International Journal on Future Revolution in Computer Science & Communication Engineering
Volume: 3 Issue: 11
ISSN: 2454-4248
568 – 572

Sine Cosine Based Algorithm for Data Clustering

Manju Bala
I. P. College for Women, Delhi University, Delhi, India
manjugpm@gmail.com

Abstract- K-Means clustering algorithm is simple and prevalent, but it has a basic problem to stuck at local optima which relies on randomly generated centroid positions. Optimization algorithms are outstanding for their capacity to lead iterative computation in looking for global optima. Clustering analysis, in today's world, is an important tool and seeking to recognize homogeneous groups of objects on the basis of values of attributes in many fields and applications. In this paper we have proposed a Sine Cosine based algorithm for data clustering (SCBAFDC). The proposed algorithm is tested on five benchmark datasets and compared with other five clustering algorithms. The results show that the proposed algorithm is giving competitive results as compared to the other algorithms in terms of quality of clustering.

*****

I. Introduction

Clustering, which is a popular analysis technique in data mining, pattern recognition and machine learning used in many applications. Based on the values of different attributes, it is used to identify the consistent groups of the same[1–3]. Clustering can be of two types- Hierarchal clustering and Partitioning clustering. Hierarchal clustering mainly functions either agglomerative or divisive. Division means splitting big clusters into smaller ones while agglomeration is joining smaller clusters into closest cluster.

In partition based clustering, compute the optimized value of objective function and update the Centre of cluster which is known as centroid.The basic goal of clustering is to find the maximum similarity within the same cluster and maximum diversity between different clusters.

K-means is the most popular center based, easy and fast clustering algorithm[9].Nevertheless, K-means algorithm majorly depends on the initial statesand always stucks to the local optima. To remove the local optimaproblem, many nature inspired algorithms are come into existences which are basically based on hierarchical clustering, partition-based clustering and density-based clusteringsuch as: statistics [10], graph theory [11],artificial neural networks [13–16], metamorphicalgorithms [17, 18], swarm intelligence algorithms [19–24] and so on.

In this work, Sine Cosine based algorithm is used for data clustering(SCBAFDC)on benchmark problems. (12 test database)We compared the performance of the SCBAFDC on clusteringwith the results of the other clustering algorithmson the same data sets that are given in [21]. We have examine the SCBAFDC and other similar clustering algorithms and also compared the performance of SCBAFDC algorithm with the large set of classification problems that are given in [21].

The paper is classified as theclustering problem is discussed in Section 2, implementation of the SCBAFDCpresented in Section 3, and later experiments and resultspresented and explained in Section 4.We conclude the paper in Section 5 by reviewing the observations and opinion about the future works.

II. The Cluster Problem Analysis

Now days, clustering of data arethe one of the most essential and famous data analysis methods, and specifies to the process of combining data items into clusters.For clustering, data within the same cluster must have higher similarities while data in different clusters have higher dissimilarities.[3,16].

Basically, for calculating the similarities between the data items, Euclidean distance calculation is applied. Specifically, the problem statement for the distance calculation is as follows: given N objects, select each item to one of K clusters and minimize the sum of squared Euclidean distances between each item and the center of the cluster that infers to every allocated item:

$$F(P, Q) = \sum_{i=1}^{N} x \sum_{j=1}^{K} W_{ij} \|(P_i - Q_j)\|$$

Where ||Pi -Qj|| is the Euclidean distance between a data itemPi and the cluster center Qj. N and K are the number of data items and the number of clusters, respectively. Wijis the associated weight of data itemPi with cluster j, which will be either 1 or 0 (if item i is assigned to cluster j; wij is 1, otherwise 0). Fuzzy clustering allows wij to take values in the interval (0, 1).

The most popular type of clustering algorithm is canonical clustering algorithms which is classified as hierarchal and partition algorithms[16,25,26]. Among the canonical clustering algorithms, K-means is the most popular algorithm due to its simplicity and efficiency [26,16]. K-means has two fundamental drawbacks: first, it needs the number of clusters (number of clusters must be given) before initialization. Secondly, it also depends on initial centroid position and may converge at local optima.For overcoming the problems related to K-means, many metaheuristic nature inspired algorithms are came into existence during last decades. Few of these

 ISSN: 2454-4248
568 – 572

metaheuristic nature inspired algorithms are as:genetic algorithms [17,12,13,16], ant colony optimization [18,17,19], particle swarm optimization algorithm [1,14,24], artificial bee colony [11], gravitational search algorithm [18,22], binary search optimization algorithm [20], firefly algorithm [19], and big bang–big crunch algorithm [17].

Clustering algorithms are now used in many fields such as image processing [7,17,15], document clustering [6,8,22], geophysics [29,11], prediction approaches[7,11], marketing and customer satisfaction[21], agriculture analysis[18], security and cyber-crime detection [13], medicine [20,23], anomaly detection [14,20] and biology [19,22].

III. Sine Cosine Algorithm

When all is said in done, all populace based advancement issues begin with irregular arrangements. At that point we over and again call the target work and enhance the goal work. Since all the population based advancement procedures search for ideal arrangements however there is no certification for discovering it in single run. Be that as it may, with adequate number of streamlining emphases and irregular arrangements, the likelihood of discovering it increments.

Sine Cosine calculation depends on the wonder of sine and cosine bend. It chips away at union of sine and cosine bend on its plane. All the nature inspired metaheuristic algorithm deals with two stages: exploitation stage and exploration stage. In the abuse stage, an advancement calculation combines the irregular arrangement with the high rate of haphazardness so that locate the promising area of the hunt space. In the misuse stage, in any case, changes in the arbitrary arrangements are less.

All things considered, the pseudo code of the SCA calculation is displayed beneath:

Initialize a set of search agents (solutions)(X)
Do
Evaluate each of the search agents by the objective function
Update the best solution obtained so far ($P=X^*$)
Update r1, r2, r3, and r4
Update the position of search agents using Eq.
While(t< maximum number of iterations)
Return the best solution obtained so far as the global optimum

The execution of the SCBAFDC calculation is contrasted and other prominent nature reused calculations, for example, Genetic Algorithm (GA), Differential Evolution (DE), Particle Swarm Optimization (PSO) on obliged and unconstrained issues [28–30]. To survey the execution of the SCBAFDC calculation we have connected it to take care of the grouping issue. As indicated by [20] when bunch size is equivalent to three, the issue progresses toward becoming NP-hard.

When we applying sine cosine calculation for information grouping, it gives 1-dimensional cluster for applicant arrangement. Each applicant arrangement is considered ask introductory cluster focuses and the individual unit in the exhibit as the group focuses measurement.

IV. Experimental Result

In this work, 12 order issues from the UCI database [33] which is a notable database storehouse are utilized to assess the execution of the SCBAFDC. The informational indexes and their elements: the number of designs, the number of inputs and the number of classes are displayed in Table 1. These 13 benchmark issues are picked precisely the same as in [26], to make a dependable examination.

From the database, the initial 75% of information is utilized as a part of preparing procedure as a training set, and the staying 25% of information is utilized as a part of testing procedure as a test set. Although, a few informational collections (glass, thyroid, and wine) classes are given in successive rundown, they are rearranged to speak to each class both in training and in testing as in [26].

4.1. Test problems

The issues believed in this work can be depicted quickly as takes after. Balance informational collection was created to demonstrate mental results come about. Every case is named having the adjust scale tip to one side, tip to one side, or be adjusted. The informational collection incorporates 4 inputs, 3 classes and there are 625 illustrations which is part into 469 for training and 156 for testing.

Cancer and Cancer-Int dataset depend on the "bosom malignancy Wisconsin - Diagnostic" and "bosom tumor Wisconsin - Original" informational collections, individually. They are determination of bosom tumor, with 2 yields (group a tumor as either kind or threatening). The previous one contains 569 examples, 30 inputs and the last one contains 699 examples, 9 inputs.

Dermatology dataset contains one of the greatest number of classes; 6 of which are psoriasis, seboreic dermatitis, lichen planus, pityriasis rosea, interminable dermatitis, and pityriasis rubra pilaris. There are 366 specimens, including 34 inputs. The diabetes informational index, a two class issue which is the determination of diabetes (regardless of whether an individual is diabetes positive or not), has 768 examples. We utilized the initial 576 examples as preparing set and the staying 192 as test set. There are 8 contributions for each example. For the issue of Escherichia coli, the first informational index has 336 cases framed of eight classes, yet

 ISSN: 2454-4248
568 – 572

three classes are spoken to with just 2, 2, 5 cases. In this manner, these 9 illustrations are precluded and 327 of aggregate, initial 245 of them in preparing and the rest of the 82 cases in testing, are utilized. The informational index contains 327 cases with 7 data sources and 5 classes.

Glass dataset is the another greatest number of classes (6 classes) in the issues that we handle. It is utilized to group glass sorts as buoy prepared building windows, non-skim handled building windows, vehicle windows, holders, flatware, or head lights. Nine sources of info depend on 9 synthetic estimations with one of 6 sorts of glass which are persistent with 70, 76, 17, 13, 9, and 29 cases of each class, separately. Add up to 214 examples are part with 161 for training and 53 for testing.

Heart dataset that is an analysis of coronary illness chooses to regardless of whether no less than one of four noteworthy vessels is diminished in breadth by over half or not. It contains 76 traits for each example, 35 of which are utilized as information qualities. The information depends on Cleveland Heart information from the archive with 303 examples

Horse dataset is utilized to foresee the destiny of a steed with a colic what's more, to characterize whether the steed will survive, will pass on, or will be euthanized. The informational index is made in view of Horse Colic information with 364 examples, each of which has 58 contributions from 27 traits and 3 yields.

Iris dataset incorporates 150 objects of blossoms from the Iris species: Setosa, Versicolor, and Virginica. Each of 50 protests in each of three classes have 4 factors; sepal length, sepal width, petal length, furthermore, petal width. Thyroid is the conclusion of thyroid whether it is hyper or hypo function. 5 sources of info are utilized to order 3 classes of thyroid capacity as being over function, typical capacity, or under function. The informational index depends on new-thyroid information and contains 215 examples.

Wine dataset which was acquired from a concoction examination of wines were gotten from three distinct cultivators. In this way, the information examination decides the three sorts of wines. There are 178 examples of wine tests with 13 inputs.

Table 1: Properties of the problem

	Data	Train	Test	Input	Class
Balance	625	469	156	4	3
Cancer	569	427	142	30	2
Cancer-Int	699	524	175	9	2
Credit	690	518	172	51	2
Dermatology	366	274	92	34	6
Diabetes	768	576	192	8	2
E coli	327	245	82	7	5
Glass	214	161	53	9	6
Heart	303	227	76	35	2
Horse	364	273	91	58	3
Iris	150	112	38	4	3
Thyroid	215	162	53	5	3
Wine	178	133	45	13	3

4.2. Algorithms and settings

The Particle Swarm Optimization calculation is a population based furthermore, swarm insight based transformative calculation for critical thinking. In the PSO calculation which recreates the social conduct of a run of flying creatures traveling to assets, the particles iteratively assess the wellness of the competitor arrangements and recollect the area which is the best. The parameters of PSO calculation are (as in [26]): n = 50, Tmax = 1000, vmax = 0.05, vmin = −0.05, c1 = 2.0, c2 = 2.0, wmax = 0.9, wmin = 0.4. With a specific end goal to make a reasonable examination, the estimations of state size and most extreme cycle number of the SCBAFDC calculation are picked same as or not as much as the estimations of swarm size and greatest emphasis number utilized as a part of PSO case, individually. For example, we chose the state estimate 20, greatest minimum cycle number (MCN) 1000, and farthest point esteem 1000. Along these lines, add up to assessment # of calculation is 20,000 where it is 50,000 for PSO calculation. We watched that in all keeps running of the calculations the outcomes don't contrast much, so that the tests are cut after 5 keeps running since they have similar outcomes.

4.3. Results and discussion

For every issue, we report the Classification Error Percentage (CEP) which is the rate of inaccurately arranged examples of the test informational collections. We arranged each example by doling out it to the class whose inside is nearest, utilizing the Euclidean separations, to the focus of the clusters. This appointed yield (class) is contrasted and the coveted yield and on the off chance that they are not precisely the same, the example is isolated as inaccurately grouped. It is computed for all test information and the aggregate inaccurately grouped example number is percentages to the extent of test informational index, which is given by-

CEP = 100 × #of misclassified illustrations

As depicted over, the information is given in two pieces: the trainingset (the initial 75%) and the test set (the last 25%). The aftereffects of the calculations SCBAFDC and PSO for the issues are given in where arrangement blunder rates (CEP qualities) are displayed. SCBAFDC calculation outflanks PSO calculation in 12 issues, though PSO calculation's outcome is superior to that of SCBAFDC calculation just for one issue (the glass issue) as far as characterization mistake. Besides, the normal clustering mistake rates for all issues are 14.13% for SCBAFDC and 15.99% for PSO.

 ISSN: 2454-4248
568 – 572

Table 2: Simulation results for clustering algorithms.

Dataset	Criteria	K-means	GA	ACO	PSO	GSA
Iris	Best	97.333	113.98	97.10	96.894	96.698
	Average	106.050	125.19	97.17	97.232	96.723
	Worst	120.450	139.77	97.80	97.897	96.764
	Std	14.631	14.563	0.367	0.347	0.0123
	NFE	120	38128	10998	4953	4628
Wine	Best	16555.68	16530.53	16530.53	16345.96	16315.35
	Average	18061.00	16530.53	16530.53	16417.47	16376.61
	Worst	18563.12	16530.53	16530.53	16562.31	16425.58
	Std	793.21	0	0	85.49	31.34
	NFE	390	33551	15473	16532	15300
Glass	Best	215.74	278.37	269.72	270.57	220.78
	Average	235.50	282.32	273.46	275.71	225.70
	Worst	235.38	286.77	280.08	283.52	229.45
	Std	12.47	4.138	3.584	4.55	3.4008
	NFE	630	199892	196581	198765	171910
CMC	Best	5842.20	5705.63	5701.92	5700.98	5698.15
	Average	5893.60	5756.59	5819.13	5820.96	5699.84
	Worst	5934.43	5812.64	5912.43	5923.24	5702.09
	Std	47.16	50.369	45.634	46.95	1.724
	NFE	270	29483	20436	21456	11796
Cancer	Best	2999.19	2999.32	2970.49	2973.50	2967.96
	Average	3251.21	3249.46	3046.06	3050.04	2973.58
	Worst	3521.59	3427.43	3242.01	3318.88	2990.83
	Std	251.14	229.734	90.500	110.80	8.1731
	NFE	180	20221	15983	16290	8262

V. Conclusion

In this work, Sine cosine based clustering algorithm, which is another, basic and simple advancement procedure, is utilized as a part of clustering of the benchmark classification problems for classification function. Clustering is a vital order strategy that accumulates information into classes (or clusters) with the end goal that the information in each cluster shares a high level of closeness while being extremely unique from information of different clusters. The execution of the SCBAFDC algorithm is compared with PSO and nine other techniques which are generally used by the researchers. The results show that SCBAFDC algorithm can successfully be applied to clustering for the purpose of classification. There are a few issues staying as the degrees for future reviews, for example, utilizing distinctive algorithms in clustering and comparing the results of SCBAFDC algorithm and the results of those algorithms.

References

[1] A. Jain, R. Dubes, Algorithms for Clustering Data, Prentice-Hall, EnglewoodCliffs, NJ, 1998.

[2] M. Sarkar, B. Yegnanafayana, D. Khemani, A clustering algorithm using an evolutionary programming-based approach, Pattern Recogn. Lett. 18 (1997) 975–986.

[3] J. Han, M. Kamber, Data Mining: Concepts and Techniques, Academic Press, 2001.

[4] H. Frigui, R. Krishnapuram, A robust competitive clustering algorithm with applications in computer vision, IEEE Trans. Pattern Anal. Mach. Intell. 21 (May (5)) (1999) 450–465.

[5] Y. Leung, J. Zhang, Z. Xu, Clustering by scale-space filtering, IEEE Trans. Pattern Anal. Mach. Intell. 22 (December (12)) (2000) 1396–1410.

[6] A.K. Jain, M.N. Murty, P.J. Flynn, Data clustering: a review,ACMComput. Surveys 31 (3) (1999) 264–323.

[7] L. Rokach, O. Maimon, Clustering methods, in: O. Maimon, L. Rokach (Eds.), Data Mining and Knowledge Discovery Handbook, Springer, New York, 2005, pp. 321–352.

[8] B. Mirkin, Mathematical Classification and Clustering, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1996.

[9] J. MacQueen, Some methods for classification and analysis of multivariateobservations, pp. 281–297, in: Proc. 5th Berkeley Symp. Math. Stat. Probability,1967.

[10] E.W. Forgy, Cluster analysis of multivariate data: efficiency versus interpretabilityof classification, Biometrics 21 (3) (1965) 768–769.

[11] C.T. Zahn, Graph-theoretical methods for detecting and describing gestalt clusters,IEEE Trans. Comput. C-20 (January (1)) (1971) 68–86.

[12] T. Mitchell, Machine Learning, McGraw-Hill, New York, 1997.

[13] J. Mao, A.K. Jain, Artificial neural networks for feature extraction and multivariatedata projection, IEEE Trans. Neural Netw. 6 (March (2)) (1995)296–317.

[14] S.H. Liao, C.H. Wen, Artificial neural networks classification and clustering ofmethodologies and applications—literature analysis from 1995 to 2005, ExpertSys. Appl. 32 (2007) 1–11.

[15] N.R. Pal, J.C. Bezdek, E.C.K. Tsao, Generalized clustering networks and Kohonen'sself-organizing scheme, IEEE Trans. Neural Netw. 4 (July (4)) (1993) 549–557.

[16] T. Kohonen, Self-Organizing Maps, vol. 30, Springer–Verlag, Berlin, Germany,1995.

[17] E. Falkenauer, Genetic Algorithms and Grouping Problems, Wiley, Chichester,UK, 1998.

[18] S. Paterlini, T. Minerva, Evolutionary approaches for cluster analysis,in: A. Bonarini, F. Masulli, G. Pasi (Eds.), Soft Computing Applications,Springer–Verlag, Germany, 2003, pp. 167–178.

[19] C.H. Tsang, S. Kwong, Ant colony clustering and feature extraction for anomalyintrusion detection, Stud. Comput. Intell. 34 (2006) 101–123.

[20] R. Younsi, W. Wang, A new artificial immune system algorithm for clustering,in: Z.R. Yang, et al. (Eds.),

 ISSN: 2454-4248
568 – 572

IDEAL 2004, LNCS 3177, Springer, Berlin, 2004, pp.58–64.

[21] I. De Falco, A. Della Cioppa, E. Tarantino, Facing classification problems withParticle Swarm Optimization, Appl. Soft Comput. 7 (3) (2007) 652–658

[22] S. Paterlini, T. Krink, Differential evolution and particle swarm optimisation inpartitional clustering, Comput. Stat. Data Anal. 50 (2006) 1220–1247.

[23] Y. Kao, K. Cheng, An ACO-based clustering algorithm, in: M. Dorigo, et al. (Eds.),ANTS, LNCS 4150, Springer, Berlin, 2006, pp. 340–347.

[24] M. Omran, A. Engelbrecht, A. Salman, Particle swarm optimization methodfor image clustering, Int. J. Pattern Recogn. Artif. Intell. 19 (3) (2005)297–322.

[25] D. Karaboga, An idea based on honey bee swarm for numerical optimization,Technical Report-TR06, Erciyes University, Engineering Faculty, ComputerEngineering Department, 2005.

[26] G. Demiroz, A. Guvenir, Classification by voting feature intervals, pp. 85–92,in: Proceedings of the Seventh European Conference on Machine Learning, 1997.

[27] Y. Marinakis, M. Marinaki, M. Doumpos, N. Matsatsinis, C. Zopounidis, A hybridstochastic genetic—GRASP algorithm for clustering analysis, Oper. Res. Int. J.(ORIJ) 8 (1) (2008) 33–46.

[28] B. Basturk, D. Karaboga, An Artificial Bee Colony (ABC) algorithm for numericfunction optimization, in: IEEE Swarm Intelligence Symposium, Indiana, USA,2006.

[29] D. Karaboga, B. Basturk, Artificial Bee Colony (ABC) optimization algorithm forsolving constrained optimization problems, LNCS: Advances in Soft Computing:Foundations of Fuzzy Logic and Soft Computing, vol. 4529, Springer–Verlag,2007, pp. 789–798.

[30] D. Karaboga, B. Basturk, A powerful and efficient algorithm for numerical function optimization: Artificial Bee Colony (ABC) algorithm, J. Global Optim. 39 (3) (2007) 171–459.