



Categorical Data Clustering Using Harmony Search Algorithm for Healthcare Datasets

Abha Sharma, University Institute of Computing, Chandigarh University, Chandigarh, India


Pushpendra Kumar, Department of Computer Science and Technology, Central University of Jharkhand, India*

 <https://orcid.org/0000-0001-7555-2625>

Kanojia Sindhuben Babulal, Department of Computer Science and Technology, Central University of Jharkhand, India

 <https://orcid.org/0000-0003-0442-8795>

Ahmed J. Obaid, Faculty of Computer Science and Mathematics, University of Kufa, Iraq

 <https://orcid.org/0000-0003-0376-5546>

Harshita Patel, School of Information Technology and Engineering, Vellore Institute of Technology, India

ABSTRACT

Healthcare analytics provide many benefits in healthcare dashboard systems. Healthcare datasets majorly contains categorical attributes. This paper proposed an optimized clustering for healthcare dataset named harmony search based categorical clustering (HSCC). The existing k-modes clustering algorithm is one of the well-known categorical data-clustering algorithm. Since the k-modes algorithm produces local optimal clusters. Generally, researchers use genetic algorithm (GA) based clustering algorithms to converge locally optimal solutions to global optimal solutions. GA has some deficiencies such as premature convergence with low speed. In this paper, harmony search (HS) optimization algorithm used to optimize clustering results. The result shows the proposed HSCC algorithm produced global optimized solution, unbiased and matured results. HSCC produces 98% accuracy for dental and 71% for lung cancer dataset. While GACC produces 95% and 65% accuracy for dental dataset and lung cancer dataset.

KEYWORDS

Categorical data, Clustering, Genetic algorithm, Harmony search, Healthcare dataset, Premature solution

INTRODUCTION

In the current few decennary, the paper-based system has been changed to the electronic system (ES) by various sectors including the healthcare sector as well. The ES system improves productivity and outcomes of the sector where it is applied (Yoo et al., 2012). Healthcare organizations and institutes are collecting electronic health data using online insurance claims, computer-based surveys, Electronic Health Record (EHR) and many other sources. EHR improved the access of patient data, which are

DOI: 10.4018/IJEHMC.309440

*Corresponding Author

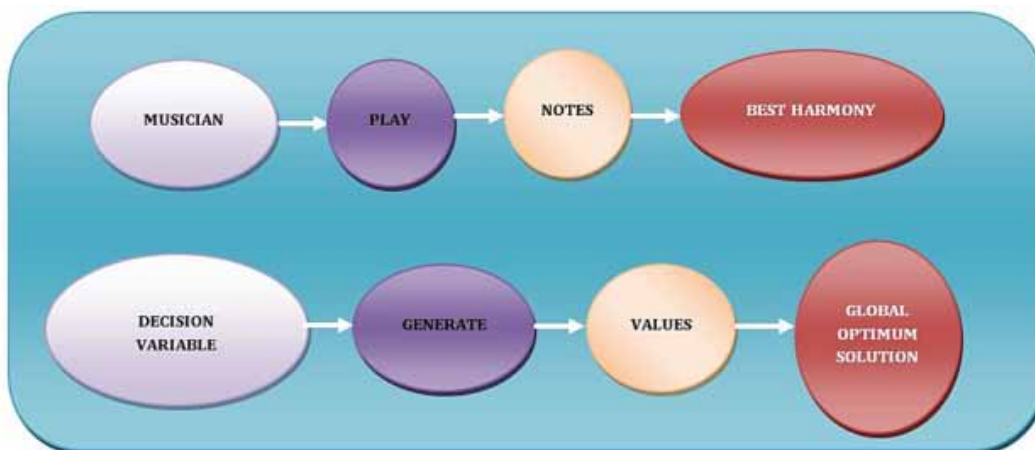
This article published as an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>) which permits unrestricted use, distribution, and production in any medium, provided the author of the original work and original publication source are properly credited.

gathered from hospitals, clinics and other health service providers (Tekieh & Raahemi, 2015). Ledger technology of Blockchain helps the healthcare researcher to facilitate secure movement of patient medical logs, handling the drug supply chain, and accelerating the safe transfer of patient medical logs (Haleem, Javaid, Singh, Suman, & Rab, 2021). Seeing the sensitivity of the medical domain, it is mandatory for computer scientists to find the right insights from the healthcare dataset. Normally healthcare (Kumar Rai, Sharma, Kumar, & Goyal, 2021; Rai & Srivastava, 2014, 2016, 2017) related datasets contain categorical attributes such as gender, age range, symptoms etc. The purpose of this paper is to locate clusters and predict Dental and Lungs Cancer disease. Various analyses are on-going to find the inherent structure or inherent patterns that exist in these datasets such as classification of diseases (P. Kumar & Thakur, 2019, 2021; Harshita Patel, Rajput, Stan, & Miclea, 2022; D. S. Rajput et al., 2021; Reddy et al., 2020), clustering, prediction, logistic regression etc. Cluster analysis (H Patel & Rajput, 2011; D. S. Rajput, 2019) is an important technique to find the inherent structure present in the datasets. Since a large quantity of categorical attributes is present in healthcare or medical datasets and to cluster categorical dataset, among other clustering algorithms k-modes algorithm is straightforward as well as fast. However, when it comes to the optimal solutions there is a need to hybridize these algorithms with optimized clustering algorithms to converge local solutions to global solutions. It means the sensitivity to the initial values is one of the challenges for clustering approaches to generate the sub optimal solutions because those algorithms are similar to hill climbing approaches, as the hill climber shifts in one single path without checking a wider search. Inaccurate choice of the initial cluster seeds usually produces detrimental clustering results. In this paper HS based Categorical data Clustering algorithm (HSCC) is proposed which is hybridization of Harmony Search (HS) (Dubey, Kumar, Kaur, & Dao, 2021) with k-modes to reduce the problem of biased results.

Harmony search (HS) is a population supported optimization algorithm which mimics the musician's behavior (V. Kumar, Chhabra, & Kumar, 2012). It is firstly motivated by the improvisation of Jazz musicians (Moh'd Alia, Al-Betar, Mandava, & Khader, 2011). In HS, decision variables are treated as musicians (Peraza, Valdez, & Castillo, 2015). As a musician plays notes, a decision variable generates values (Shi, Han, & Si, 2013). Through the notes, the musician creates best harmony in the same way decision variables generate values for global optimum solution. The similarity between musicians and HS shown in Figure 1.

Many authors used HS to optimize the following problem such as: gene expression data to improve performance of GACC (Sharma & Thakur, 2017), document clustering (D. Rajput, Reddy,

Figure 1. Harmony search and musicians' behavior



& Shrivastava, 2016; D. S. Rajput, Thakur, & Thakur, 2014), truss structure, time table, single and multi-stage expansion in space shuttle etc. In this research paper, HS is used towards optimizing the clustering results concerning healthcare datasets. However, HS based categorical data clustering is untouched. This paper presents HS based clustering for categorical data and compares its improvement over GACC.

LITERATURE SURVEY

HS algorithms are in a way homogenous to the extemporize process by a trained musician. Numerous researchers have incorporated these techniques to obtain an efficient result in their venture.

(Geem, Kim, & Loganathan, 2001) in 2001 originally proposed HS algorithm and it has been a well intuitive nature-inspired meta-heuristic optimization algorithm and imitating is the extemporization mechanism of music players. It proved victorious in a wide variety of optimization problems.

(V. Kumar et al., 2012) presented revisions in prevailing harmony search, by selecting suitable values of HMCR as well as PAR, then allowing them to change dynamically while improvisation. The impacts of these constant parameters are evaluated.

(Moh'd Alia et al., 2011) explored the search space using HS of the prescribed dataset to locate the optimal cluster centers. Cluster centers evaluated by using c-means, further the finest cluster centers utilized as the initial cluster centers for the c-means algorithms.

(Peraza et al., 2015) presents a paper on Harmony Search (HS) Algorithm and its variants. Later it equated with some existing techniques of optimization like genetic algorithms.

(Shi et al., 2013) has discussed low convergence speed and prematurity of GA. They offered hybridization of HS and GA to form a new algorithm; the proposed algorithm has shown improvement in performance in comparison of GA in respect of result quality along with convergence speed.

(Arunanand, Nazeer, Palakal, & Pradhan, 2014) presented a comparison-based study of some known clustering algorithms. As well as the hybridization of HS was performed accompanied by K-means and Fuzzy c-means for giving superior cluster quality.

(George, Gopakumar, Pradhan, Nazeer, & Palakal, 2015) presented a latest hybrid algorithm employing Harmony Search to calculate the optimal dimension of a SOM grid.

(Malaki & Abolhassani, 2008) presented two clustering algorithms for space shuttle dataset. One algorithm was based on Fuzzy Harmony Search which was suitable for discovering near global regions, yet inferior than fuzzy c-means at fine-tuning across the same amount of time. In second algorithm, Fuzzy c-means was hybridized with FHSClust and found good convergence speed. The FHSClust generate much higher quality clusters with optimized objective functions.

(Al-Betar & Khader, 2012) applied HS algorithm to deal with the university course timetable problem for converging the (near) optimal solution. This paper also developed a improved harmony search algorithm (MHSA), which proposes two modifications of the fundamental HAS, is presented: (i) memory consideration is altered (ii) pitch adjustment operators functionality is enhanced by converting the acceptance rule from 'random walk' to 'first improvement' and 'side walk'. The results of MHSA were better with the basic HSA, but the computational time needed for MHSA was longer enough.

(Cheng, Prayogo, Wu, & Lukito, 2016) developed a Hybrid Harmony Search (HHS) algorithm of truss structure optimization problems. This paper proposes two brand new features supported by Global-best PSO search and neighborhood search which excludes the randomization in original HS.

(Zebalah, Hadjeri, Chatelet, & Massim, 2010) presented a very interesting problem of electrical expansion-planning power design optimization frequently detected in the energy sector. HS is aimed at electing an optimal series-parallel electrical power system design for EPP, that reduces complete contributing cost bounded up with reliability constraints.

(Haleem et al., 2021) reviewed the notable application of blockchain for healthcare. Blockchain can play a pivotal role in handling duplicity in clinical trials for finer healthcare outcomes.

Harmony Search Based Clustering

Many meta-heuristic techniques have been developed to optimize clustering depending on the characteristics of the data for decades. HS is a music inspired optimization mechanism in which the best solution is obtained in terms of objective function similar to the musician establishing the best harmony in terms of aesthetics. HS works in the way musicians improvise the harmony (Nazeer, Sebastian, & Kumar, 2013) to discover the outstanding harmony in regards of notes. HS improvise the harmony from the pool of (Harmony Vector) HV until the best harmony is achieved.

To optimize any problem using HS followings steps have to follow:

Step 1: Initialization of Problem

Various optimization functions have been proposed which calculate the fitness of the HV. In general, fitness functions like squared error function and average distance between data points is used in clustering.

Step 2: Creation of Harmony Memory

The Harmony Memory (HM) is the collection of randomly generated HV where the HV is itself a complete solution of the optimization problem. Many ways have been developed to initialize harmony memory and to improve the convergence rate of the harmony memory. HM is created in a similar fashion as the population is created in the GACC algorithm.

Step 3: Fitness Calculation

Counting on the objective function of the problem the fitness of each HV is calculated and stored. If the fitness of improvised HV is exceeding the worst fit HV in the HM then substitute the existing HV from HM with improvised HV.

Step 4: Improvisation Based on HMCR, PAR and Random Selection

Improvisation of HV is performed using HMCR, PAR along with Random selection such that the optimized value is achieved.

- **HMCR:** HMCR is the rate of selecting one value of improvised HV out of harmony memory, whereas 1-HMCR is the rate of selecting one value out of universe.
- **PAR:** PAR is a constant user assigned value between [0,1]. Decision for adjusting the pitch depends on the Eqn. (1):

$$x_{old} = x_{new} \pm rand() \times b_w \quad (1)$$

where b_w is an arbitrary distance bandwidth, $rand()$ is a random number between 0 and 1.

Step 5: Updating the Harmony Memory

Once the new HV is generated, compare the fitness value of all the HV in the HM. If the new HV finds superior, update the HM by replacing the worst fit HV from improvised HV.

Step 6: Termination Criteria

The whole process will continue until the user defined maximum number of iterations has been achieved.

PROPOSED HSCC METHODOLOGY

This paper developed the optimized clustering algorithm for categorical dataset. Since traditional clustering algorithms such as k-means and k-modes are excellent. However, initial seed selection is difficult.

This paper used the numerous benefits of harmony search. The reason behind this is that the clustering of categorical dataset is not simple, as very few distance measures are present in the literature. Since it works on modes (as central tendency) the results give premature and biased results.

HSCC algorithm is especially for categorical datasets. To the best of our knowledge, clustering of categorical data has not been done with HS. This paper calculated the clustering results of two categorical healthcare datasets; dental dataset and lung cancer dataset. HSCC comprises two stages; the first stage is HS searches for the global optimal centers of the clusters. Whereas in the second phase, these optimized cluster centers are utilized as the initial cluster centers for *k*-modes algorithm (existing categorical data clustering).

This section explains each step of the proposed HSCC algorithm to cluster categorical data.

1. Problem Formulation Using ADDC.

To apply HS on any problem, it should be initially formulated as an objective function that needs to be minimized or maximized. In this proposed approach ADDC (Average Distance of Data points to cluster Centroids) is used as the objective function as shown in Eqn. (2) that has to be minimized (Shi et al., 2013):

$$f(x_i) = \frac{\sum_{j=1}^{n_i} \frac{D(c_i, d_{ij})}{n_i}}{k} \quad (2)$$

subject to constraint:

$$LB_i \leq x_i \leq UB_i$$

LB_i and UB_i are lower and upper bounds of HV. Along with other parameters of HS such as *HMS*, *PAR*, *HMCR*, Number of improvisations.

x_i is one of the solutions in solution space containing n decision variables, where K is the number of clusters, n_i is the number of data objects in cluster i .

D is distance function, d_{ij} is the j^{th} data object of cluster, and c_i is i^{th} cluster.

2. Creation of Harmony Memory.

Harmony memory for categorical data clustering is created using randomly generated solutions. The size of the *HM* is the size of the *HM* (*HMS*) \times number of data objects (m) in the dataset (Nazeer

et al., 2013). It implies all row of harmony memory are a complete clustering solution of the dataset as shown in Eqn (3):

$$\begin{bmatrix} x_1^1 & x_2^1 & \bullet & x_n^1 \\ x_1^2 & x_2^2 & \bullet & x_n^2 \\ \bullet & \bullet & \bullet & \bullet \\ x_1^{HMS} & x_2^{HMS} & \bullet & x_n^{HMS} \end{bmatrix} \quad (3)$$

3. Improvised Harmony.

The new clustering solution, $(x_1', x_2', \dots, x_n')$ is produced using the following rules: memory consideration, pitch adjustment along with random selection. In this step we require a method to generate an updated HV which contains sufficient possible information with a latest clustering answer.

Every decision variable is randomly chosen against HM accompanied by probability HMCR and selected from a set $\{1, 2, 3, \dots, K\}$ with probability $(1 - HMCR)$.

HMCR varies from 0 to 1 shown in Eqn (4), higher the value of *HMCR*, more the number of values will be chosen from the *HM*. *HMCR* decides whether the value should be pitch adjusted or not:

$$x_i = \begin{cases} x_i^i \in (HMCR) \\ x_i^i \in (1 - HMCR) \end{cases} \quad (4)$$

The process of adjusting the pitch is shown in Eqn (4). To store the solutions and to avoid prematurity of the solution this pitch adjustment process is developed. In this paper the PAR is applied to data object x_i which replaces its cluster center with a new cluster based on the following probability distribution (Eqn. 5):

$$p_j = \frac{[c_m d_{\max}(x_i) - d(x_i, z_j)]}{\sum_{i=1}^k c_m d_{\max}(x_i) - d(x_i, z_j)} \quad (5)$$

- $C_m > 1$ is a constant.
- $d(x_i, z_j)$ is the simple matching distance amongst x_i and $d_{\max}(x_i) = \max_{i \leq j \leq k} d(x_i, z_j)$.
- $d(x_i, z_j)$ is 0 if the j^{th} cluster is empty.

4. Updated Harmony Memory

The improvised harmony vector substitutes the worst harmony from the *HM*, provided if its fitness is beyond that of the worst harmony among all (Arunanand et al., 2014).

For example, consider two HV in HM: $HV_1 = [1, 2, 3, 1, 2, 3, 1, 1]$ and $HV_2 = [1, 2, 3, 3, 2, 2, 3, 1]$. Let the fitness values of HV_1 and HV_2 are 10 and 13 respectively. Improvised HV (HV_i) is achieved using HMCR and random selection methods. Further, the process of pitch adjustment takes place after calculating the probability of the data objects of HV_i to be in cluster 1, 2 or 3 using the Equation (3). Therefore, three probabilities are achieved for each.

5. Stopping Criteria.

All the HS based algorithms cease once the maximum number of improvisations is achieved. The stopping criteria varies in proposed HS based categorical data clustering (Al-Betar & Khader, 2012). Pseudo code of proposed Harmony Search algorithm for Categorical data Clustering is shown in Algorithm 1.

DATASET

The experimental results performed on dental and lung cancer datasets because:

1. Dental and lung cancer datasets are categorical datasets and this paper is fully focused on categorical datasets.
2. Dental and lung cancer dataset is healthcare datasets.
3. This paper compared proposed HSCC and existing GACC. In the GACC paper dental and lung cancer datasets were used therefore, we used dental and lung cancer datasets for comparison.

Algorithm 1. Proposed Harmony Search algorithm for Categorical data Clustering (HSCC)

HS Based Categorical data Clustering Algorithm
<p>Input: D: Data set k: Number of cluster centers P: Size of Harmony Memory HMCR: Harmony Memory Considering Rate PAR: Pitch Adjustment Rate MI: Maximum number of iterations</p> <p>Method Steps:</p> <ol style="list-style-type: none"> 1. Generate the harmony memory randomly ($LB_i - UB_i$) 2. While MI 3. for each i [1, N] do 4. If $\text{rand}(0,1) \leq \text{HMCR}$ 5. then 6. begin 7. $x'_i = x_i^j$ 8. if $\text{rand}(0,1) \leq \text{HMCR}$ 9. Improvised HV using $[\text{rand}(1-\text{HMS})]$ 10. if $\text{rand}(0, 1) \leq \text{PAR}$ //Pitch adjustment 11. begin adjusting the pitch using following probability function 12. $p_j = \frac{[c_m d_{\max}(x_i) - d(x_i, z_j)]}{\sum_{i=1}^k c_m d_{\max}(x_i) - d(x_i, z_j)}$ 13. end if 14. else //random selection from $\text{rand}(1 \text{ to } k)$ 15. end if 16. if fitness ($\text{HV}_{\text{new}} > \text{HV}_{\text{worst}}$ from HM) 17. Replace the HV_{worst} with HV_{new}. 18. else 19. Go to step 3 20. End <p>Output: Optimal Modes of the clusters</p>

The explanation of dental and lung cancer is discussed in the following subsection in detail.

Dental Dataset


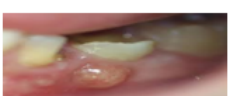
The dataset was collected from Sagar Dental Clinic, Hoshangabad, Madhya Pradesh, India, to predict the dental diseases. In this study, five most common dental diseases: Dental caries, Gingivitis, Periodontitis, Pulpitis abscess or dental abscess and pericoronitis with impaction abscess, considered shown in Fig. 2. Dental data is a purely categorical data set and the patients or data objects recognized as OPD numbers. This dataset contains 24 symptoms and due to the large number of attributes, sample size = 50 is used. The Performa converted into the structured dataset for all the patients and treated as categorical data objects and all the symptoms became the input for both existing GACC and proposed HSCC algorithms.

Lung Cancer Dataset

Identification of various types of lung cancer has been determined from either fluid formation around the lungs or from samples of lung tissue taken for biopsy test or from a sputum sample. This sample is then taken to the lab for a microscopic investigation and the final diagnosis is made based on some characteristics of the cell. Lung cancer is considered of three main categories that are small cell lung cancer, non-small cell lung cancer and nodule lung cancer.

Small cell lung cancer is a tiny cell, which has very less distance between the two sides that means its nucleus, and cell walls are closed to each other. These cells divide and spread over the body very quickly.

Figure 2. Various dental diseases

Dental caries	
Gingivitis	
Pericoronitis with impaction	
Periodontitis	
periapical abscess.	

Non-small cell lung cancer has a huge distance between the nucleus including the outer wall and this macro cell has the function of secreting a mutant that keeps the lungs moist if the cell becomes a cancerous called Adenocarcinoma. There is only a 5-10 percent chance that a lung nodule is cancerous.

Lung cancer dataset is downloaded from UCI repository (Asuncion & Newman, 2007). This dataset contains 56 nominal features decoded data values from 0-3. It has three classes, which describe three varieties of pathological lung cancers (small cell lung cancer, non-small cell lung cancer and nodule lung cancer).

The data taken as the sample size of 32. The description of the dental and lung cancer datasets is given in Table 1. Sample of lung cancer images are provided in Fig.3.

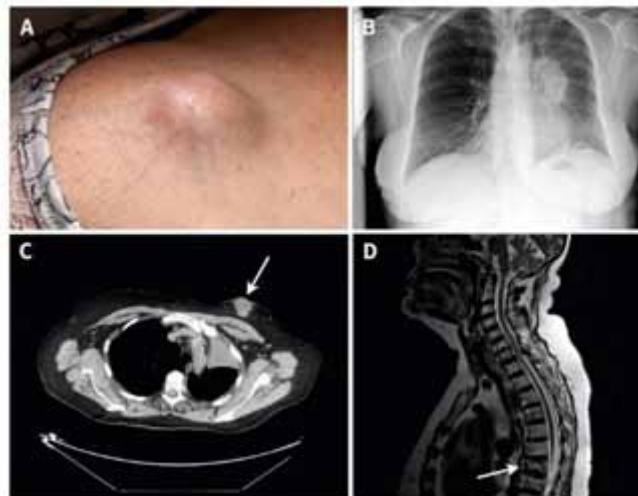
EXPERIMENTAL RESULTS

The proposed HS based clustering algorithm implemented in Java. The experiments were performed on a computing machine with 4 GB RAM on Windows 8 operating system. This algorithm compared with the existing GACC clustering algorithm on dental and lung cancer dataset. Blockchain is used here to secure both lung and dental dataset because if the data is stored physically, data may be corrupted or privacy of the data may be compromised. The decentralized approach and public key

Table 1. Dataset description

Dataset	Features	Instances	Classes
Dental	24	50	5 classes (or 5 Diseases) Class1: dental caries, Class 2: gingivitis, Class 3: pericoronitis with impaction, Class 4: periodontitis, Class 5: periapical abscess.
Lung Cancer	56	32	3 classes (or 3 Diseases of lung) Lung nodules, Non-Small cell lung cancer, Small cell lung cancer.

Figure 3. Lung cancer scan report



used in blockchain prevents data hacking, theft and breaching. The HSCC algorithm executed many times with variations in parameters. The *HMCR* set to 0.80 and *PAR* taken as constant value 0.3 for this experiment is presented in Table 2.

Table 3 along with Table 4 demonstrates the clustering outcomes of dental dataset and lung cancer dataset respectively. For example, in a dental dataset patient having OPD no. 1 has dental disease number 5 i.e., periapical abscess disease. Patient having OPD no. 2 has dental disease gingivitis and so on.

Similarly, Table 4 having OPD no.1 has lung nodules disease in lung cancer dataset. Patients having OPD number 2 have Lung nodules disease and so on. After achieving these following assignments, the accuracy is calculated based on the total number of correctly classified patients shown in Eqn. (6).

Accuracy Calculation

Accuracy is an extrinsic validation measure used to validate the results of any clustering algorithm having approximately equal number of data objects of each class:

$$\text{Accuracy} = \frac{\text{No. of data objects predicted correctly in all the clusters}}{\text{Total number of data objects}} \quad (6)$$

Table 2. Parameters used in the experiment

P. No.	Parameter	Value
1	<i>HMS</i>	100
2	<i>HMCR</i>	0.80
3	<i>PAR</i>	0.30
4	<i>k</i>	2,3,4,5
5	<i>MI</i>	500
6	<i>C_m</i>	3

Table 3. Assignment of classes to each OPD after the implementation of HSCC for dental dataset

OPD No.	Class	OPD No.	Class	OPD No.	Class	OPD No.	Class	OPD No.	Class
1.	5	11.	3	21.	4	31.	2	41.	3
2.	2	12.	2	22.	3	32.	3	42.	3
3.	2	13.	5	23.	2	33.	5	43.	1
4.	2	14.	4	24.	3	34.	3	44.	4
5.	3	15.	5	25.	2	35.	1	45.	5
6.	2	16.	1	26.	4	36.	5	46.	4
7.	4	17.	5	27.	1	37.	1	47.	5
8.	5	18.	3	28.	2	38.	4	48.	2
9.	3	19.	3	29.	3	39.	4	49.	1
10.	2	20.	4	30.	3	40.	2	50.	3

Table 4. Assignment of classes to each OPD after the implementation of HSCC for lung cancer dataset

OPD No.	Class	OPD No.	Class	OPD No.	Class	OPD No.	Class	OPD No.	Class
1.	1	8.	3	15.	1	22.	3	29.	2
2.	3	9.	3	16.	1	23.	2	30.	2
3.	1	10.	2	17.	3	24.	3	31.	3
4.	3	11.	3	18.	3	25.	3	32.	3
5.	3	12.	3	19.	2	26.	2		
6.	1	13.	3	20.	2	37.	3		
7.	2	14.	3	21.	1	28.	3		

ADDC Calculation

It predicts the correct number of clusters based on the minimum value of $f(x)$ because clustering is a minimization problem calculated using Eqn. 2. In GACC the optimum number of clusters present in the dataset has been found using TWCV (Total within cluster variation). Table 5 shows the comparison of GACC and HSCC clustering algorithms w.r.t. accuracy and cost for dental dataset.

The results show that the HS based clustering for categorical data gives better results in terms of accuracy and *ADDC*. HSCC finds better k partitions with HS operators such as *PAR*, *HMCR* and random Selection etc. HS always gives mature solutions and does not require complex calculations as found in literature.

Advantage of Proposed HSCC

1. The HSCC clustering algorithm for categorical data exhibits the improved performance because it carries the simplicity of k -modes and robustness of HS to get the globally optimum partitions of the datasets. HS explored categorical data first time and produced good performance in HSCC.
2. The implementation of the proposed algorithm is very easy because it is truly simple to fine tune HM, PAR, HMCR (only 2-3 parameters).
3. This paper presented results of a proposed algorithm using some validation measures for real life categorical datasets. This paper compared the accuracy and *ADDC* of the partitions with the GACC algorithm. The result shows the HSCC algorithm is better in terms of accuracy over the GACC algorithm.
4. This algorithm is working flawlessly with a higher number of features also such as 24, 56 in dental and lung cancer dataset respectively.

Table 5. Accuracy and cost of datasets

Dental dataset		
	GACC	Proposed HSCC
Accuracy	95%	98%
Optimal number of clusters (Minimum cost value)	5 (calculate based on TWCV)	5 (calculated based on ADDC)
Lung Cancer dataset		
Accuracy	65%	71%
Optimum number of clusters (Minimum cost value)	3 (calculated based on TWCV)	3 (calculated based on HSCC)

Limitations of Proposed HSCC

1. The data objects in the dental dataset (50) and lung dataset (32) is comparatively small, this algorithm can be implemented with larger datasets.
2. This paper can be extended by using more datasets and many other fields.

CONCLUSION

Due to the associated problems with GA such as low convergence speed and prematurity, this paper explored HS because of less effort to fine tune with very few parameters such as PAR, HMCR and HMS. The experimental outcome exhibits that the proposed algorithm and its improvements are preferably better than GA in convergence speed, solution quality and other indicators. The accuracy of HSCC exceeds 98% (for dental dataset) and 71% (for Lung Cancer dataset) which is higher accuracy when compared to the existing models such as GACC. Proposed HSCC can be extended for a larger number of big categorical datasets such as COVID-19 and other healthcare datasets. In future work patients and dentists may participate in the blockchain tools through mobile and web applications and may share their honest feedback. Dentacoins (coins for dental data) could be earned by providing genuine feedback which will support the dentist and researchers community to strengthen the quality of treatment. These coins could be further utilized for dental treatment in future.

CONFLICTS OF INTEREST

The authors declare that there is no conflict of interests regarding the publication of this paper.

FUNDING STATEMENT

This study did not receive any funding in any form.

REFERENCES

- Al-Betar, M. A., & Khader, A. T. (2012). A harmony search algorithm for university course timetabling. *Annals of Operations Research*, 194(1), 3–31. doi:10.1007/s10479-010-0769-z
- Arunanand, T., Nazeer, K. A., Palakal, M. J., & Pradhan, M. (2014). A nature-inspired hybrid fuzzy c-means algorithm for better clustering of biological data sets. *International Conference on Data Science & Engineering (ICDSE)*. doi:10.1109/ICDSE.2014.6974615
- Asuncion, A., & Newman, D. (2007). *UCI machine learning repository*.
- Cheng, M. Y., Prayogo, D., Wu, Y. W., & Lukito, M. M. (2016). A Hybrid Harmony Search algorithm for discrete sizing optimization of truss structure. *Automation in Construction*, 69, 21–33. doi:10.1016/j.autcon.2016.05.023
- Dubey, M., Kumar, V., Kaur, M., & Dao, T.-P. (2021). A systematic review on harmony search algorithm: Theory, literature, and applications. *Mathematical Problems in Engineering*, 2021, 2021. doi:10.1155/2021/5594267
- Geem, Z. W., Kim, J. H., & Loganathan, G. V. (2001). A new heuristic optimization algorithm: harmony search. *Simulation*, 76(2), 60–68.
- George, A. J., Gopakumar, G., Pradhan, M., Nazeer, K. A., & Palakal, M. J. (2015). A self organizing map-harmony search hybrid algorithm for clustering biological data. *IEEE International Conference on Signal Processing, Informatics, Communication and Energy Systems (SPICES)*. doi:10.1109/SPICES.2015.7091532
- Haleem, A., Javaid, M., Singh, R. P., Suman, R., & Rab, S. (2021). Blockchain technology applications in healthcare: An overview. *International Journal of Intelligent Networks*, 2, 130–139. doi:10.1016/j.ijin.2021.09.005
- Kumar, P., & Thakur, R. S. (2019). Diagnosis of Liver Disorder Using Fuzzy Adaptive and Neighbor Weighted K-NN Method for LFT Imbalanced Data. *International Conference on Smart Structures and Systems (ICSSS)*. doi:10.1109/ICSSS.2019.8882861
- Kumar, P., & Thakur, R. S. (2021). An Approach Using Fuzzy Sets and Boosting Techniques to Predict Liver Disease. *CMC-COMPUTERS MATERIALS & CONTINUA*, 68(3), 3513–3529. doi:10.32604/cmc.2021.016957
- Kumar, V., Chhabra, J. K., & Kumar, D. (2012). Effect of harmony search parameters' variation in clustering. *Procedia technology*, 6, 265–274.
- Kumar Rai, B., Sharma, S., Kumar, A., & Goyal, A. (2021). Medical Prescription and Report Analyzer. *Thirteenth International Conference on Contemporary Computing (IC3-2021)*. doi:10.1145/3474124.3474165
- Malaki, M., & Abolhassani, H. (2008). A Combinatory Approach to Fuzzy Clustering with Harmony Search and its Applications to Space Shuttle data. *SCIS & ISIS*.
- Moh'd Alia, O., Al-Betar, M. A., Mandava, R., & Khader, A. T. (2011). Data clustering using harmony search algorithm. *International Conference on Swarm, Evolutionary, and Memetic Computing*. doi:10.1007/978-3-642-27242-4_10
- Nazeer, K. A., Sebastian, M., & Kumar, S. M. (2013). A novel harmony search-K means hybrid algorithm for clustering gene expression data. *Bioinformation*, 9(2), 84–88. doi:10.6026/97320630009084 PMID:23390351
- Patel, H., & Rajput, D. (2011). Data mining applications in present scenario: A review. *International Journal of Soft Computing*, 6(4), 136–142. doi:10.3923/ijscmp.2011.136.142
- Patel, H., Rajput, D. S., Stan, O. P., & Miclea, L. C. (2022). A New Fuzzy Adaptive Algorithm to Classify Imbalanced Data. *CMC-Computers Materials & Continua*, 70(1), 73–89. doi:10.32604/cmc.2022.017114
- Peraza, C., Valdez, F., & Castillo, O. (2015). *A harmony search algorithm comparison with genetic algorithms Fuzzy Logic Augmentation of Nature-Inspired Optimization Metaheuristics*. Springer.
- Rai, B. K., & Srivastava, A. (2014). Security and Privacy issues in healthcare Information System. [IJETTCS]. *International Journal of Emerging Trends & Technology in Computer Science*, 3(6), 248–252.
- Rai, B. K., & Srivastava, A. (2016). Pseudonymization techniques for providing privacy and security in EHR. *International Journal of Emerging Trends & Technology in Computer Science*, 5(4), 34–38.

Rai, B. K., & Srivastava, A. (2017). Patient controlled Pseudonym-based mechanism suitable for privacy and security of Electronic Health Record. *International Journal of Research in Engineering, IT and Social Sciences*, 588(7.2).

Rajput, D., Reddy, P. K., & Shrivastava, D. (2016). Mining Frequent termset for Web Document Data using Genetic Algorithm. *Published in International Journal of Pharmacy and Technology*, 8(2), 4038–4054.

Rajput, D. S. (2019). Review on recent developments in frequent itemset based document clustering, its research trends and applications. *International Journal of Data Analysis Techniques and Strategies*, 11(2), 176–195. doi:10.1504/IJDATS.2019.098818

Rajput, D. S., Basha, S. M., Xin, Q., Gadekallu, T. R., Kaluri, R., Lakshmanna, K., & Maddikunta, P. K. R. (2021). Providing diagnosis on diabetes using cloud computing environment to the people living in rural areas of India. *Journal of Ambient Intelligence and Humanized Computing*, 1–12.

Rajput, D. S., Thakur, R. S., & Thakur, G. S. (2014). An integrated approach and framework for document clustering using graph based association rule mining. *Proceedings of the Second International Conference on Soft Computing for Problem Solving (SocProS 2012)*. doi:10.1007/978-81-322-1602-5_144

Reddy, G. T., Reddy, M. P. K., Lakshmanna, K., Rajput, D. S., Kaluri, R., & Srivastava, G. (2020). Hybrid genetic algorithm and a fuzzy logic classifier for heart disease diagnosis. *Evolutionary Intelligence*, 13(2), 185–196. doi:10.1007/s12065-019-00327-1

Sharma, A., & Thakur, R. S. (2017). GACC: Genetic algorithm-based categorical data clustering for large datasets. *International Journal of Data Mining. Modelling and Management*, 9(4), 275–297.

Shi, W. W., Han, W., & Si, W. C. (2013). *A hybrid genetic algorithm based on harmony search and its improving Informatics and Management Science I*. Springer.

Tekieh, M. H., & Raahemi, B. (2015). Importance of data mining in healthcare: a survey. *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. doi:10.1145/2808797.2809367

Yoo, I., Alafaireet, P., Marinov, M., Pena-Hernandez, K., Gopidi, R., Chang, J. F., & Hua, L. (2012). Data mining in healthcare and biomedicine: A survey of the literature. *Journal of Medical Systems*, 36(4), 2431–2448. doi:10.1007/s10916-011-9710-5 PMID:21537851

Zeblah, A., Hadjeri, S., Chatelet, E., & Massim, Y. (2010). Efficient harmony search algorithm for multi-stages scheduling problem for power systems degradation. *Electrical Engineering*, 92(3), 87–97. doi:10.1007/s00202-010-0165-3

Abha Sharma is an Associate Professor in the University Institute of Computing, Chandigarh University. She did her PhD from NIT Bhopal in 2016 in the field of data mining. She did MCA from RGPV, Bhopal in 2007. Her area of interests are clustering techniques and nature inspired algorithms.

Pushpendra Kumar is an Assistant Professor in the Department of Computer Science & Technology at Central University of Jharkhand, Ranchi, India. He received his Ph.D from National Institute of Technology, Bhopal (MP). He has published research work in various reputed journals. His area of interest includes Data Mining, Machine Learning and Deep Learning.

Kanojia Sindhuben Babulal is an Assistant Professor in the Department of Computer Science & Technology at Central University of Jharkhand, Ranchi, India. She has 10 years of teaching experience. Dr. Sindhu has published research work in various reputed journals. Her area of interest includes Machine Learning, Computer Vision, Energy Efficient Wireless Sensor Networks, MANETS, Cross Layer Designs, 5G Communication.

Ahmed J. Obaid, is a Asst. Professor at the Department of Computer Science, Faculty of Computer Science and Mathematics, University of Kufa, Iraq. Dr. Ahmed holds a Bachelor in Computer Science, degree in – Information Systems from College of Computers, University of Anbar, Iraq (2001-2005), and a Master Degree (M. TECH) of Computer Science Engineering (CSE) from School of Information Technology, Jawaharlal Nehru Technological University, Hyderabad, India (2010-2013), and a Doctor of Philosophy (PhD) in Web Mining from College of Information Technology, University of University of Babylon, Iraq (2013-2017). He is a Certified Web Mining Consultant with over 14 years of experience in working as Faculty Member in University of Kufa, Iraq. He has taught courses in Web Designing, Web Scripting, JavaScript, VB.Net, MATLAB Toolbox's, and other courses on PHP, CMC, and DHTML from more than 10 international organizations and institutes from USA, and India. Dr. Ahmed is a member of Statistical and Information Consultation Center (SICC), University of Kufa, Iraq. His main line of research is Web mining Techniques and Application, Image processing in the Web Platforms, Image processing, Genetic Algorithm, information theory, and Medical Health Applications. Ahmed J. is Associated Editor in Brazilian Journal of Operations & Production Management (BJO&PM) and Editorial Board Members in: International Journal of Advance Study and Research Work (IJASRW), Journal of Research in Engineering and Applied Sciences(JREAS), GRD Journal for Engineering (GRDJE), International Research Journal of Multidisciplinary Science & Technology (IRJMST), The International Journal of Technology Information and Computer (IJTIC), Career Point International Journal Research (CPIJR). Ahmed J. was Editor in Many International Conferences such as: ISCPs_2020, MAICT_2020, IHICPS_2020, IICESAT_2021, IICPS_2020, ICPAS_2021, etc. (Scopus Indexed Conferences). He has edited Some books, such as Advance Material Science and Engineering (ISBN: 9783035736779, Scientific. net publisher), Computational Intelligence Techniques for Combating COVID-19 (ISBN: 978-3-030-68936-0 EAI/ Springer), A Fusion of Artificial Intelligence and Internet of Things for Emerging Cyber Systems (ISBN: 978-3-030-76653-5 Springer). Ahmed J. has supervised several final projects of Bachelor and Master in his main line of work, and authored and co-authored several scientific publications in journals, Books and conferences with more than 65+ Journal Research Articles, 5+ book Chapters, 15+ Conference papers, 10+ Conference proceedings, 8+ Books Editing, 2+ Patent. Ahmed J. also Reviewer in many Scopus, SCI and ESCI Journals e.g., CMC, IETE, IJAACS, IJIPM, IJKBD, IJBSR, IET, IJUFKS and many others. Dr. Ahmed Attend and participate as: Keynote Speakers (40+ Conferences), Webinars (10+), Session Chairs (10+), in many international events in the following countries: India, Turkey, Nepal, Philippines, Vietnam, Thailand, Indonesia and other countries.

Harshita Patel is presently working as an Assistant Professor Senior in the School of Information Technology & Engineering, VIT, Vellore, India. She has received her PhD degree in 2017 from Maulana Azad National Institute of Technology, Bhopal, Madhya Pradesh, India. Additionally, she has qualified national level competitive exams like GATE and UGC-NET. She has more than 13 years of teaching and research experience. She has published a good number of research papers in various conferences and journals of international repute which are indexed in SCOPUS and SCI. She has also delivered expert lectures in reputed workshops. Her areas of research include Data Mining, Machine Learning and Artificial Intelligence.