential customers, and for use in market segmentation, Product positioning, New product development and Selecting test markets.

6) Grouping of shopping items

Clustering can be used to group all the shopping items available on the web into a set of unique products. For example, all the items on eBay can be grouped into unique products. (eBay doesn't have the concept of a SKU)

7) Social network analysis

In the study of social networks, clustering may be used to recognize communities within large groups of people.

8) Search result grouping

In the process of intelligent grouping of the files and websites, clustering may be used to create a more relevant set of search results compared to normal search engines like Google. There are currently a number of web based clustering tools such as Clusty.

9) Slippy map optimization

Flickr's map of photos and other map sites use clustering to reduce the number of markers on a map. This makes it both faster and reduces the amount of visual clutter.

10) Software evolution

Clustering is useful in software evolution as it helps to reduce legacy properties in code by reforming functionality that has become dispersed. It is a form of restructuring and hence is a way of directly preventative maintenance.

11) Image segmentation

Clustering can be used to divide a digitalimage into distinct regions for border detection or object recognition.

12) Evolutionary algorithms

Clustering may be used to identify different niches within the population of an evolutionary algorithm so that reproductive opportunity can be distributed more evenly amongst the evolving species or subspecies.

13) Recommender systems

Recommender systems are designed to recommend new items based on a user's tastes. They sometimes use clustering algorithms to predict a user's preferences based on the preferences of other users in the user's cluster.

Markov chain Monte Carlo methods

Clustering is often utilized to locate and characterize extrema in the target distribution.

©IJRASET: All Rights are Reserved

14) Crime analysis

Cluster analysis can be used to identify areas where there are greater incidences of particular types of crime. By identifying these distinct areas or "hot spots" where a similar crime has happened over a period of time, it is possible to manage law enforcement resources more effectively.

15) Educational data mining

Cluster analysis is for example used to identify groups of schools or students with similar properties.

16) Typologies

From poll data, projects such as those undertaken by the Pew Research Center use cluster analysis to discern typologies of opinions, habits, and demographics that may be useful in politics and marketing.

17) Field robotics

Clustering algorithms are used for robotic situational awareness to track objects and detect outliers in sensor data.

18) Mathematical chemistry

To find structural similarity, etc., for example, 3000 chemical compounds were clustered in the space of 90 topological indices.[38]

19) Climatology

To find weather regimes or preferred sea level pressure atmospheric patterns.

20) etroleum geology

Cluster analysis is used to reconstruct missing bottom hole core data or missing log curves in order to evaluate reservoir properties. Physical geography

he clustering of chemical properties in different sample locations.

Cluster Arrangement or basically clustering is the procedure of collection the arrangement of articles in such a way, to the point that protests in the same gathering called cluster that are more comparable in some sense or another to one another than to those in different gatherings or clusters. It is an unmistakable and compulsory assignment of exploratory information mining, and a typical procedure for factual information investigation utilized as a part of numerous fields, including machine learning, example acknowledgment, picture examination, data recovery, and bioinformatics. Cluster investigation itself is not one particular calculation, but rather the general errand to be fathomed. It can be accomplished by different calculations that contrast fundamentally in their thought of what constitutes a cluster and how to effectively discover them. Prevalent ideas of clusters incorporate gatherings with little separations among the cluster individuals, thick zones of the information space, interims or specific factual dispersions. Clustering can accordingly be detailed as a multi-target streamlining issue.

REFERENCES

[1] Achtert, E.; Böhm, C.; Kröger, P. (2006). "DeLi-Clu: Boosting Robustness, Completeness, Usability, and Efficiency of Hierarchical Clustering by a Closest Pair Ranking". LNCS: Advances in Knowledge Discovery and Data Mining. Lecture Notes in Computer Science 3918: 119–128. doi:10.1007/11731139_16. ISBN 978-3-540-33206-0.
[2] Aditya Desai, Himanshu Singh, Vikram Pudi, 2011. DISC: Data-Intensive Similarity Measure for Categorical Data, Pacific-Asia Conferences on Knowledge Discovery Data Mining
[3] Andre Baresel, Harmen Sthamer, Michael Schmidt,2002. Fitness Function Design to improve Evolutionary Structural Testing
[4] Andrew L.Nelson, Gregory J.Barlow, Lefteris Doitsidis,2008 .Fitness Functions in Evolutionary Robotics: A Survey and Analysis
[5] Can, F.; Ozkarahan, E. A. (1990). "Concepts and effectiveness of the cover-coefficient- based clustering methodology for text databases". ACM Transactions on Database Systems 15 (4): 483. doi:10.1145/99935.99938
[6] Hans-Peter Kriegel, Peer Kröger, Jörg Sander, Arthur Zimek (2011). "Density-based Clustering". WIREs Data Mining and Knowledge Discovery 1 (3): 231–240. doi:10.1002/widm.30.
[7] He Zengyou, Xu Xiaofei, Deng Shenchun, 2002. Squeezer: An Efficient Algorithm for Clustering Categorical Data,Journal of Computer Science and Technology,Vol. 17, No. 5,pp 611-624
[8] He Zengyou, Xu Xiaofei, Deng Shenchun, 2003. Discovering Cluster Based Local Outliers,Article Published in Journal Pattern Recognition Letters,

©IJRASET: All Rights are Reserved

Volume 5 Issue VI, June 2017
ISSN: 2321-9653

Volume 24. Issue 9-10,pp 1641-1650,01 June 2003

[9] He Zengyou, Xu Xiaofei, Deng Shenchun, 2006. Improving Categorical Data Clustering Algorithm by Weighting Uncommon Attribute Value Matches, ComSIS Vol.3,No.1

[10] Jerzy Stefanowski, 2009, Data Mining - Clustering, University of Technology, Poland Lloyd, S. (1982). "Least squares quantization in PCM". IEEE Transactions on information Theory 28 (2): 129–137. doi:10.1109/TIT.1982.1056489.

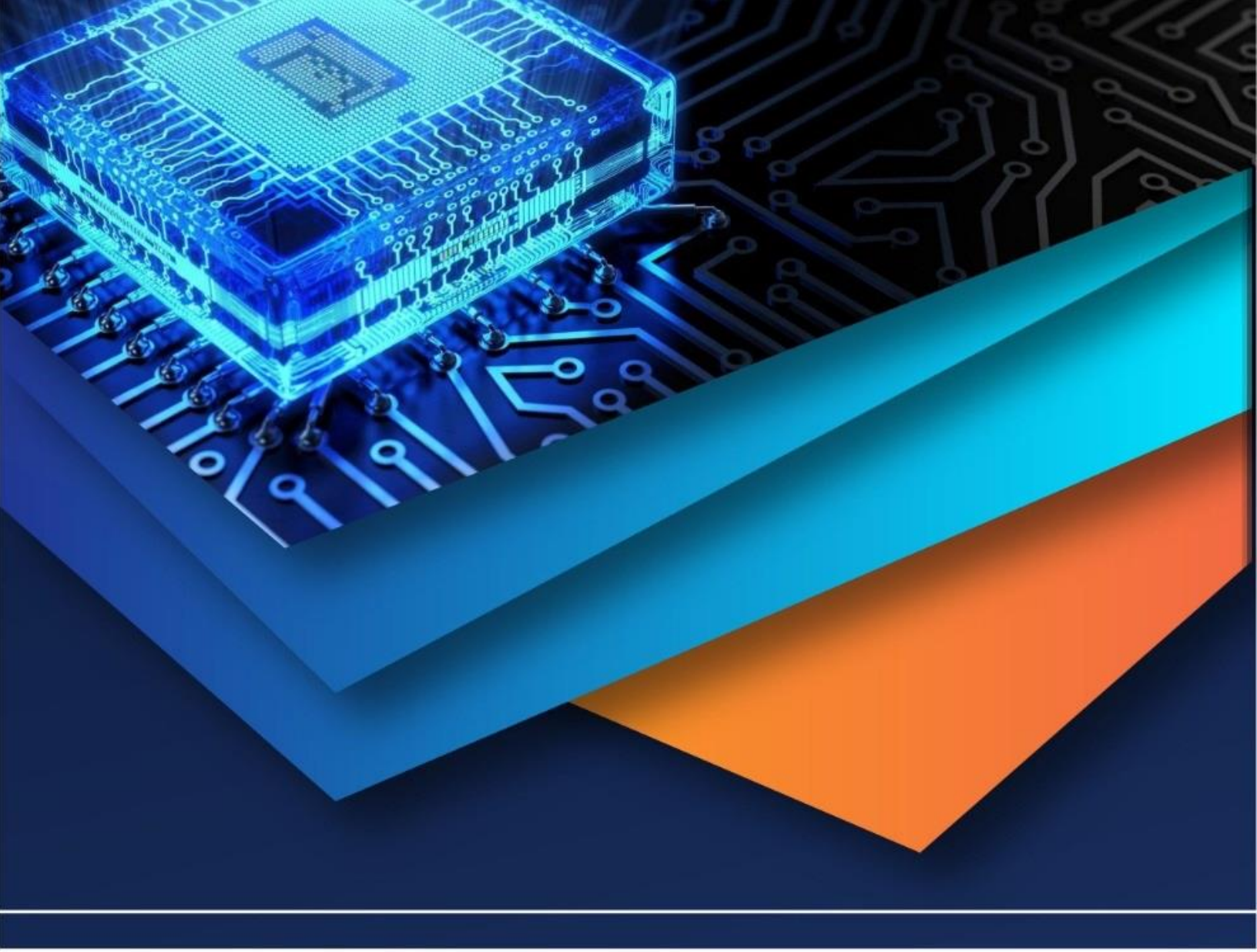
[11] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu (1996). "A density- based algorithm for discovering clusters in large spatial databases with noise".

[12] In Evangelos Simoudis, Jiawei Han, Usama M. Fayyad. Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96). AAAI Press. pp. 226–231. ISBN 1-57735-004-9.

©IJRASET: All Rights are Reserved

Volume 5 Issue VI, June 2017
ISSN: 2321-9653

©IJRASET: 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)