Journal of Advanced Computing and Communication Technologies *(ISSN:* 2347 - 2804*)*
Volume No.4 Issue No.3. July 2016

Swarm Intelligence Technique on Ant Colony Optimization Cluster Algorithm

By

A.Haritha Deepthi, V.Arun, A.Mohan Kumar
Department of Information Technology,
PSG PolytechnicCollege, Coimbatore, India
harithadeepthibtech@gmail.com, itarunpsg@gmail.com, mohanshankercbe@gmail.com

ABSTRACT

This project proposes an swarm intelligence technique on ant colony optimization cluster algorithm. It introduces new adjustment process, which adjusts the cluster generated by the carry process iteratively. Cluster analysis is a data mining technology designed to derive a good understanding of data to solve clustering problems by extracting useful information from a large volume of mixed data elements. This algorithm can improve the efficiency and the convergence of the cluster theoretically. Ant colony algorithm can form more accurate and stable clusters than the K-modes Algorithm, Information Entropy-Based Cluster Algorithm and LF Algorithm Scalability experiments show that the running time has an obvious linear relationship with the size of data set. This Algorithm can handle large datasets. The system aim is to increase the classification performance through the Ant Colony Optimixation Algorithm and to decrease the false alarm rate.

KEYWORDS

Data Mining, Ant clustering, Swarm intelligence, Clustering

1. INTRODUCTION

Cluster analysis is a data mining technology designed to derive a good understanding of data to solve clustering problems by extracting useful information from a large volume of mixed data elements. Recently researchers have aimed to derive clustering algorithms from nature's swarm behaviors. Ant-based clustering is an approach inspired by the natural clustering and sorting behavior of ant colonies. In this research an optimized ant-based clustering method is presented with new modifications to the original ant colony clustering model ACC to enhance the operations of ants, picking up and dropping off data items. Data experiments show that the optimized ant colony cluster algorithm can form more accurate and stability clusters than the FRECCA algorithm. Scalability experiments show that the running time has an obviouslinear relationship with the size of data set. This algorithm can handle large dataset more rapidly, accurately and effectively, which helps to maintain the good scalability at the same time.

1.1. Datamining

Generally, data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, orboth. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases.

1.2 Working of Datamining

While large-scale information technology has been evolving separate transaction and analytical systems, data mining provides the link between the two. Data mining software analyzes relationships and patterns in stored transaction data based on open-ended user queries. Several types of analytical software are available: statistical, machine learning, and neural networks. Generally, any of four types of relationships are sought:
Classes: Stored data is used to locate data in predetermined groups. For example, a restaurant chain could mine customer purchase data to determine when customers visit and what they typically order. This information could be used to increase traffic by having daily specials.

Clusters: Data items are grouped according to logical relationships or consumer preferences. For example, data can be mined to identify market segments orconsumer affinities.

Associations: Data can be mined to identify associations. The beer-diaper example is an example of associative mining.

Sequential patterns: Data is mined to anticipate behavior patterns and trends. For example, an outdoor equipment retailer could predict the likelihood of a backpack being purchased based on a consumer's purchase of sleeping bags and hiking shoes.

1.3 Clustering

Clustering is concerned with the Division of data into homogenous subgroups. Informally, the Objective of this division is twofold: data items within one cluster are required to be similar to each other, while those within different clusters should be dissimilar. Problems of this type arise In a Variety of disciplines ranging from sociology and psychology to commerce, biology and Computer science, and algorithms for tackling them continue to be the subject of active research. Consequently, there Exists a multitude Of clustering methods, which Differ Not only in the principles of the algorithm used (which of course determine runtime Behavior and scalability), But also in Many of their most basic properties, such as the data

handled (numerical vs. Categorical and proximity data), assumptions on the shape Of the clusters (e.g., spherically shaped), the form of the final partitioning (hard vs. Fuzzy assignments) or the parameters that Have to be Provided (e.g., the correct number of clusters).The four main Classes of clustering algorithms Available in The literature Are partitioning methods ,Hierarchical methods, density-based clustering and grid-based clustering For the purpose of our Comparative study we select Two of the most popular and well-studied algorithms From the literature:-means, a
Representative of the class of partitioning methods, And agglomerative average link clustering, which is a hierarchical approach. Additionally, we compare against one-dimensional self-organising maps.

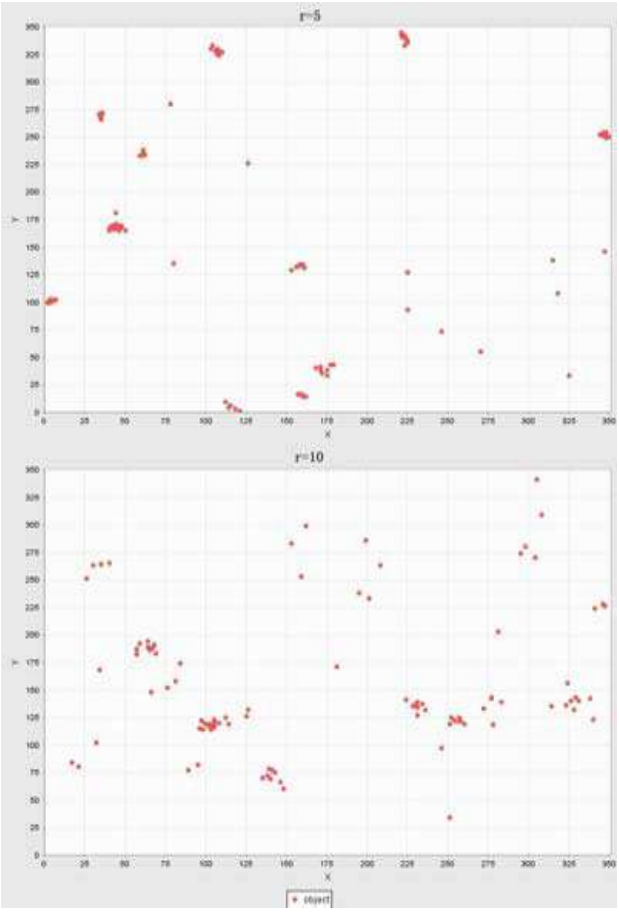


Fig 1: Data that are formed into clusters

1.4 Advantages and Disadvantages of Cluster

Open space can provide community members with larger recreation areas and create a sense of openness that many people desire. Open space can benefit the environment by providing habitat for wildlife, naturally filtering stormwater, reducing storm water runoff from impervious surfaces, and protecting the natural features of a site. Linking the open space of several conservation design subdivisions can help develop larger and more effective “environmental corridors” within and between communities. Developers may benefit because these designs usually reduce the costs of site development and increase the market price of individual plots in comparison with traditional subdivisions. These designs can benefit rural areas by reinforcing the policy of maintaining the local rural character that is included in many comprehensive land use plans.
Perhaps most important, local officials, developers, and the community may be predisposed toward traditional development designs because they are familiar and well understood. An education effort may be necessary to help these groups understand the goals and advantages of cluster development. During the planning phases, lot and home layout may take extra work to ensure that while homes are Cluster/Conservation Development - Land Use Planning - Local Community located closer together, they still take advantage of the open-space goals of the design. Methods to protect and maintain the open space must be carefully developed, implemented, and monitored. Although not necessarily a restricting disadvantage, the management of waste water must be carefully designed for smaller lots. While these disadvantages should be acknowledged and addressed, none should preclude the use of cluster development.

2. SWARM INTELLIGENCE

Swarm intelligence (SI) is the collective behavior of decentralized, self-organized systems, natural or artificial. The concept is employed in work on artificial intelligence. The expression was introduced by Gerardo Beni and Jing Wang in 1989, in the context of cellular robotic systems**.**

SI systems consist typically of a population of simple agents or boids interacting locally with one another and with their environment. The inspiration often comes from nature, especially biological systems. The agents follow very simple rules, and although there is no centralized control structure dictating how individual agents should behave, local, and to a certain degree random, interactions between such agents lead to the emergence of "intelligent" global behavior, unknown to the individual agents. Examples in natural systems of SI include ant colonies, bird flocking, animal herding, bacterial growth, fish schooling and Microbial intelligence. The definition of swarm intelligence is still not quite clear. In principle, it should be a multi-agent system that has self-organized behavior that shows some intelligent behavior.
The application of swarm principles to robots is called swarm robotics, while 'swarm intelligence' refers to the more general set of algorithms. 'Swarm prediction' has been used in the context of forecasting problems.Swarm Intelligence-based techniques can be used in a number of applications. The U.S. military is investigating swarm techniques for controlling unmanned vehicles. The European Space Agency is thinking about an orbital swarm for self-assembly and interferometry. NASA is investigating the use of swarm technology for planetary mapping. A 1992 paper by M. Anthony Lewis and George A. Bekey discusses the possibility of using swarm intelligence to control nanobots within the body for the purpose of killing cancer tumors. Conversely al-Rifaie andAber have used Stochastic Diffusion Search to help locate tumours. Swarm intelligence has also been applied for data mining.

3. RELATED WORKS

Ant Colony Optimization (ACO) is a population-based approach for solving combinatorial optimization problems that is inspired by the foraging behavior of ants and their inherent ability to find the shortest path from a food source to their nest. ACO is the result of research on computational intelligence approaches to combinatorial optimization. ACO Algorithm has been applied to a broad range of hard combinatorial problems. Among them, we have the classic Traveling Salesman Problem (TSP), where an individual must find the shortest route by which to visit a given

number of destinations.

The original AntNet algorithm, Ants are only capable of simple stochastic decisions influenced by the availability of previously laid stigmergic trails. The chemical denoting a stigmergic trail is subject to decay over time and reinforcement proportional to the number of ants taking the same path. Trail building is naturally a bidirectional process, ants need to reach the food and make a successful return path, in order to significantly reinforce a stigmergic trail. Moreover, the faster the route, the earlier the trail is reinforced. An ant on encountering multiple stigmergic trails will probabilistically choose the route with the greatest stigmergic reinforcement. Naturally, this will correspond to the 'fastest' route to the food

4. ANTCOLONY OPTIMIZATION ALGORITHM

An Ant Colony Optimization algorithm ACO is essentially a system based on agents which simulate the natural behavior of ants, including mechanisms of cooperation and adaptation. The fundamental approach underlying ACO is an iterative process in which a population of simple agents repeatedly construct candidate solutions; this construction process is probabilistically guided by heuristic information on the given problem instance as well as by a shared memory containing experience gathered by the ants in previous iteration. In the use of this kind of system as a new meta-heuristic was proposed in order to solve combinatorial optimization problems. This new meta-heuristic has been shown to be both robust and versatile in the sense that it has been successfully applied to a range of different combinatorial optimization problems.

4.1. System Analysis

System Design is the process of making the newly designed system fully operational and consistent in performance. The following steps have been followed in the implementation of the system. Implementation in planning and User Training. As the part of implementation, the system is taken the site and loaded on to client's computer. Some of the user's level, exposure to computer etc. These users are trained first and they run the system for a month. A detailed documentation is prepared for the employees and they trained to access the software. These users are trained first and they can run the system for a month. After installation of software, the hardware specifications are checked. If hardware specifications are satisfactory, then the software is loaded for pilot run. User training starts at this time itself. Users will be given a user manual, which documents how to use the system and all the exception handling procedures.

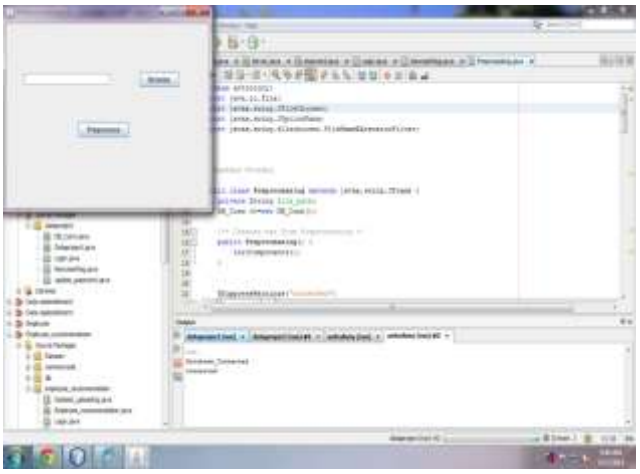


Fig 2: Shows the process for entering the existing data into the database

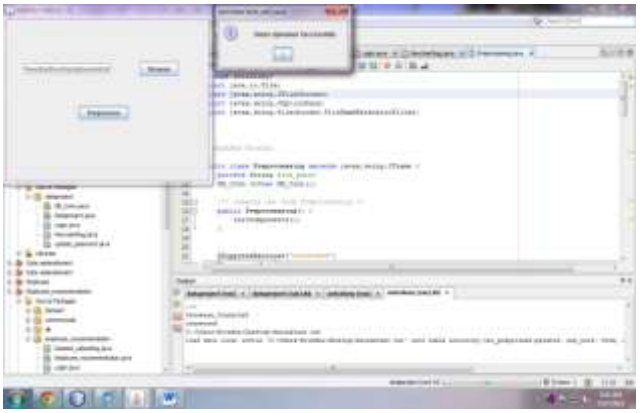


Fig 3: Notifications for uploading the data into the dataset is displayed

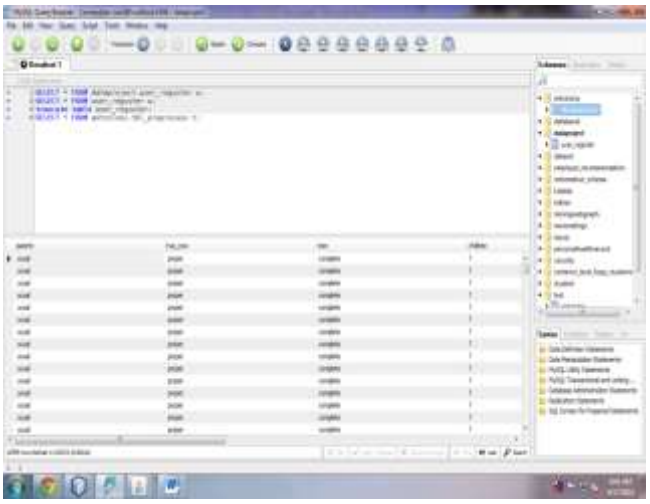


Fig 4: The database where the entered datas are stored this are used for further process

4.2. Input and Output Design

Input design is the part of overall system design which requires very careful attention. Often the collection of input data is the most expensive part of the system, in terms of both the equipment used and the number of people involved; it is the

point of most contact for the users with the computer system; and it is prone to error. If data going into the system are incorrect, then the processing and output will magnify these errors. In this system inputs are given in two ways, the Existing users can directly enter into the system using login form, and new users have to register all their details in the registration form provided.Input design is the very important part in the project and should be concentrated well as it is prone to error. The data that are to be inserted are to be inserted with care as this plays a very important role. In order to get the meaningful output and to achieve good accuracy the input should be acceptable and understandable by the user. Output design plays a very important role in a system. Getting a correct output is a task that has to be concentrated, as a system is validated as a correct one only if it gives the correct output according to the input. Here in this project in all the three days of inductions if the employee has completed all his/her input, then the output shows the status as completed or his status will be pending.

5. TEST PLAN

Testing is a process of checking whether the developed system is working according to the original objectives and requirements.Testing is a set of activities that can be planned in advance and conducted systematically. Testing is vital to the success of the system. System testing makes a logical assumption that if all the parts of the system are correct, the global will be successfully achieved. In adequate testing if not testing leads to errors that may not appear even many months. This creates two problems, the time lag between the cause and the appearance of the problem. The effect of the system errors on the files and records within the system. A small system error can conceivably explode into a much larger problem. Effective testing early in the purpose translates directly into long term cost savings from a reduced number of errors. Another reason for system testing is its utility, as a user-oriented vehicle before implementation. The best programs are worthless if it produces the correct outputs. No other test can be more crucial. Following this step, a variety of tests are conducted.

5.1. Unit Testing

A program represents the logical elements of a system. For a program to run satisfactorily, it must compile and test data correctly and tie in properly with other programs. Achieving an error free program is the responsibility of the programmer. Program testing checks for two types of errors: syntax and logical. Syntax error is a program statement that violates one or more rules of the language in which it is written. An improperly defined field dimension or omitted keywords are common syntax errors. These errors are shown through error message generated by the computer. For Logic errors the programmer must examine the output carefully. When a program is tested, the actual output is compared with the expected output. When there is a discrepancy the sequence of instructions must be traced to determine the problem. The process is facilitated by breaking the program into self-contained portions, each of which can be checked at certain key points .The idea is to compare program values against desk-calculated values to isolate the problems.

Test case no	Description	Expected result
1	Test for application window properties	All the properties of the windows are to be properly aligned and displayed
2	Test for mouse operations	All the mouse operations like click, drag, etc. must perform the necessary operations without any exceptions

5.2. Functional Testing

Functional testing of an application is used to prove the application delivers correct results, using enough inputs to give an adequate level of confidence that will work correctly for all sets of inputs. The functional testing will need to prove that the application works for each client type and that personalization function work correctly

Test case no	Description	Expected result
1	Test for all peers	All peers should communicate in the group.
2	Test for various peer in a distributed network framework as it display all users available in the group	The result after execution should give the accurate result.

5.3. Non-Functional Testing

This testing used to check that an application will work in the operational environment. On-functional testing includes: Load testing, Performance testing, Usability testing, Reliability testing, and Security testing

5.4. Load Testing

Test case no	Description	Expected result
1	It is necessary to ascertain that the application behaves correctly under loads when 'Server busy' response is received.	Should designate another active node as a Server.

5.5. Performance Testing

Test case no	Description	Expected result
1	This is required to assure that an application perforce adequately, having the capability to handle many peers, delivering its results in expected time and using an acceptable level of resource and it is an aspect of operational management.	Should handle large input values, and produce accurate result in a expected time

5.6. Reliability Testing

Test case no	Description	Expected result
1	This is to check that the server is rugged and reliable and can handle the failure of any of the components involved in provide the application.	In case of failure of the server an alternate server should take over the job

5.7. Security Testing

It is necessary to check that the application's data is secured.

Test case no	Description	Expected result
1	Checking that the user identification is authenticated	In case failure it should not be connected in the framework
2	Check whether group keys in a tree are shared by all peers	The peers The peers should know group key in the same group

5.8. White Box Testing

White box testing, sometimes called glass-box testing is a test case design method that uses the control structure of the procedural design to derive test cases. Using white box testing method, the software engineer can derive test cases

Test case no	Description	Expected result
1	Exercise all logical decisions on their true and false sides	All the logical decisions must be valid
2	Execute all loops at their boundaries and within their operational bounds.	All the loops must be finite
3	Exercise internal data structures to ensure their validity.	All the data structures must be valid

.

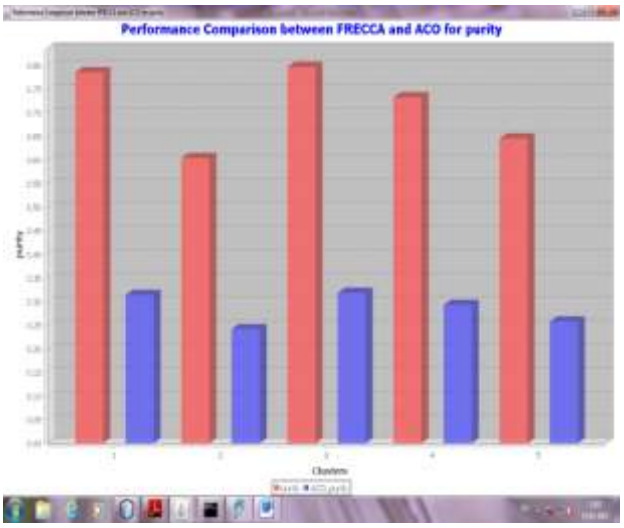
5.9. Black Box Testing

Black box testing, also called behavioral testing, focuses on the functional requirements of the software. That is, black testing enables the software engineer to derive sets of input conditions that will fully exercise all functional requirements for a program. Black box testing is not alternative to white box techniques. Rather it is a complementary approach that is likely to uncover a different class of errors than white box methods. Black box testing attempts to find errors in the following categories

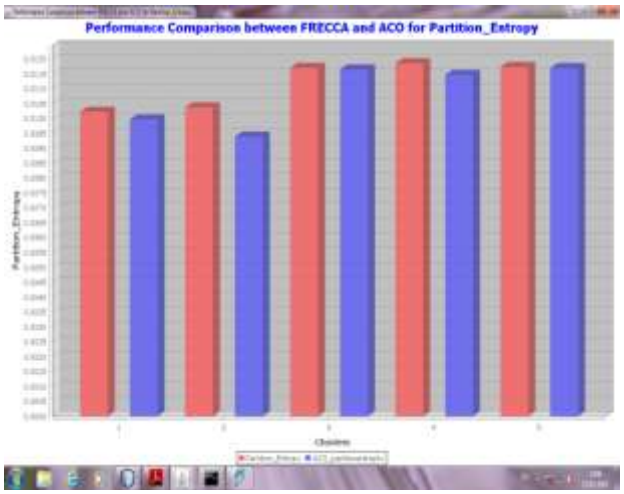
Test case no	Description	Expected result
1	To check for incorrect or missing functions	All the functions must be valid
2	To check for interface errors	All the interface must function normally
3	To check for errors in a data structures or external data base access.	The database updation and retrieval must be done
4	To check for initialization and termination errors.	All the functions and data structures must be initialized properly and terminated normally

6. Graphical Comparison

This following graph shows the comparison between FRECCA algorithm and ACO algorithm for purity and partition entropy respectively.



a) Purity



b) Partition entropy

7. CONCLUSION

In this paper, a new Ant Colony Optimization based algorithm, termed Optimized Ant Colony Clustering Algorithm (OACA), is proposed for data clustering. We proposed new techniques to increase the working performance of the ant colony optimization algorithm .OACA extends Ant Colony Optimization through the use of a quadratic metric, the Sum of K Nearest Neighbor Distances metric, together with constrained addition of pheromone and shrinking range strategies to better partition data sets.

8. REFRENCES

[1] M. H. Dunham, Data Mining: Introductory and Advanced Topics, Prentice Hall, 2003

[2] A. A. Freitas, S. H. Lavington, Mining very large databases with parallel processing. Dordrecht, the Netherlands, Kluwer Academic Publishers, 1998

[3] I. Foster, C. Kesselman, the Grid: Blueprint for a New Computing Infrastructure. Morgan Kaufmann, Elsevier Press, 2004, pp. 593-620

[4] T. G. Dietterich, An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting and randomization. Machine Learning Vol.40, 2000, pp.139-158

[5] Cheng-Fa Tsai, Chun-Yi Sung, "DBSCALE: An Efficient Density Based Clustering Algorithm for Data Mining in Large Databases" (PACCS 2010) Second Pacific-Asia Conference on Circuits, Communications and System, 91201 Pingtung, Taiwan, October 2010.

[6] R. Agrawal, J. C. Shafer, Parallel mining of association rules IEEE Transactions on Knowledge and Data Engineering, Vol 8., 1996, pp.962-969

[7] E. Januzaj, H-P. Kriegel, M. Pfeifle, DBDC: Density-Based Distributed Clustering Proc. 9th Int. Conf. on Extending Database Technology(EDBT), Heraklion, Greece 2004, pp. 88-105