ISSN 2278-3091
Volume 9, No.4, July – August 2020
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse12942020.pdf
https://doi.org/10.30534/ijatcse/2020/12942020

Novel Multistage approach of Medical Data using Machine Learning with Clustering Technique

[1]K. Sasirekha, [2]Dr. V. Kathiresan
[1]Research Scholar, Department of Computer Science, Dr. SNS Rajalakshmi College of Arts & Science, Coimbatore.
[2]Director, Department of Computer Applications (PG), Dr. SNS Rajalakshmi College of Arts & Science, Coimbatore.

ABSTRACT

Medical knowledge collection from other resources, classification of information in keeping with their belongings, characteristics and arranging the tips the sub team is essential. Due to this reason why, to into effect the data mining and gadget leaning techniques are helpful to get the consumer decisions. Machine learning techniques can also be categorized into two categories, known as supervised learning and unsupervised studying. This analysis work, focus on to investigate clusters of medical information acquired by the use of unsupervised clustering ways and examine the efficiency via classification set of rules at the medical knowledge.

New characteristic variety is accomplished in the class of the dimensionality aid supervised approach referred to as igPCA that attempts to choose a subset of the predictor features in line with the guidelines achieve. After that the easy k-means clustering set of rules randomly chooses K segments as centroids of clusters at the initialization stage. This proposed manner introduce the consensus serve as with the combo of CSPA, HGPA, and MCLA, which might be used to combine unbiased clustering into one final consensus clustering. The result of consensus clustering is used to train the classification set of rules known as SVM. The performance of the classification algorithm on the ultimate degree is crucial for figuring out the effectiveness of the classification set of rules. Performance analysis may also be thought to be a sign of the reliability, quality and balance of the combined consensus clustering.

Key words: - Medical data, Clustering, Consensus function, Classification, machine learning.

1. INTRODUCTION

The research of scientific knowledge is these days a key matter in bio statistics and system studying programs to medical research. The effective knowledge acquisition is matter to many stumbling blocks. In hospital therapy a number of restrictions rise up from moral and experimental problems. The high prices associated with scientific analysis lead to less widespread tests, which in turn results in longitudinal knowledge this is most often sparse and incomplete, with unequal and long sampling classes, which would possibly critically hamper the analysis and the proper identification of significant covariance. In reality, maximum of this information translates into time sequence, because the corresponding sufferers are medically adopted for a time frame, which results in the relevancy of longitudinal information research strategies so as to higher perceive and interpret medical information [8].

Medical area has huge quantity of information that require processing and analysis in an effort to extract useful data that now and again would possibly save a human existence. Medical information come with patient data, test effects, or some type of pictures similar to X-rays, MRI, ECG, EEG and CT scans. In order to analyze this information, supervised and unsupervised learning tactics are vital to facilitate knowledge handling and determination making. This analysis concerns best about supervised and unsupervised finding out algorithms. These techniques have been closely hired in scientific box. In the previous work, fuzzy clustering is employed to generate the weights of each example within the dataset to which class it belongs to introduce additional vital options added to the data [9]. The information is then fed to change SVM for classification procedure. The effects aren't in the expected level of the researcher and additional it may be advanced using the opposite means of the proposed one. It overcomes the other works offered in the literature.

This analysis displays that the proposed system can be employed as an impressive tool to facilitate final choice of medical diagnosis and will also be successfully applied for more than a few medical data classification. Although the result of this research is promising, selection of basic directions remains open to increase this work. This research can be prolonged to research other real-world problems of various domain names. Also, checking out the scalability of the improved multistage machine is a fascinating matter. Furthermore, the dataset used on this research is a well-known two-class problem; and in our work will also be evaluating the efficiency of the proposed system on other multi class issues [3]

In our proposed multistage approach together with more than a few procedures for dimensionality reduction, consensus clustering of randomized samples, followed by the use of a quick supervised classification set of rules. The performance of the classification algorithm at the final level is an important for the effectiveness of this technique. It will also be thought to be an indication of the reliability, high quality and stability of the blended consensus clustering.

This paper is organized as follows. In section 2, a short-lived overview of the similar work presented. Section 3 describes concerning the methodology of proposed research work.

Section 4 incorporates experimental results comparing with the prevailing one. Section 5 is dedicated to the conclusion and future route.

2. RELATED WORK

There are several approaches are recognized in this literature evaluate for solving this research problem. But every way has its own features and shortcomings additionally. This phase summarizes the literature evaluation of the problem.

The packages of information mining is disease prognosis for this purpose one wishes scientific dataset to spot hidden patterns and finally extracts useful wisdom from scientific database. Recently, researchers have used different classification and clustering algorithms for diagnosing diseases. This paper provides survey on two other complex diseases which contains the guts disease and Cancer illness, paper severely noticed the existing literature paintings to determine significant wisdom in this area and summarize other approaches utilized in disease diagnosing, additional discussed the equipment to be had for processing and classification of data [1].

Investigates the prevailing practices and potentialities of scientific information classification in response to data mining tactics. It highlights major advanced classification approaches used to beef up classification accuracy. Past analysis has supplied literature on medical knowledge classification using knowledge mining tactics. From in depth literature research, it is discovered that information mining techniques are very efficient for the task of classification. This paper analyzed relatively the present advancement in the classification of medical knowledge. The findings of the find out about showed that the existing classification of medical information can also be improved. , there will have to be more analysis to ascertain and reduce the ambiguities for classification to realize precision [2].

Due to the rapid building in era these days, massive amount of knowledge are available. In medicine, choice making is completely in line with the hidden knowledge in that massive information. For that explanation why, data mining and machine learning technologies provide powerful tools for knowledge discovery inside of data. Two main techniques are used interchangeably: clustering and classification. In system studying, clustering is an unsupervised studying technique while classification is a supervised finding out approach. These techniques are capable of extracting helpful patterns and data which assist the process of data analysis and medical decisions. This analysis gifts a recent study of those ways in the scientific box all over the previous five years. , this paper proposes a hybrid multistage fuzzy clustering system implemented to medical knowledge classification. In the proposed machine, two fuzzy clustering algorithms particularly FCM and GK have been to begin with employed to procure the club values. These weights are then utilized in the second level of the system as informative options to reinforce the classification procedure completed by means of SVM algorithm [3].

A unique system finding out means applying consensus clustering together with classification for the knowledge mining of very large and extremely dimensional ECG information units. To obtain robust and stable clustering, consensus functions may also be carried out for clustering ensembles combining a mess of unbiased preliminary clustering. Direct programs of consensus purposes to highly dimensional ECG information sets stay computationally expensive and impracticable. We introduce a multistage approach including various procedures for dimensionality reduction, consensus clustering of randomized samples, adopted by way of a fast supervised classification algorithm. Applying the Hybrid Bipartite Graph Formulation combined with rank ordering and SMO [4].

Dimensionality Reduction (DR) permits the development of a decrease dimensional space (embedding) from a high dimensional function area while keeping object-class discriminability. Several popular DR approaches suffer from sensitivity to selection of parameters and/or presence of noise within the data. In this paper, we present a novel DR technique known as consensus embedding that goal to conquer those problems through generating and combining more than one low-dimension a embeddings, hence exploiting the variance amongst them in a fashion very similar to ensemble classifier approaches comparable to Bagging. We display theoretical homes of consensus embedding which display that it'll lead to a single stable embedding solution that preserves knowledge more than as in comparison to any individual embedding and code parallelization are utilized to supply for an efficient implementation of the way [5].

Therefore, the proposed multistage approach combines quite a lot of procedures for dimensionality reduction, consensus clustering of randomized samples and fast supervised classification algorithms for processing of the highly dimensional medical datasets

.

3. PROPOSED METHODOLOGY

In this segment, a novel multistage manner is describes using the high-dimensional medical dataset. The proposed multistage manner consists of the three stages,

- **Phase 1: - Dimensionality Reduction (DR).**
- **Phase 2: - Clustering with Consensus Function.**
- **Phase 3: - Classification the use of SVM.**

The detailed descriptions of the above stages are given in the following subsections and the proposed framework (figure 1) presentations the running concept of the multistage way.

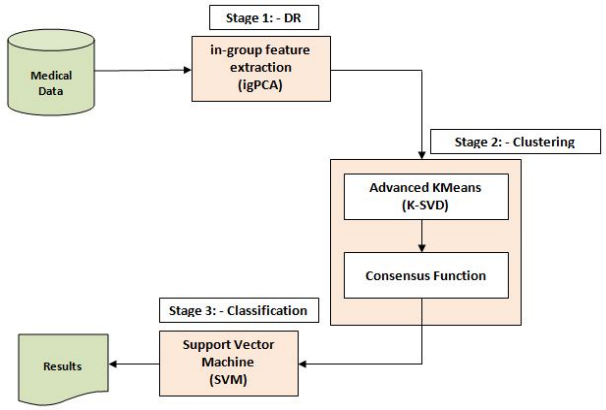

Figure 1: Proposed Framework

3.1. Dimensionality Reduction using PCA

Reduction of large datasets will also be carried out through decreasing the number of analyzed parameters (dimensions) or by way of reducing the number of analyzed instances. The dimensionality reduction can be carried out via statistical strategies, basically Principal Component Analysis (PCA) [11] or via using characteristic variety techniques [14, 15]. Dataset cardinality aid can also be accomplished through sampling, grouping or example selection methods [16].

In this analysis, we suggest a modification to the applying of PCA way called igPCA (in-group Principal Component Analysis). It introduces the pre-processing phase that arranges the comparable features into teams of identical distribution. Further, applied our approach to scale back knowledge derived from ECG alerts to strengthen storage and inference process in fixing arrhythmia classification problem. The proposed in-group feature extraction means (igPCA) is in line with main part research incorporating variety in distribution of quite a lot of parameters. The steps can be introduced as follows:

The process starts with data preparation, which aims at adjusting original datasets to analysis needs.

1. Feature grouping based on statistical analysis of distribution is carried out and groups of similar characteristics are distinguished:
   - The first group consists of binary features (B).
   - The rest of the features are divided into five groups based on skewness of the frequency distributions and ranging:
     a. a highly positive distribution,
     b. a moderately positive distribution,
     c. a symmetric distribution,
     d. a moderately negative distribution,
     e. a highly negative distribution.
2. The principal component analysis is performed for each of six separate groups of features and results in six sets of new features:
   - F_{SB} - for binary features,
3. The final result set of new features is a sum denoted as (3):

$$F_{igPCA} = F_{SB} \cup F_{S1} \cup F_{S2} \cup F_{S3} \cup F_{S4} \cup F_{S5} \quad \text{.... equ. (1)}$$

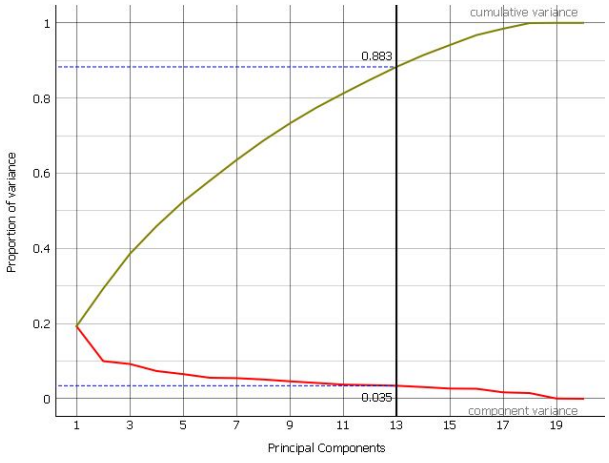

Figure 2: Principal Component Analysis

The original heart disease scientific dataset contained parameters of uniform values (normally zeros) or attributes where the number of rows with non-zero values was once below 10 (< 2% of all recordings) (figure 2).

Data instances: 303
Features: 13
Meta attributes: 14
Target: Class 'diameter narrowing'

	diameter narrowing	Selected	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12
1	0	No	1.37672	0.819689	5.32805	0.164384	-2.89838	3.78746	0.864347	0.981341	-0.451652	0.701081	-0.466098	0.680617
2	1	No	3.32402	1.61588	-0.51415	-1.11795	0.598914	-0.229254	-0.288242	0.83722	0.887038	1.55577	0.0876092	-1.98731
3	1	No	4.04827	-0.117373	-0.66669	-0.97122	0.143153	-0.783403	-0.262062	-0.0484561	0.363656	-0.000557622	-0.407619	-0.862771
4	0	No	-1.51479	-1.17509	0.428781	0.997866	-2.43361	2.3778	1.33497	-2.87064	0.281303	-1.5375	-0.743697	-1.11276
5	0	No	-2.97117	1.0307	-0.487136	-1.62776	0.285397	1.15261	0.166793	0.507577	-0.895723	-1.52696	-0.0669712	0.224433
6	0	No	-2.81747	-1.71279	-0.206248	-0.236512	0.136902	0.460074	0.582288	1.37711	-0.173729	-0.413103	0.481053	-1.2647
7	1	No	0.601348	3.05371	-0.0869615	-0.580601	-0.971629	1.94905	2.71231	-0.7041	0.899374	-1.3559	-2.56796	-0.8832
8	0	No	-0.772758	0.964735	-2.22154	0.684444	2.97342	0.868437	0.631274	-0.411942	-0.294933	0.909718	0.152798	0.0177062
9	1	No	1.9953	-0.0481745	-0.0217388	-1.34756	-1.01207	-1.6877	0.205517	0.455639	0.152597	-0.717509	-0.638658	0.538983
10	1	No	3.33991	-0.283667	3.28151	0.310101	0.926147	2.917	0.847958	-2.06768	-0.695387	-1.98481	-1.04807	-0.0903842
11	0	No	0.810056	-1.33557	-0.378487	0.949457	-1.91112	0.965569	-1.13926	2.51053	1.1761	0.552065	0.16773	1.85763
12	0	No	-1.26935	2.64212	-0.929515	-1.02875	-0.758119	0.0165168	-0.22254	1.28329	-1.65285	-1.52834	0.647555	-0.810614
13	1	No	2.26719	0.691746	3.29345	0.858946	-0.0331027	1.04565	-3.50104	0.422331	0.993377	0.132118	2.04496	0.227622

Table 1: Principle Components

The key element of the primary element analysis is resolution of the collection of primary elements to be stored for further research.

- 13 essential elements after PCA carried out for the whole dataset (table 1).
- No main element for binary options.
- 3 important components for every other workforce.

3.2. Clustering with Consensus Function

Clustering is a technique of dividing a dataset D into a suite of clusters such that the objects in the similar cluster are very similar to every rather than the items within the other clusters [12]. The proposed multistage means the usage of advanced k-means algorithm. The fundamental objective of this algorithm is to decide the optimum and minimum number of clusters in combination with K-SVD as a dictionary finding out approach and CS principle as a random sampling means [13]. The K-SVD improves the K-Means clustering process for adapting dictionaries in an effort to reach sparse medical information sign representation.

Therefore, the principle purpose of KSVD is to coach the proper dictionary to generate compressed medical indicators. The K-Means algorithm in response to K-SVD approach states that the sparse pattern of a dataset can be recovered from a suite of low-dimensional random linear measurements. In the K-SVD learning approach together with CS idea, slightly than measuring each pattern after which computing a compressed representation, we will measure and gather a compressed representation of the dataset without delay.

$$c = \Sigma_i \Sigma_j \|X_j - v_j(Xi)\|^2 \quad \text{.... equ.(2)}$$

The steps for proposed advanced k-means algorithm is given below,

Input: - $X = \{X_i : X_i \varepsilon R^N\}$

(K : the number of Clusters, M = the number of random linear measurements)

Output: - Optimal partitioning based on minimum number of Clusters.

Step 1: - For a given price of k, the enter dataset partitions into k clusters;
Step 2: - Iterate over all knowledge vectors to determine clusters in response to nearest subspace in line with (2);
Step 3: - Compute the information vector's contribution to the full residual by re-assignment of the input knowledge vectors to their found clusters;
Step 4: - The set of rules is repeated via exchange packages of Steps 2 and 3 till convergence.

In each re-assignment step, the new clusters change into constant and in ultimate case of non-overlapping clusters, the knowledge vector that belongs to one decided on cluster will get more contribution to a given point's residual. Thus, the figure 3 shows, the coefficients of information vector not belonging to this cluster are minimized.

While the existing K-Means approach applies mean calculations to evaluate the final clustering, the K-Means based on CS theory generates the update dictionary by training operation to provide the optimal and minimum number of clusters [10].

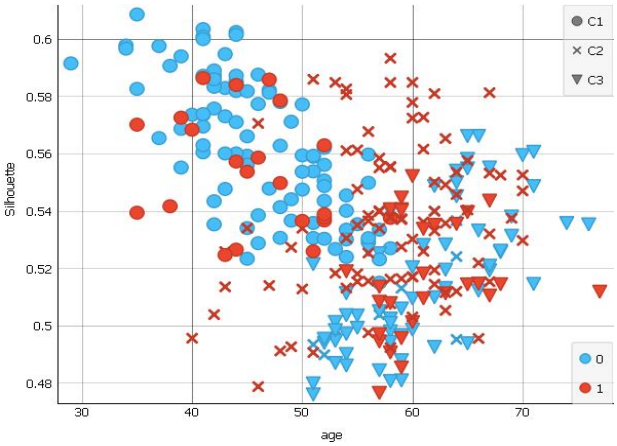

Figure 3: Optimal No. of Clusters

3.2.1. Consensus functions

Clustering ensembles have emerged as an impressive manner for bettering each the robustness as well as the steadiness of unsupervised classification solutions [10]. However, discovering a consensus clustering from multiple partitions is a difficult downside that can be approached from graph-based, combinatorial or statistical views. This learns about extends earlier analysis on clustering ensembles in several respects [7].

This proposed multistage means, introduce a unique consensus serve as is proposed in response to a twin similarity measure - the similarity between initial clusters and the similarity between candidate clusters and unsure objects. Our consensus serve as does now not require a given cluster quantity as a parameter. The set of rules is hence referred to as Dual-Similarity Consensus Function, DSCF. In this segment, we address this problem and describe the consensus purposes which might be used to combine independent clustering into one ultimate consensus clustering. Here, additional investigated the performance of three consensus clustering algorithms included in the proposed multistage scheme. These methods are cluster-based similarity partitioning algorithm (CSPA), hypergraph partitioning set of rules (HGPA), and the metaclustering algorithm (MCLA).
Let us denote the data set being investigated by,
$DS = \{ds_1, ds_2, \ldots, ds_n\}$. …. equ. (3)
The clustering ensemble on this data set will be denoted by
$CL = \{CL^{(1)}, CL^{(2)}, \ldots, CL^{(k)}\}$, …. equ. (4)
Where for each clustering $CL^{(i)}$, the whole set *DS* is a disjoint union of the classes in this clustering so that
$C(i) = \{C^{(i)}_1, C^{(i)}_2, \ldots, C^{(i)}_{ki}\}$, …. equ. (5)

DSCF uses the set of all clusters in all clustering's of the ensemble as features for its feature vectors. For each element $ds \in DS$ and each cluster $C^{(i)}_j$, the $C^{(i)}_j$th component of the feature vector of *d* is set to 1 if *ds* belongs to $C^{(i)}_j$, and it is set to 0 otherwise. The dual similarity clustering function is then used to cluster this set of feature vectors in order to find the consensus clustering.

3.3. Classification using SVM

Classification is without doubt one of the device studying tasks while gadget finding out algorithms are used to gather some knowledge from the set of given knowledge, seek for positive patterns and after that make their very own choices in accordance with the discovered facts. Supervised learning is a process where the goal is to generate a choice fashion that may appropriately classify unknown instances based on a style construct at the coaching set the place the classes of cases are identified. Besides settling on the best characteristic set, classification accuracy depends upon the classifier [6]. Each classification approach has some parameters that affect the accuracy of the created type. In this analysis, the give a boost to vector device was used for classification. SVM parameters and their position in classification are defined in the following section. SVM is in reality a binary classifier however in this day and age there are other techniques which can be enabled to a binary classifier to accomplish multi-classification.

The strengthen vector device separates circumstances from different classes via hyperplane which exists if the dataset contains linearly separable cases. The optimal hyperplane divides datasets so all circumstances from the similar class are on the one facet of the hyperplane and it is so far as imaginable from all cases.

SVM classifiers are based on the class of hyperplanes, $(w.x) + b = 0$ $w \in RN$, $b \in R$, corresponding to decision functions $f(x) = \text{sign}((w.x) + b)$. We can show that the optimal hyperplane, defined as the one with the maximal margin of separation between the two classes. In practical use, the user specifies the kernel function; the transformation $\varphi(.)$ is not explicitly stated. Given a kernel function K(xi, xj), the transformation $\varphi(.)$ is given through its Eigen purposes (a concept in useful analysis). Eigen purposes can be tough to build explicitly. This is why we specify the kernel function without being concerned concerning the actual transformation. There exists another view that, kernel serve as being an inner product, is really a similarity measure between the objects.

4. EXPERIMENTAL RESULTS

The proposed model is experimented with 6 different datasets in more than a few domains and sizes. These datasets are from the UCR Time-series Data Mining Archive. The measurement of time-series in those datasets levels from 28 to 6,000 records, and their exact dimension. This set is chosen as a result of it's of more than a few numbers of clusters, other cluster shapes and density accommodates noise points, and used in many articles within the literature as a benchmark. These datasets have two units within the repository, particularly TRAIN and TEST.

In this study the TEST set is used as it includes large datasets and the TRAIN sets is used to visualise the results for the sake of simplicity. The actual global datasets decided on from the UCR repository had been examined for clustering by other researchers and used as benchmarking for comparability.

4.1. Performance Evaluation Metrics

i) Reduction ratio

Dimensionality reduction refers to tactics that scale back the choice of enter variables in a dataset. More enter features steadily make a predictive modelling job more difficult to style, more typically known as the curse of dimensionality. Figure 4 shows the comparison of reduction ratio between the existing and proposed system.

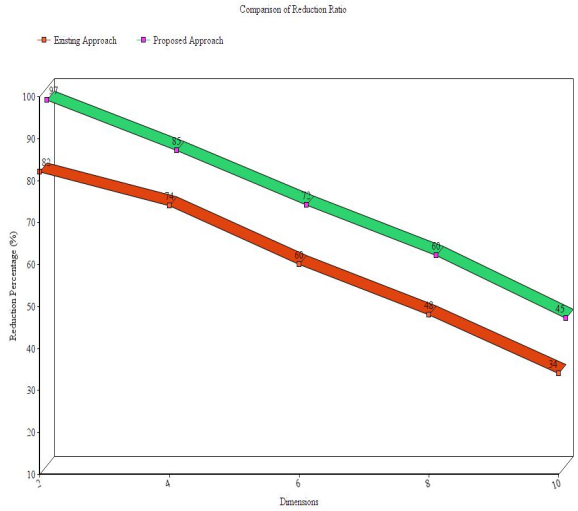

Figure 4: Comparison of reduction ratio

ii) Number of Clusters

Getting the optimum number of clusters could be very significant within the research. If K is too top, every level will extensively get started representing a cluster and if K is simply too low, then knowledge points are incorrectly clustered. Finding the optimal collection of clusters ends up in granularity in clustering. Thus, figure 5 shows result of the comparison of cluster minimization between existing and proposed approach.

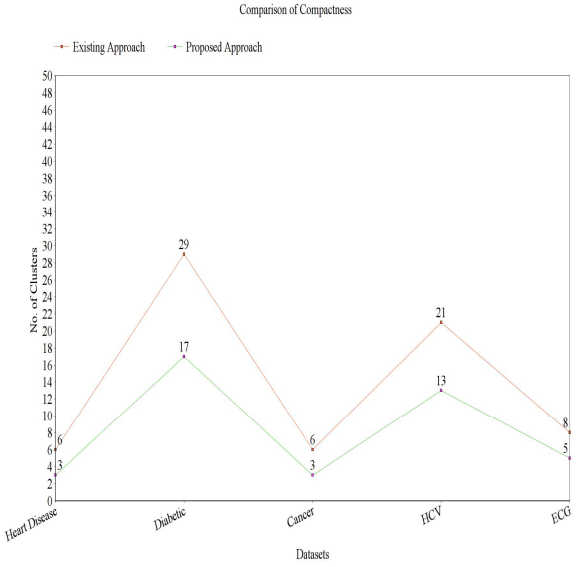

Figure 5: Comparison of Cluster Minimization

iii) Accuracy

It determines how close the dimension comes to the real worth of the quantity. So, it indicates the correctness of the end result. Maximum effort has to be taken to procure accuracy in information. The high quality of the dimension depends on the accuracy of the entire data. It can be restricted via elements like board answer or environmental noise. Thus, figure 6, illustrate the result of accuracy.

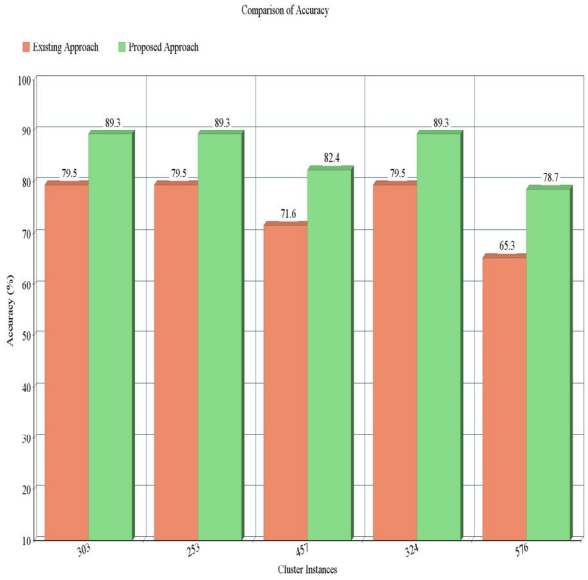

Figure 6: Accuracy based on Cluster Instance

iv) Compactness

Compactness measures the typical pairwise distances between points in the same cluster: All of our test data units have external true labels which aren't utilized by the clustering process. Thus, figure 7 shows the result of compactness between the existing and proposed approach.

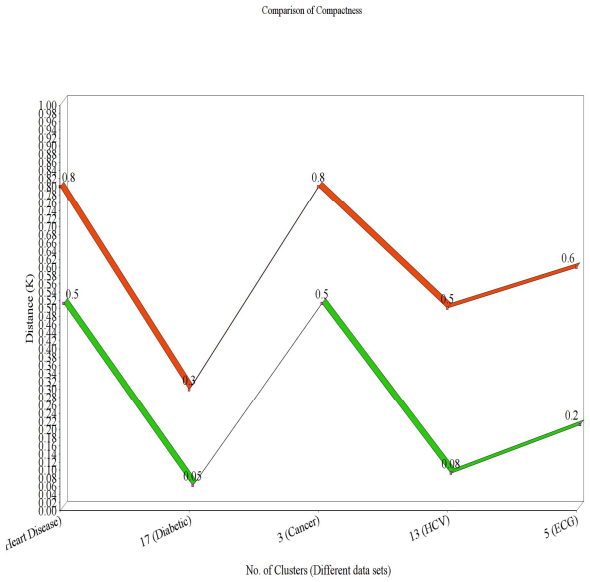

Figure 7: Comparison of Compactness

Moreover, the proposed system was once compared to different comparable paintings current within the literature. The Comparison shows that the proposed gadget has achieved the most efficient efficiency over the opposite methods.

5. CONCLUSION

This analysis introduces the very important wishes of Supervised and unsupervised studying for medical information classification. The proposed a multistage approach in accordance with consensus serves as clustering for medical knowledge classification. The gadget involves 3 major stages. In the first phase implies dimensionality reduction algorithm, and the second section applied the multiple unsupervised clustering algorithms and after all, those fast classification algorithms categorized the entire knowledge set in this sort of approach that environment friendly and accurate profiling of very large and extremely dimensional medical knowledge sets can be accomplished.

REFERENCES

[1] Sharma, R., Singh, S. N., & Khatri, S. (2016, February). Medical data mining using different classification and clustering techniques: a critical survey. In 2016 Second International Conference on Computational Intelligence & Communication Technology (CICT) (pp. 687-691). IEEE. https://doi.org/10.1109/CICT.2016.142

[2] Lashari, S. A., Ibrahim, R., Senan, N., & Taujuddin, N. S. A. M. (2018). Application of Data Mining techniques for medical data classification: A review. In MATEC Web of Conferences (Vol. 150, p. 06003). EDP Sciences.

[3] Abdullah, M., Al-Anzi, F., & Al-Sharhan, S. (2018, March). Hybrid multistage fuzzy clustering system for medical data classification. In 2018 International Conference on Computing Sciences and Engineering (ICCSE) (pp. 1-6). IEEE.

[4] Kelarev, A., Stranieri, A., Yearwood, J., & Jelinek, H. (2012, June). Empirical investigation of consensus clustering for large ecg data sets. In 2012 25th IEEE International Symposium on Computer-Based Medical Systems (CBMS) (pp. 1-4). IEEE. https://doi.org/10.1109/CBMS.2012.6266364

[5] Viswanath, S., & Madabhushi, A. (2012). Consensus embedding: theory, algorithms and application to segmentation and classification of biomedical data. BMC bioinformatics, 13(1), 26. https://doi.org/10.1186/1471-2105-13-26

[6] Janardhanan, P., & Sabika, F. (2015). Effectiveness of support vector machines in medical data mining. Journal of communications software and systems, 11(1), 25-30. https://doi.org/10.24138/jcomss.v11i1.114

[7] Fern, X. Z., & Brodley, C. E. (2006). Cluster ensembles for high dimensional clustering: An empirical study.

[8] Aghabozorgi, S., Ying Wah, T., Herawan, T., Jalab, H. A., Shaygan, M. A., & Jalali, A. (2014). A hybrid algorithm for clustering of time series data based on affinity search technique. The Scientific World Journal, 2014.

[9] Liu, S., Triantis, K. P., Zhao, L., & Wang, Y. (2018). Capturing multi-stage fuzzy uncertainties in hybrid system dynamics and agent-based models for enhancing policy implementation in health systems research. PloS one, 13(4), e0194687.

[10] Donoso, F. I., Figueroa, R. L., Lecannelier, E. A., Pino, E. J., & Rojas, A. J. (2013, July). Clustering of atrial fibrillation based on surface ECG measurements. In 2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) (pp. 4203-4206). IEEE.

[11] Wosiak, A. (2019). Principal Component Analysis based on data characteristics for dimensionality reduction of ECG recordings in arrhythmia classification. Open Physics, 17(1), 489-496. https://doi.org/10.1515/phys-2019-0050

[12] Liu, H., Zhao, R., Fang, H., Cheng, F., Fu, Y., & Liu, Y. Y. (2017). Entropy-based consensus clustering for patient stratification. Bioinformatics, 33(17), 2691-2698.

[13] Balouchestani, M., & Krishnan, S. (2016). Advanced K-means clustering algorithm for large ECG data sets based on a collaboration of compressed sensing theory and K-SVD approach. Signal, Image and Video Processing, 10(1), 113-120.

[14] Abdi, H., & Williams, L. J. (2010). Principal component analysis. Wiley interdisciplinary reviews: computational statistics, 2(4), 433-459.

[15] Wosiak, A., & Zakrzewska, D. (2014, September). Feature selection for classification incorporating less meaningful attributes in medical diagnostics. In 2014 Federated Conference on Computer Science and Information Systems (pp. 235-240). IEEE.

[16] Wosiak, A., & Zakrzewska, D. (2018). Integrating correlation-based feature selection and clustering for improved cardiovascular disease diagnosis. Complexity, 2018.