International Journal of Emerging Technology in Computer Science & Electronics (IJETCSE)
ISSN: 0976-1353 Volume 3 Issue 1 –NOVEMBER 2013.

A Stride towards Developing Efficient Approaches for Data Clustering Based on Evolutionary Programming

P.Senthilkumar [#1] and Dr. N.Suthanthira Vanitha[2]*,

 # Research Scholar,Anna University

**Professor, Department of EEE, Knowledge Institute of Technology*

***Abstract*- The primary objective of my research is to devise and develop an effective approach for optimized accomplishment of clustering. The research has been inspired by the need for an efficient approach to improve the performance of the clustering approach and their associated applications. It is not an easy task if the collected data contains more noise and irrelevant items that affects the accuracy of the clustering approach.**

***Index Terms*- Data Clustering, evolutionary programming, data mining**

I. INTRODUCTION

Across a wide variety of fields, data are being collected and accumulated at a dramatic pace. There is an urgent need for a new generation of computational theories and tools to assist humans in extracting useful information (knowledge) from the rapidly growing volumes of digital data. These theories and tools are the subject of the emerging field of knowledge discovery in databases (KDD) [47]. Knowledge discovery in databases uses methods and techniques that are derived from the areas of statistical and data analysis, decision support and machine learning. A formal definition of Knowledge discovery in databases is the non-trivial process of identifying valid, previously unknown and potentially useful information from large datasets [48, 49]. The Knowledge Discovery in Databases process comprises of a few steps leading from raw data collections to some form of new knowledge. The iterative process consists of the following steps [50]:

Data cleaning: also known as data cleansing, it is a phase in which noise data and irrelevant data are removed from the collection. Data integration: at this stage, multiple data sources, often heterogeneous, may be combined in a common source. Data selection: at this step, the data relevant to the analysis is decided on and retrieved from the data collection. Data transformation: also known as data consolidation, it is a phase in which the selected data is transformed into forms appropriate for the mining procedure. Data mining: it is the crucial step in which clever techniques are applied to extract patterns potentially useful. Pattern evaluation: in this step, strictly interesting patterns representing knowledge are identified based on given measures. Knowledge representation: it is the final phase in which the discovered knowledge is visually represented to the user. This essential step uses visualization techniques to help users understand and interpret the data mining results. Among the Knowledge Discovery in Databases process, data cleaning is an emerging domain that aims at improving data quality through the detection and elimination of data artifacts. These data artifacts comprise of errors, discrepancies, redundancies, ambiguities, and incompleteness that hampers the efficacy of analysis or data mining [1]. Data cleaning, also called data cleansing or scrubbing, is an automated method for examining the data, detecting missing and incorrect values and correcting them [5]. It focuses on eliminating variations in data contents and reducing data redundancy aimed at improving the overall data consistency [4]. Data cleaning first detects dirty records by determining whether two or more records represented differently refer to the same real world entity, and then, it cleans the dirty records by either (i) collapsing them to get a consolidated whole devoid of missing parts, (ii) unifying them with a single entity identity and (iii) retaining only one copy of records that are exact duplicates [3]. Data cleansing comprises the identification and removal of errors in existing data sets to enhance the overall data quality [1]. Data quality problems are present in single data collections, such as files and databases, e.g., due to misspellings during data entry, missing information or other invalid data. In order to provide access to accurate and consistent data, consolidation of different data representations and elimination of duplicate information becomes necessary [2]. Duplicate record detection is the process of identifying different or multiple records that refer to one unique real world entity or object. Typically, the process of duplicate detection is preceded by a data preparation stage during which data entries are stored in a uniform manner in the database, resolving (at least partially) the structural heterogeneity problem. The data preparation stage includes a parsing, a data transformation, and a standardization step. The approaches that deal with data preparation are also described under using the term ETL (Extraction, Transformation,

Loading). These steps improve the quality of the in-flow data and make the data comparable and more usable [61, 16]. Most existing data cleaning methods focus on removing noise that is the result of low-level data errors that result from an

imperfect data collection process, but data objects that are irrelevant or only weakly relevant can also significantly hinder data analysis. Consequently, there is a need for data cleaning techniques that remove both types of noise. Because data sets can contain large amounts of noise, these techniques also need to be able to discard a potentially large fraction of the data [44]. After cleaning the data, data mining is crucial because data mining is a core component of the KDD process. The commercial and research interests in data mining is increasingly rapidly, as the amount of data generated and stored in databases of organizations is already enormous and continuing to grow very fast. This large amount of stored data normally contains valuable hidden knowledge, which, if harnesses, could be used to improve the decision making process of an organization. Data mining engine is ideally consists of a set of functional modules for tasks such as characterization, association, classification, cluster analysis, and evolution. The concept of clustering has been around for a long time. It has several applications, particularly in the context of information retrieval and in organizing web resources [13]. Clustering is the unsupervised classification of patterns into groups. Clustering is the task of grouping a set of objects into different subsets such that objects belonging to the same cluster are highly similar to each other [14]. Clustering is a vital process for condensing and summarizing information, since it can provide a synopsis of the stored data. Clustering has many applications, including part family formation for group technology, image segmentation, information retrieval, web pages grouping, market segmentation, and scientific and engineering analysis [15]. The clustering problem has been studied in many disciplines such as statistics, pattern recognition,

signal processing (e.g., vector quantization), biology, and so on. As a consequence numerous clustering algorithms had been proposed in these different communities, spanning different clustering paradigms such as partitional [7, 8], hierarchical [11], spectral [12], density-based [9], mixture-modeling [10], and so on. Partitional clustering techniques create a one-level (unnested) partitioning of the data points. If K is the desired number of clusters, then partitional approaches typically find all K clusters at once. In recent years, it has been recognized that the partitional clustering technique is well suited for clustering a large dataset due to their relatively low computational requirements [41, 42]. The time complexity of the partitioning technique is almost linear, which makes it widely used. The best known partitioning clustering algorithm is the K-means algorithm and its variants [43]. This algorithm is simple, straightforward and is based on the firm foundation of the analysis of variances. Recently published studies have shown that partitional clustering algorithms that optimize certain criterion functions, which measure key aspects of inter- and intra-cluster similarity, are very effective in producing hard clustering solutions for datasets and out perform traditional partitional and agglomerative algorithms. But it is prone to convergence near local optima. Since stochastic optimization approaches are good in preventing convergence to local optimum solutions, these methods could be used to find global optima [51, 52]. Tremendous research effort has gone in the past few years to evolve the clusters in complex data sets through evolutionary computing techniques. Most of the existing clustering techniques, based on evolutionary algorithms, accept the number of classes K as an input instead of determining the same on the run. For example, while clustering a set of data arising from the query to a search engine, the number of classes K changes for each set of data that result from an interaction with the search engine. Also, if the data set is described by high-dimensional feature vectors (which is very often the case), it may be practically impossible to visualize the data for tracking its number of clusters [45].

II. MOTIVATION FOR THE RESEARCH

Data warehouses require and provide extensive support for data cleaning. They load and continuously refresh huge amounts of data from a variety of sources so the probability that some of the sources contain “dirty data” is high. Due to the wide range of possible data inconsistencies and the sheer data volume, data cleaning is considered to be one of the biggest problems in data warehousing [29, 30]. A data cleaning approach should satisfy several requirements. First of all, it should detect and remove all major errors and inconsistencies both in individual data sources and when integrating multiple sources. The approach should be supported by tools to limit manual inspection and programming effort and be extensible to easily cover additional sources. Furthermore, data cleaning should not be performed in isolation but together with schema-related data transformations based on comprehensive metadata [2]. A main problem for cleaning data from multiple sources is to identify overlapping data, in particular matching records referring to the same real-world entity (e.g., customer). This problem is also referred to as the object identity problem, duplicate elimination or the merge/purge problem [28]. Frequently, the information is only partially redundant and the sources may complement each other by providing additional information about an entity. Thus duplicate information should be purged out and complementing information should be consolidated and merged in order to achieve a consistent view of real world entities [2]. The duplicate elimination task is typically performed after most other transformation and cleaning steps, especially after having cleaned single-source errors and conflicting representations. It is performed either on two cleaned sources at a time or on a single already integrated data set. Duplicate elimination requires to first identify (i.e. match) similar records concerning the same real world entity. In a second step, similar records are merged into one record containing all relevant attributes without redundancy. Furthermore, redundant records are purged [31]. Clustering

[33] has been studied extensively for more than forty years and across many disciplines due to its broad applications. Clustering is the process of assigning data objects into a set of disjoint groups called clusters so that objects in each cluster are more similar to each other than objects from different clusters. The literature presents with an enormous number of algorithms for efficient clustering of data. These algorithms can be categorized into nearest neighbor clustering, fuzzy clustering, partitional clustering, hierarchical clustering, artificial neural networks for clustering, statistical clustering algorithms, density-based clustering algorithm and so on. In these methods, hierarchical and partitional clustering algorithms are two primary approaches of increasing interest in research communities. Hierarchical clustering algorithms can usually find satisfiable clustering results. Although the hierarchical clustering technique is often portrayed as a better quality clustering approach, this technique does not contain any provision for the reallocation of entities, which may have been poorly classified at the early stage. Furthermore, most of the hierarchical algorithms are very computationally intensive and require much memory space [32]. In recent times, partitioning clustering techniques have been widely used among the researches suited for clustering a large datasets. The partitional clustering algorithms construct a single partition of a database D of n objects into a set of k clusters, such that the objects in a cluster are more similar to each other than to objects in different clusters. As a traditional clustering algorithm, k-Means is a popular partitional clustering algorithm for its simplicity in implementation, and it is commonly applied in diverse applications. Although widely used, the k-means algorithm suffers from some drawbacks, including: 1) Dependency on the input order 2) The tendency to result in local minimum 3) Limited applicability to only the data set consisting of isotropic clusters (i.e., a circle in the 2D domain or sphericity in the 3D domain) 4) Choosing the k initial clusters. 5) Degeneracy: Degeneracy means that the clustering may end with some empty clusters [20]. One of the important problems in partitional clustering is to find a partition of the given data, with a specified number of clusters, which minimizes the total within cluster variation (TWCV). In general, artitional clustering algorithms are iterative and hill climbing and usually they converge to a local minimum. As a consequence, it is very difficult to find an optimal partition of the data using hill climbing techniques. The algorithms based on combinatorial optimization such as integer programming, dynamic programming and, branch and bound methods are expensive ever for moderate number of data points and moderate number of clusters. A detailed discussion on clustering algorithms can be found in [10]. The simplest and most popular among iterative and hill climbing clustering algorithms is the K-means algorithm (KMA). As mentioned above, this algorithm may converge to a suboptimal partition. Since stochastic optimization approaches are good at avoiding convergence to a locally optimal solution, these approaches could be used to find a globally optimal solution. The stochastic approaches used in clustering include those based on simulated annealing, genetic algorithms, evolution strategies and evolutionary programming [53 -60]. The algorithmic task can be stated as an optimization problem for which the objective is to maximize the similarities among objects within the same clusters while minimizing the dissimilarities between different clusters. However, it produces a local optimal solution that strongly depends on its initial seeds. Bad initial seeds can also cause the splitting or merging of natural clusters even if the clusters are well separated [6]. In order to overcome local optima problem lots of studies have done in clustering. So, in recent ears, many clustering algorithms based on evolutionary algorithms such as genetic algorithm, tabu search and SA have been introduced [46]. The problem of partitional clustering has been approached from diverse fields of knowledge, such as statistics (multivariate analysis) [34], graph theory [35], expectation– maximization algorithms [36], artificial neural networks [37], evolutionary computing [38], and so on. Many partitional clustering algorithms that have been introduced in recent years are based on genetic algorithm (GAS) [40], which are stochastic search heuristics inspired by Darwinian evolution and genetics. An important advantage of these algorithms is their ability to cope with local optima by maintaining, recombining and comparing several candidate solutions simultaneously [39].

.

III. REVIEW OF RELATED RESEARCH

Literature presents enormous amount of works for data cleaning and partitional based clustering. In recent times, data cleaning, clustering using partitional based methods and clustering with the aid of evolutionary programming have gained enormous popularity among the researchers. A brief review of some of the recent researches is presented below: Rehman, M. and Esichaikul, V. [17] have focused on one of the major issue of data cleansing i.e. "duplicate record detection" which arises when the data is collected from various sources. As a result, comparison among standard duplicate elimination algorithm (SDE), sorted neighborhood algorithm (SNA), duplicate elimination sorted neighborhood algorithm (DE-SNA), and adaptive duplicate detection algorithm(ADD) was provided. A prototype was also developed which shows thatadaptive duplicate detection algorithm is the optimal solution for the problem of duplicate record detection. For approximate matching of data records, string matching algorithms (recursive algorithm with word base and recursive algorithm with character base) have been implemented and it is concluded that the results are much better with recursive algorithm with word base. Patrick Lehti and Peter Fankhauser [18] have presented an unsupervised, domain independent approach that starts with a broad alignment of potential duplicates, and analyses the distribution of observed distances among potential duplicates and among non-duplicates to iteratively refine the initial alignment.

Evaluations showed that the approach supersedes other unsupervised approaches and reaches almost the same accuracy as even fully supervised, domain dependent approaches. Duplicate detection is usually performed in two phases, an efficient blocking phase that determines few potential candidate duplicates based on simple criteria, followed by a second phase performing an in-depth comparison of the candidate duplicates. Patrick Lehti and Peter Fankhauser [19] have evaluated a precise and efficient approach for the blocking phase, which requires only standard indices, but performs as well as other approaches, based on special purpose indices, and outperforms other approaches based on standard indices. The key idea of the approach was to use a comparison window with a size that depends dynamically on a maximum distance, rather than using a window with fixed size. Mohammad Al Hasan et al. [6] have proposed, ROBIN, a method for initial seed selection in k-means types of algorithms. It imposes constraints on the chosen seeds that lead to better clustering when k-means converges. The constraints make the seed selection method insensitive to outliers in the data and also assist it to handle variable density or multi-scale clusters. Furthermore, the constraints make the method deterministic, so only one run suffices to obtain good initial seeds, as opposed to traditional random seed selection approaches that need many runs to obtain good seeds that lead to satisfactory clustering. They did a comprehensive evaluation of ROBIN against state-of-the art seeding methods on a wide range of synthetic and real datasets. Domenico Daniele Bloisi and Luca Iocchi et al. [21] have presented a clustering method based on k-means that has been implemented on a video surveillance system. Rek-means does not require specifying in advance the number of clusters to search for and is more precise than kmeans in clustering data coming from multiple Gaussian distributions with different covariances, while maintaining real-time performance. Experiments on real and synthetic datasets were presented to measure the effectiveness and the performance of the proposed method. Mingwei Leng et al. [22] have presented an algorithm, called an efficient k-means clustering based on influence factors, which is divided into two stages and can automatically achieve the actual value of k and select the right initial points based on the datasets characters. They presented influence factor to measure similarity of two clusters, using it to determine whether the two clusers should be merged into one. In order to obtain a faster algorithms theorem was presented and proofed, using it to accelerate the algorithm. Experimental results from Gaussian datasets were generated as in Pelleg and Moore, showed the algorithm has high quality and obtains a satisfying result. Amir Ahmad and Lipika Dey [23] have presented a clustering algorithm based on k-mean paradigm that works well for data with mixed numeric and categorical features. They proposed cost function and distance measure based on co-occurrence of values. The measures also take into account the significance of an attribute towards the clustering process. They presented a modified description of cluster center to overcome the numeric data only limitation of k-mean algorithm and provide a better characterization of clusters. The performance of the algorithm has been studied on real world data sets. K-means algorithm is the most popular partitional clustering algorithm; its fuzzy, rough, probabilistic and neural network is also popular. However, a major problem with the K-means algorithm and its variants is that they may not reach the globally optimal solution of the associated clustering problem. Genetic algorithms (GAs) are attractive to solve the partitional clustering problem. M. N. Murty et al. [24] have explained the GA based clustering approaches and presented an efficient scheme for clustering high-dimensional large scale data sets using GAs based on the well-known CF-Tree data structure. Most of the classical clustering algorithms are strongly dependent on, and sensitive to, parameters such as number of expected clusters and resolution level. To overcome this drawback, a Genetic Programming framework, capable of performing an automatic data clustering, was presented by I. De Falco et al. [25]. Moreover, a way of representing clusters which provides intelligible information on patterns was introduced together with an innovative clustering process. The effectiveness of the implemented partitioning system was estimated on a medical domain by means of evaluation indices. Qin Chen and Jinping Mo [26] have proposed an improved ant clustering algorithm based on K-means, which optimizes the rules of ant clustering algorithm. They decided the proper values of parameters Pdel and Iter by training the training datasets before the clustering process. Experimental results demonstrated that the proposed method has a good performance. In Dervis Karaboga and Celal Ozturk's [27] research, Artificial Bee Colony (ABC) is used for data clustering on benchmark problems and the performance of ABC algorithm is compared with Particle Swarm Optimization (PSO) algorithm and other nine classification techniques from the literature. 13 of typical test data sets from the UCI Machine Learning Repository were used to demonstrate the results of the techniques. The simulation results indicated that ABC algorithm can efficiently be used formultivariate data clustering.

IV. PROPOSED METHODOLOGY

The primary objective of my research is to devise and develop an effective approach for optimized accomplishment of clustering. The research has been inspired by the need for an efficient approach to improve the performance of the clustering approach and their associated applications. It is not an easy task if the collected data contains more noise and irrelevant items that affects the accuracy of the clustering approach. For this reason, I have intended to split up my research work into three phases: Data cleaning for detecting duplicate records, developing partition-based clustering approach and clustering assisted by evolutionary programming. Phase 1: The primary goal of this phase is to minimize the errors present in the database and to obtain

relevant and meaningful data that is well suited for performing the data mining task, namely, clustering. With this intention, I will devise an effective method for data cleaning, so as to improve the quality of the data. The proposed approach will be capable of providing accurate data records by removing some of the errors such as missing values, illegal values, inconsistent data and duplicate records, which usually arise when data is warehoused from external sources. Phase 2: In this phase, the refined data records will be grouped by performing a clustering process. I will devise an efficient approach for clustering which will first detect the duplicate records, so as to diminish the discovery of inaccurate and useless knowledge. Furthermore, the devised approach will be based on the well-known data analysis technique, namely, partitional clustering. The well-known partitional clustering techniques have some disadvantages which ultimately affect the performance of the clustering process. In order to prevent these drawbacks, Develop an efficient clustering approach which is an improvement of the traditional partitional clustering method. Phase 3: In this phase, I will extend the second phase by incorporating the evolutionary programming technique in the clustering method.

V. CONCLUSION

Generally, clustering of data records assisted by evolutionary programming technique is a promising research area for enhancing the performance and effectiveness of the clustering processes. So, by incorporating evolutionary programming technique, an effective clustering approach with improved efficiency and accuracy will be developed.

VI. REFERENCES

[1] Heiko Muller, "Semantic Data Cleansing in Genome Databases", Proceedings of the VLDB 2003 PhD Workshop Colocated with the 29th International Conference on Very Large Data Bases VLDB 2003, Vol: 76, 2003.

[2] Erhard Rahm, Hong Hai Do, "Data Cleaning: Problems and Current Approaches", IEEE Techn. Bulletin on Data Engineering, 2000.

[3] Timothy E. Ohanekwu, C.I. Ezeife, "A Token-Based Data Cleaning Technique for Data Warehouse Systems", IEEE Workshop on Data Quality in Cooperative Information Systems,Siena, Italy, January 2003.

[4] B. Delvin. "Data warehouse from architecture to implementation", Addison-Wesley, 1997.

[5] E. Simoudis and B. Livezey and R. Kerber, "Using Recon for Data Cleaning", In Proceedings of KDD 1995, pp. 282-287, 1995.

[6] Mohammad Al Hasan, Vineet Chaoji, Saeed Salem, and Mohammed J. Zaki, "Robust Partitional Clustering by Outlier and Density Insensitive Seeding", Pattern Recognition Letters, Vol: 30, No: 11, pp: 994-1002, 2009.

[7] S. P. Lloyd, "Lease square quantization in pcm", IEEE Transactions on Information Theory,Vol: 28, No: 2, pp. 129-136, 1982.

[8] J. B. MacQueen, "Some Method for Classification and Analysis of Multivariate Observations", Proc. of Berkeley Symp. on Mathematical Statistics and Prob., Berkeley, U. of California Press, Vol: 1, pp: 281-297, 1967.

[9] M. Ester, H. Kriegel, J. Sander, and X. Xu, "A Density-based Algorithm for Discovering Clusters in Large Spatial Databases with Noise", Proc. of KDD, 1996.

[10] A. K. Jain, and R. C. Dubes, "Algorithms for Clustering data", Pentice-Hill, 1988.

[11] C. J. Jardine, N. Jardine, and C. Sibson, "The structure and construction of Taxonomic Hierarchies", Math. Bio-science. Vol: 1, No: 2, pp. 173-179, 1967.

[12] J. Shi, and J. Malik, "Normalized Cuts and Image Segmentation", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol: 22, No: 8, 2000.

[13] I. K. Ravichandra Rao, "Data Mining and Clustering Techniques", DRTC Workshop on Semantic Web, Bangalore, 8th – 10th December, 2003.

[14] Aynur Dayanik, Craig G. Nevill-Manning, "Clustering in Relational Biological Data",

[15] ICML-2004 Workshop on Statistical Relational Learning and Connections to Other Fields, pp: 42-47, 2004.

[16] Pham, D.T. and Afify, A.A. "Clustering techniques and their applications in engineering", Proceedings- Institution of Mechanical Engineers Part C Journal of Mechanical Engineering Science, Vol: 221; No: 11, pp: 1445-1460, 2007.

[17] Rundensteiner. Elke, "Special Issue on Data Transformation", IEEE Data Engineering Bulletin, vol. 22, no.1, 1999.

[18] Rehman, M. Esichaikul, V., "Duplicate Record Detection for Database Cleansing", Second International Conference on Machine Vision, ICMV '09, pp: 333 - 338, Dec. 2009.

[19] Patrick Lehti and Peter Fankhauser, "Probabilistic Iterative Duplicate Detection", Lecture Notes in Computer Science, Springer, Berlin, Vol: 3761, pp: 1225-1242, 2005.

[20] Patrick Lehti, Peter Fankhauser, "A Precise Blocking Method for Record Linkage", Lecture Notes in Computer Science, Springer Berlin, Vol: 3589, pp: 210-220, 2005.

[21] Cheng-Ru Lin, Ming-Syan Chen, "Combining Partitional and Hierarchical Algorithms for Robust and Efficient Data Clustering with Cohesion Self-Merging," IEEE Transactions on Knowledge and Data Engineering, vol. 17, no. 2, pp. 145-159, Feb. 2005,

[22] Domenico Daniele Bloisi and Luca Iocchi, "Rek-Means: A k-Means Based Clustering Algorithm", Lecture Notes in Computer Science, Springer Berlin, Vol: 5008, pp: 109-118, 2008.

[23] Mingwei Leng, Haitao Tang, Xiaoyun Chen, "An Efficient K-means Clustering Algorithm Based on Influence Factors", Eighth ACIS International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing, Vol: 2, pp: 815 - 820, 2007.

[24] Amir Ahmad and Lipika Dey, "A k-mean clustering algorithm for mixed numeric andcategorical data", Data & Knowledge Engineering Vol: 63, No: 2, pp: 503-527, 2007.

[25] M. N. Murty, Babaria Rashmin, Chiranjib Bhattacharyya, "Clustering Based on Genetic Algorithms", Studies in Computational Intelligence, Springer Berlin, Vol: 98, pp: 137-159, 2008.

[26] I. De Falco, E. Tarantino, A. Della Cioppa and F. Fontanella, "An Innovative Approach to Genetic Programming-based Clustering", Advances in Soft Computing, Springer Berlin, Vol: 34, pp: 2006.

[27] Qin Chen, Jinping Mo, "Optimizing the Ant Clustering Model Based on K-Means Algorithm", World Congress on Computer Science and Information Engineering, Vol: 3, pp: 699 - 702, 2009.

[28] Dervis Karaboga, Celal Ozturk, "A Novel Clustering Approach: Artificial Bee Colony(ABC) Algorithm", Applied Soft Computing, 2009.

[29] Hernandez, M.A.; Stolfo, S.J., "Real-World Data is Dirty: Data Cleansing and the Merge/Purge Problem", Data Mining and Knowledge Discovery, Vol: 2, No: 1, pp: 9-37, 1998.

[30] Jarke, M., Lenzerini, M., Vassiliou, Y., Vassiliadis, P., "Fundamentals of Data Warehouses", Springer, 2000.

[31] Chaudhuri, S., Dayal, U., "An Overview of Data Warehousing and OLAP Technology", ACM SIGMOD Record, Vol: 26, No: 1, 1997.

[32] Monge, A. E. "Matching Algorithm within a Duplicate Detection System", IEEE Techn.Bulletin Data Engineering Vol: 23, No: 4, 2000.

[33] Hesam Izakian, Ajith Abraham, Vaclav Snasel, "Fuzzy Clustering Using Hybrid Fuzzy cmeans and Fuzzy Particle Swarm Optimization",

World Congress on Nature and Biologically Inspired Computing (NaBIC 2009), India, IEEE Press, pp. 1690-1694, 2009.
[34] Swagatam Das, Ajith Abraham, Amit Konar, "Automatic Clustering Using an Improved Differential Evolution Algorithm", IEEE Transactions on Systems, Man, and Cybernetics—Part A: Systems And Humans, Vol. 38, No. 1, 2008.
[35] E. W. Forgy, "Cluster analysis of multivariate data: Efficiency versus interpretability of classification," Biometrics, Vol. 21, No. 3, pp. 768–769, 1965.
[36] C. T. Zahn, "Graph-theoretical methods for detecting and describing gestalt clusters," IEEE Trans. Comput., Vol. C-20, No. 1, pp. 68–86, Jan. 1971.
[37] T. Mitchell, Machine Learning. New York: McGraw-Hill, 1997.
[38] J. Mao and A. K. Jain, "Artificial neural networks for feature extraction and multivariate data projection," IEEE Trans. Neural Netw., Vol. 6, No. 2, pp. 296–317, 1995.
[39] E. Falkenauer, "Genetic Algorithms and Grouping Problems", Chichester, U.K.: Wiley, 1998.
[40] Sandra Paterlini, Thiemo Krink, "High Performance Clustering with Differential Evolution", IEEE Congress on Evolutionary Computation, CEC 2004, Vol. 2. pp: 2004-2011,2004.
[41] Holland J.H., "Adaptation in Natural and Artificial Systems", University of Michigan Press, Ann Harbor, 1975.
[42] M. Steinbach, G. Karypis, V. Kumar, "A Comparison of Document Clustering Techniques", Text Mining Workshop, KDD, 2000.
[43] Y. Zhao and G. Karypis, "Empirical and Theoretical Comparisons of Selected Criterion Functions for Document Clustering", Machine Learning, Vol: 55, No: 3, pp. 311-331, 2004.
[44] J. A. Hartigan, "Clustering Algorithms", John Wiley and Sons, Inc., New York, NY,1975.
[45] Hui Xiong, Gaurav Pandey, Michael Steinbach, Vipin Kumar, "Enhancing Data Analysis with Noise Removal", IEEE Transactions on Knowledge and Data Engineering, Vol: 18, 2006.
[46] Ajith Abraham, Swagatam Das and Sandip Roy, "Swarm Intelligence Algorithms for Data Clustering", Soft Computing for Knowledge Discovery and Data Mining, Springer Verlag, Germany, pp. 279-313, 2007.
[47] Taher Niknam, Bahman Bahmani Firouzi and Majid Nayeripour, "An Efficient Hybrid Evolutionary Algorithm for Cluster Analysis", World Applied Sciences Journal, Vol: 4, No: 2, pp: 300-307, 2008.
[48]
[49] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth, "From Data Mining to Knowledge Discovery in Databases", AI Magazine, vol. 17, pp. 37-54, 1996.
[50] K. Yacoben and L. Carmichael, "Applying the Knowledge Discovery in Databases (KDD) Process to Fermilab Accelerator Machine Data"
[51] Frawley, W., Piatetsky-Shapiro, G., and Matheus, C. "Knowledge Discovery in Databases: An Overview". AI Magazine, pp: 57-70, 1992.
[52] Osmar R. Zaiane, "Chapter I: Introduction to Data Mining", CMPUT690 Principles of Knowledge Discovery in Databases, 1999.
[53] Youping Deng, Dheeraj Kayarat, Mohamed O. Elasri, Susan J. Brown, "Microarray Data Clustering Using Particle Swarm Optimization K-means Algorithm", Pattern Recognition and Machine Intelligence, Vol: 1, pp: 114-119, 2006.
[54] K. Krishna, and Murthi, "Genetic k-means algorithm", IEEE Transactions on systems, man, and cybernetics-Part B, 29, 433-4, 1999.
[55] G. P. Babu, "Connectionist and evolutionary approaches for pattern clustering," Dept. Comput. Sci. Automat., Indian Inst. Sci., Bangalore, Apr. 1994.
[56] R. W. Klein and R. C. Dubes, "Experiments in projection and clustering by simulated annealing," Pattern Recognit., vol. 22, pp. 213–220, 1989.
[57] S. Z. Selim and K. Alsultan, "A simulated annealing algorithm for the clustering problem," Pattern Recognit., vol. 10, no. 24, pp. 1003–1008, 1991.
[58] G. P. Babu and M. N. Murty, "Simulated annealing for selecting initial seeds in the k-means algorithm," Ind. J. Pure Appl. Math., vol. 25, pp. 85–94, 1994.
[59] J. N. Bhuyan, V. V. Raghavan, and V. K. Elayavalli, "Genetic algorithm for clustering with an ordered representation," in Proc. 4th Int. Conf. Genetic Algorithms. San Mateo, CA: Morgan Kaufman, 1991.
[60] D. R. Jones and M. A. Beltramo, "Solving partitioning problems with genetic algorithms," in Proc. 4th Int. Conf. Genetic Algorithms. San Mateo, CA: Morgan Kaufman, 1991.
[61] G. P. Babu and M. N. Murty, "A near-optimal initial seed selection in K-means algorithm using a genetic algorithm," Pattern Recognit. Lett., vol. 14, pp. 763–769, 1993.
[62] Phanendra Babu G., Narasimha Murty M. "Clustering with evolution strategies," Pattern Recognition., vol. 27, no. 2, pp. 321–329, 1994.
[63] Kimball, Ralph; Joe Caserta, "The data warehouse ETL toolkit: Practical techniques for extracting, cleaning, conforming, and delivering data", John Wiley & Sons, 2004.