IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 7, Issue 1, January 2018

Design and Implementation of Clustering Algorithm based on Multi-Swarm Intelligence

Priyanka Panwar[1], Amit Vajpayee[2], Harish Patidar[3]

PG Student, Computer Science Department, LNCT, Indore, India [1]
Assistant Professor, Computer Science Department, LNCT, Indore, India [2]
HOD, Computer Science Department, LNCT, Indore, India [3]

Abstract: Clustering remains an active field of different domain research currently. No single algorithm is identified, which can group every real world datasets resourcefully and without error. Our proposed technique be different from the classical ant system in the sense that here the pheromone trail are simplified in behaviour. The capability of Multi Swarm Optimization (MSO), heuristic technique for search of optimal solutions based on the perception of swarm, to powerfully face classification of multiclass database instances. it is evident that the proposed method outperforms other methods compared. The experimental result show the proposed technique is efficient.

Keywords: Ant Colony Clustering Algorithm, Swarm Intelligence Algorithm, multi Swarm Optimization.

I. INTRODUCTION

Clustering is a technique based on unsupervised learning, which separate datasets into a number of dissimilar sets according to positive criterion. through the development of information technology, clustering has been useful to a lot of fields such as machine learning, data mining and so on Data clustering is the procedure of grouping data into a numeral of clusters. Clustering, which plays an significant role in data mining, intend at grouping a set of data into two or additional mutually exclusive unknown groups. The objective of data clustering is to create the data in the similar cluster contribute to a high degree of resemblance while the data in dissimilar cluster being extremely dissimilar to data from other clusters. Clustering algorithms have been useful to a wide range of problems, such as data mining [1], data analysis, and pattern recognition and image segmentation. The major four kind of clustering algorithms are partitioning process, hierarchical technique; density based clustering and grid-based clustering. Partitioned clustering algorithm divide data records into a predefined number of clusters by optimizing a quantity of convinced criterion. In the field of clustering, K-means algorithm is the generally popularly used algorithm to discover a partition that minimizes mean square error (MSE) compute. In the past three decades, K-means clustering Algorithm has been use in a variety of domains. Though, K-means algorithm is sensitive to the initial states and forever converges to the local optimum solution. In order to conquer this problem, a lot of process have been proposed. Over the previous decade, additional and more stochastic, population based optimization algorithms have been functional to clustering problems. For instance, an evolutionary algorithm based on ACO algorithm for clustering problem has been introduce in [1] and presents PSO to resolve the clustering problem. In during the research on the particle swarm optimization with compression factor, the algorithm is enhanced by configuring the optimal parameters and controlling convergence rate, which can successfully get better accuracy and the global convergence of MSO. In [2], the thought of genetic algorithm is use to build the objective function of k-means clustering algorithm. This customized clustering algorithm collective with swarm intelligence obtains enhanced consequence in practical application. Initially, when ants build a tour they close modify the amount of pheromone on the visit edges by a local updating role. Secondly, subsequent to each the multi swam have build their individual tours, a global updating rule is useful to modify the pheromone level on the limits that fit in to the best data classification results. The rest of the paper is prepared as follows. The rest of the paper is organized as follows. In Section II and discuss the PSO algorithm and related work in Section. The details of data clustering algorithm based on PSO will be given in section III. Section IV corresponds to the clustering algorithms conclusions.

II. RELATED WORK

The perception of Particle Swarms, though initially introduce for simulate human social behaviours, has turn into very popular these days as a creative every and optimization technique. Xiaojie Zhu et al[1] In this paper, to afford an successful method to complete resourceful multi-keyword ranked search in excess of encrypted cloud data. Meanwhile, an algorithm HCSF is intended to apply this way. In the index building phase of HCSF, utilize swarm intelligence to create initial clusters and recommend fuzzy k-means to process the initial clusters. in addition, a split

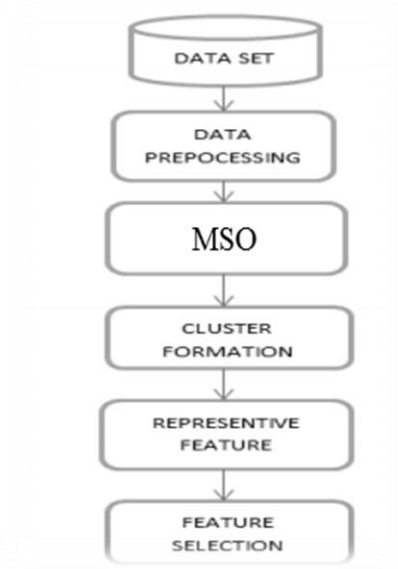
IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319-5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 7, Issue 1, January 2018

method is proposed to control the size of clusters and a hierarchical structure is construct to organize documents by combine the split mechanism. Antim Jaiswal1 et al[2]challenge for enhancement and innovation in the rising field of genome sequence alignment. Algorithms for position and mapping have to correct to accommodate continually altering genomic input data. One significant piece the conniving of novel alignment algorithm which will be based on Map Reducer like and Hive, new succession mapping algorithm should be intended with the help of last Map Reducers like Jqal. Qinghua Lu t et al[3] The giving of this work is mostly two fold. primary, a appropriate estimation component is proposed to predict the performance of Hadoop clusters when executing dissimilar jobs which can be used by GAs. instant, with the efficient information that the evaluation module give, a genetic algorithm based job scheduling replica for geo-distributed data are put forward. D. Asir Antony Gnana Singh et al[4]proposed a genetic algorithm support wrapper feature selection for medical data classification. The research is conduct with four experimental strategies. The proposed system contains the genetic algorithm (GA) to search and outline the feature subsets and the Naive Bayes classifier as the assessment tool to choose the important feature subset. The performance of the proposed process is evaluated with the well traditional classifiers such as Naive Bayes.

Dr.S.Santhosh Baboo et al[5] The improved k-means algorithm engaged additional steps but consequence is accurate algorithmic finishing stage; iteration also frequent extremely minimum times. The concluding discussion of this gather average centroid clustering for every previously selected values and present selected clustering data. The consequence of this gave the comparative study of the k-means, improved k means algorithms and AC clustering values.

III. PROPOSED METHODOLOGY

In this paper we studied the dissimilar clustering technique and the algorithms based on these methods. Clustering is extensively used in a number of applications. Every clustering technique is having its possess pros and cons. it is clear that none of the clustering algorithms converse, perform well for all the leading factors.PSO (partitioning based) is the simplest of every the algorithms. But its utilize is classified to numeric data values simply. The presentation of the PSO algorithm increases with the boost as the number of clusters increase. Our proposed clustering algorithm form nested clusters by split or integration of data points based technique is considered for building clusters of random shapes. It build clusters automatically no require to talk about the number of clusters and naturally removes outliers. To develop a collective model using clustering and categorization concepts. To put together the developed model inside parallel programming architecture of Map reduces. Performance optimization of data analysis with the proposed technique Clustering consequence depends on two facet inner distance and outer distance. The less significant inner distance and larger outer distance guide to enhanced effect. as well, accuracy is an important evaluation index since we can leverage original standard classification of dataset to compute accuracy value. Allowing for above every, we moderator the performance of PSO. we propose has enhanced clustering performance when it processed dataset. Show in figure demonstrates when fitness value is improved with iterations, the modified algorithm requirements to iterate less times than others in the case of the similar fitness value. And it as well illustrate that when iterative process is up to meet, compare the convergence value, fitness value of advanced algorithm is better than others. The result illustrate the dataset is clustered additional efficiently and precisely than traditional clustering algorithms.



Working of proposed algorithm

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 7, Issue 1, January 2018

Step 1
Procedure for PSO
Input: Randomly initialized position and velocity of the particles: Xi(0) and Vi(0)
Output: Position of the approximate global optima X¤
1: while terminating condition is not reached do
2: for
i = 1 to number of particles do
3: Evaluate the ¯tness: =f(Xi(t));
4: Update P(t)and g(t);
5: Adapt velocity of the particle using Equation 2;
6: Update the position of the particle;
7: end for
8: end while

a) Data pre-processing: afraid with redundant feature removal and judgment applicable feature to the target class.
b) Construction of minimum PSO for the build for the data set subsequent to pre-processing it.
c) construction of the clusters of the features.
d) Selection of features which are additional relevant to the target class.

Since of altering updating rule of data classification using clustering, in addition, owing to simulated thought applied to modernize the thought adjusts parameters used for learning and updating and the consequence lead to less iterations and run time. More than every, the advanced algorithm is superior to traditional clustering algorithm in comprehensive data classification quality and efficiency. In appropriate features and redundant features involve the accuracy of the learning machines. The feature selection with clustering includes following steps. Our proposed novel clustering technique for feature selection from big data. The configuration of clusters reduces the dimensionality and assist in selection of the significant features for the target class. The data pre-processing concerned removes the redundant and inappropriate features. The formation of clusters find from minimum error rate reduces the complexity for the working out of feature selection. The proposed feature selection technique forms clusters of features and a representative feature can be selected from it but this does not assurance that the exacting feature from the subset is additional applicable to the target class. The classification of features which belong to exacting class can be complete in this case. Thus, the feature subset obtain from can be specified to Supervised Learners for classification purpose. Ensemble technique can be used for voting that give the most excellent feature for target appropriate class. The major advantage of feature selection is that the distinctiveness of the selected features can give insights into the nature of the problem at hand. Consequently, the feature selection is an significant step in resourceful learning of large multi-featured datasets.
to perform the experiment 4GB RAM 120GB Space . to used the simulation tools Weka Tools.

IV. "RESULT AND ANALYSIS"

In this research to use for performance and investigational analysis of the proposed algorithm implemented. Valuation Metric the algorithm instigated was assessed on the subsequent metrics Accuracy ,Precision ,Recall ,F measure

Accuracy: Accuracy rate or percent correct is distinct as the quantity of correct cases separated by the complete number of cases.

Precision: it is used for retrieved instances that are applicable or it is the percentage of certain items that are correct Recall: it is applicable instances that are recovered or it is the percentage of accurate items that are selected.

F Measure: A metric that associations precision and recall metrics, it is the weighted can be measured as a collective measure that measures the precision recall trade off.

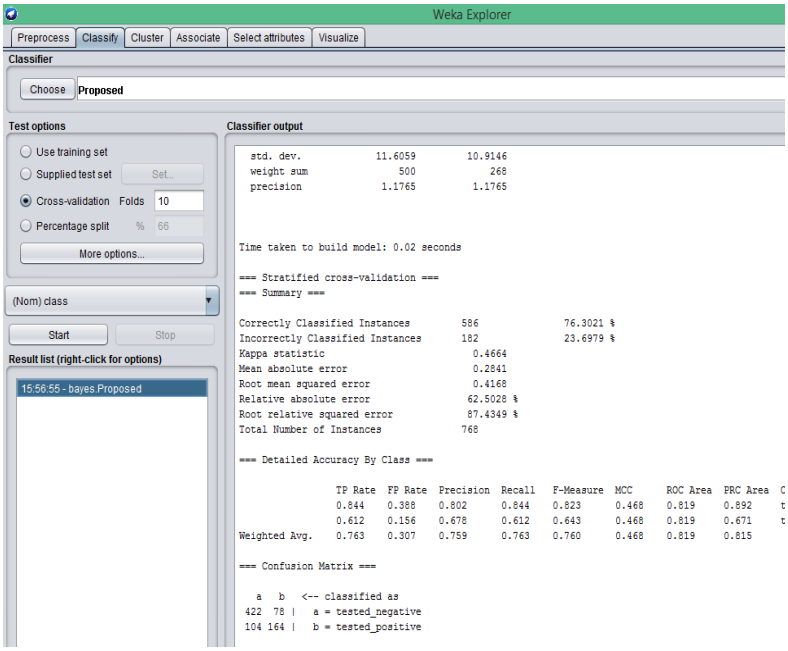
To used different the formulas discoursed below.
TP Is A TRUE POSITIVE Correct Result
FN is a false negative Missing result
FP is a false positive unexpected result
TN is a TRUE NEGATIVE Correct absenteeism of result
Accuracy = (TP + TN) / (TP + FP + FN + TN)
Precision = TP / (TP + FP)
Recall = TP / (TP + FN)
F1 = 2 * P * R / (P + R)

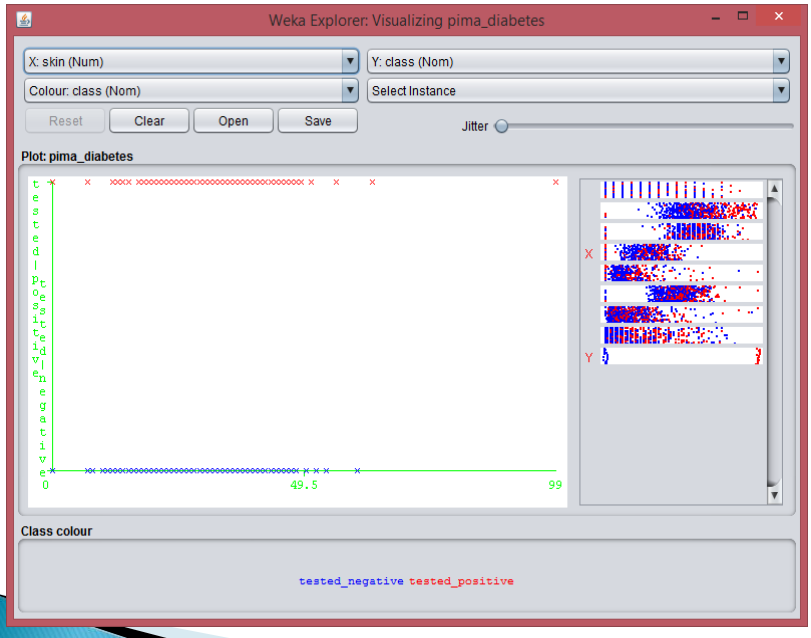
IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 7, Issue 1, January 2018

Correctly classified instances=2185
52.3228 % Correct
In Correctly classified instances=1991
47.6772%
Root mean Squared error =0.4694
Total number of instances=4176

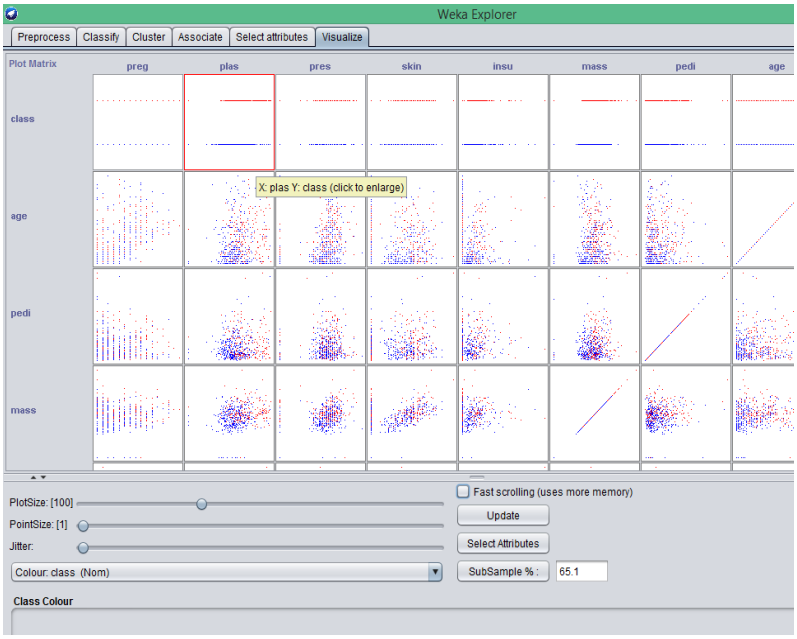




IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 7, Issue 1, January 2018



Number of Experiment	Proposed Approach	Precision	Recall	F-score
1	0.802	0.651	0.41	0.749
2	0.678	0.54	0.51	0.524
3	0.759	0.628	0.524	0.708
4	0.503	0.404	0.401	0.402
5	0.526	0.401	0.303	0.301

Table 1: Tabular form of Accuracy

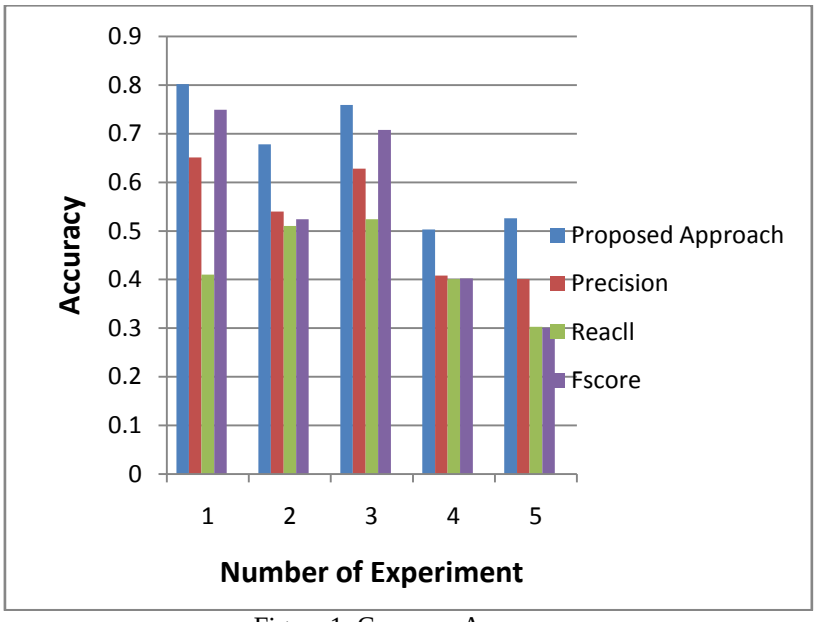


Figure 1: Compare Accuracy

DOI 10.17148/IJARCCE.2018.7142

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319-5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 7, Issue 1, January 2018

Number of Experiment	Proposed Approach	Precision	Recall	F-score
1	0.3	0.455	0.729	0.560
2	0.31	0.481	0.171	0.252
3	0.33	0.190	0.234	0.211
4	0.2	0.333	0.308	0.320
5	0.32	0.571	0.889	0.696

Table 2: Tabular form of Time Values

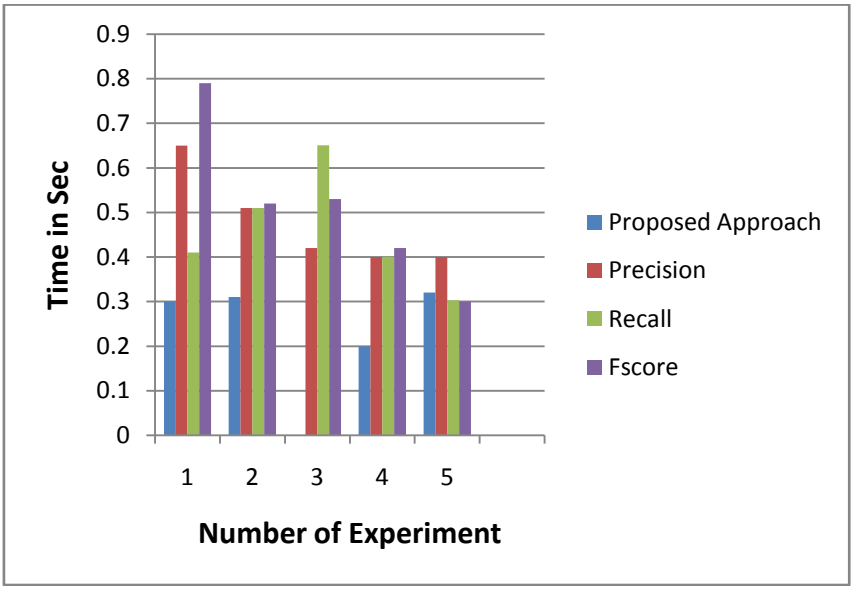


Table 2: Compare Time consumption

time performance of: time vs. Initial sub sets the results through the graph to through the experiment On the other hand, as the cluster size enlarge, the classification for smaller data sets strength underutilize the cluster resources. This result indicate that there is a threshold above which increasing the number of nodes does not create considerable performance gains.

V. CONCLUSION

In this paper, design and implementation of clustering algorithm based on multi-swarm intelligence introduced a quantity of of the preliminary concepts of Swarm Intelligence (SI) with an importance on PSO and MSO. Then the essential data clustering terminologies are introduced. In this paper, we as well illustrated a quantity of of the past and ongoing works, which be relevant dissimilar hadoop tools to clustering problems. As a result, the our proposed algorithms can productively be applied to data clustering with performance and efficiency. There are several issues remaining as the scopes for future studies such as with MSO algorithms in data stream clustering.

REFERENCES

[1] Xiaojie Zhu, Chi Chen ,XueTian, Jiankun Hu," HCSF: A hierarchical clustering algorithm based on swarm intelligence and fuzzy logic for cipher text search" 978-1-4799-8389-6/15/ IEEE -2015.
[2] Antim Jaiswal1, Arvind Upadhyay2 "An Enhanced Framework of Genomics Using Big Data Computing" IEEE International Conference on Computer, Communication and Control (IC4-2015).
[3] Qinghua Lu, ShanshanLi,Weishan Zhang," Genetic Algorithm based Job Scheduling for Big Data Analytics" International Conference on Identification, Information, and Knowledge in the Internet of Things 2015.
[4] D. Asir Antony Gnana Singh , E. JebamalarLeavline, R. Priyanka and P. Padma Priya," Dimensionality Reduction using Genetic Algorithm for Improving Accuracy in Medical Diagnosis" I.J. Intelligent Systems and Applications, 2016, 1, 67-73 Published Online January 2016 in MECS (http://www.mecs-press.org/) DOI: 10.5815/ijisa.2016.01.08.
[5] Dr.S.SanthoshBaboo,K.tajudin" Clustering Centroid Finding Algorithm (CCFA) using Spatial Temporal Data Mining Concept" Proceedings of the 2013 International Conference on Pattern Recognition, Informatics and Mobile Engineering (PRIME) February 21-22.

[5] Zhang, Zili, and Pengyi Yang, "An ensemble of classifiers with genetic algorithm Based Feature Selection," IEEE intelligent informatics bulletin, vol. 9, pp. 18-24, 2008.

[6] Zhuo, Li, Jing Zheng, Xia Li, Fang Wang, Bin Ai, and Junping Qian. "A genetic algorithm based wrapper feature selection method for classification of hyper spectral images using support vector machine," Geo informatics 2008 and Joint Conference on GIS and Built Environment: Classification of Remote Sensing Images, pp. 71471J71471J, 2008.

[7] Aziz, Amira Sayed A., Ahmad Taher Azar, Mostafa A. Salama, Aboul Ella Hassanien, and SE-O. Hanafy, "Genetic algorithm with different feature selection techniques for anomaly detectors generation." IEEE Federated Conference on Computer Science and Information Systems (Fed CSIS), pp. 769-774, 2013.

[8] LijieXu, ಯMapReduce Framework Optimization via Performance Modelingರ, Proc. of Parallel and Distributed Processing Symposium Workshops & PhD Forum (IPDPSW), 2012.

[9] Jungkyu Han, Masakuni Ishii and Hiroyuki Makino. “A Hadoop Performance Model for Multi-Rack Clusters”, the 5th International Conference on Computer Science and Information Technology (CSIT), 2013.

[10] P. Kumar and A. Verma, “Independent task scheduling in cloud computing by improved genetic algorithm”, International Jour-nal of Advanced Research in Computer Science and Software Engineering ,vol.2,no.5,pp.111ರ114,2012.