Received: December 30, 2016

INASS

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

Hybrid Approach for Data Classification in E-Health Cloud

Thangavel Muthamilselvan[1]* **Balamurugan Balusamy[1]**

[1]*School of Information Technology & Engineering, VIT University, Vellore, Tamilnadu, India.*
* Corresponding author's Email: tmuthamilselvan@vit.ac.in

Abstract: The growth of IT industry technology is absorbed by the cloud service technology, which leads to secure connectivity and availability of services to cloud users. This paper proposes an e-health data classification system that gives the services to e-health cloud user for their disease prediction requirements. The tree bagging method is suitable to select better weighted attributes and the careful seeding of K-means ++ algorithm improves the accuracy and speed of clustering. Most of the tree classification methods use only information gain as the strategy to select suitable attributes for classification. We used information gain in bagging technique to improve accuracy. In this research article, we proposed a method to Coalescing Decision tree classification algorithm, bagging technique, and K-means++ algorithm to build a better classifier for the e-health cloud users. It was evaluated with the standard data sets such as Diabetes, breast cancer, liver disorders and cardiotocography.

Keywords: Classifiers, cloud user, C4.5 decision trees, K-means++, prediction accuracy, clustering.

1. Introduction

The cloud computing becomes popular in the IT industry because of its easy implementation and its nature of cost effective maintenance to the cloud users. Any user can get their required software, platform or even virtual machine from the cloud environment and they can use and pay as per their use. The cloud environment has other benefits too other than cost; those are flexibility, scalability and less maintenance [1,2]. In recent years, the healthcare providers understand that providing their services by e-health cloud environment is best in terms of cost, fast communication and easy administration of their health care environment and also IT infrastructure of their surroundings. In overall, the e-health cloud provides excellent opportunity to give better services to their customers in affordable treatment cost [3]. Naturally, the e-health environment provides a significant amount of health records, and it is a perfect place for analyzing the large health data and extracting the hidden knowledge, this task is termed as knowledge discovery or data mining. The knowledge discovery involves many stages such as preparing the data to a suitable format for data analysis, creating models for several hidden pattern extraction, training the models using the training set data, testing the models to check whether it gives desired level of accuracy results or not. Then these models can be used to predict the new patterns or association among the given new data set. In knowledge discovery science, there are several techniques used to predict the new patterns such as classification, clustering, and association rule mining, etc. In this paper, we are proposing a hybrid approach for better classification of health data, which involves decision tree classification learning technique and clustering technique. The decision tree is used to find the best set of attributes, and then the best set is utilized in the clustering process. Bagging technique also used to address the large set of data. The proposed system is a combination idea from bagging technique, C4.5 decision tree classifier, and K-means++ algorithm.

One of the data set in our experiment is Diabetes PIMA data set. Diabetes is not a disease; it is a metabolic disorder by that blood glucose level increase unconditionally. Insulin is a hormone which is produced by pancreatic beta cells; this hormone is mixed with blood and travels to body

cells. The insulin helps in consuming of glucose from the blood by the body cells [4]. If there is a defect in insulin production, then the glucose consumption of the body cells from the blood is affected. So there is an increase in blood glucose. Diabetes leads to a variety of health problems like heart disorders, blood pressure, vision defects or blindness and many. Our proposed system experimented with diabetes data set and other data sets too.

E-health environment provides a significant number of health related data; there we use knowledge discovery techniques to analyze the data. Classification is one of the essential knowledge discovery techniques. To strengthen the e-health computational intelligent environment, we need to enhance the classification process.

The proposed classification system is for large e-health data set. The many existing classification algorithms are not able to perform well [17] due to the well-known process of attribute selection. Our method combines bagging algorithm and decision tree C4.5 for attribute selection for enhanced attribute selection. The new feature of the proposed system is eliminating the overriding of attribute selection in large training data set. By The proposed combination of bagging and decision tree in the weighted attributes selection overcomes the problem. Assignment of weights to attributes leads to attribute reduction. Because of the attribute reduction and attribute selection, prediction accuracy increased compared to other classification methods. Then k-means++ algorithm is applied to predict the class label of new instances.

The rest of the paper is organized as follows. Section 2 discusses the literature survey. In section 3, discussion about bagging and cross-validation, Decision tree and clustering are given. Section 4 deals with the proposed system. Section 5 provides about class label prediction of new objects. Section 6 presents excremental results and discussion. In section 7, the paper is concluded.

2. Literature survey

Healthcare data management in e-cloud is pioneer one; many types of research are going on this field. One of the systems (CBIHCS) cloud-based intelligent health care systems [14] adapts real-time monitoring in healthcare environment. Body sensor components are used to monitor regularly, and it is stored in the cloud repository for the further analysis. Various data intelligent systems are attached to extract hidden information to help the users of the system.

Information technology grows faster and produces an enormous amount of data [15]. The advancement in IT infrastructure leads to the introduction of cloud establishments. The dynamic streams of data from various automated data collection or generation sources populate vast amount of data. Those data are heterogeneous in nature. These set of data is called Big data. It has the characteristics of value, velocity, veracity, volume. A lot of electronic health records (HER) generated from the cloud-based health servers become the health care Big data (HBD).

When the age of people increases, the problem of heart disorders also increases. One of the equipment used to analyze the rhythms and disorder of heart is ECG (electrocardiography). It gives only the measurement of heart rhythms and functions. Cardiac guard cloud services [4] used the hybrid classification approach using SVM (support vector machine) and RT (random tree) for analyzing the heart disorders. It can classify six types of cardiac disorders. Cardiograph (ECG) is used as the critical equipment to detect heart disease and to give treatment.

Knowledge identification from huge data is finding of various patterns hidden in the data. The decision tree has the vital place in the knowledge discovery, but the simple decision tree has the sharp decision boundaries. In the real world, many situations has fuzziness in taking the decision, so that we have to identify a decision-making tool which gives the correct decision for the fuzzy dataset. PCA (Principle Component Analysis) is a technique for the dimensionality reduction. After dimensionality reduction, decision tree improves its performance. Modified fuzzy SLIQ technique utilized the modified Gini index based fuzzy and PCA to classify the given dataset [16].

Constructing a decision tree for a large amount of data provides a poor decision. The technique called bagging [5], by which the training dataset is split into several partitions, and then one decision tree is created for each partition. Finally, all the trees are combined into one. This method gives us the opportunity to use parallel processing to create the tree models faster. A minimum number of decision trees creation and combining them increases classification performance in the bagging method.

From the above analysis, there are only a few works are carried out in the classification of e-health large data set. In most of the tree based classification techniques use only information gain for attribute selection. This motivated us to deal with e-health data in cloud environment with the new approach for weighted attribute selection and utilizing

improved clustering in predicting class label of new objects. So we proposed a system to assist the authenticated cloud users in classifying their diseases.

3. Bagging and cross-validation

The prediction accuracy of any classifier is based on classification algorithm and training data set used. In bagging technique [5], the overall training dataset is partitioned into different subsets, then a particular classification algorithm is applied to each training sub-dataset, and classification models are generated. Finally, all the generated classification models are combined and created as a final classifier. Boosting is another approach to handling training dataset [6]. The desired classification algorithm is trained on each training sub-datasets. Then each data models of each training sub-datasets are tested. The incorrectly classified examples are given more weight to get more importance. Again the weighted examples are used to construct data models. Finally, all the generated classification models are combined and created final classifier [7].

The cross validation is a strategy of how to divide the overall training dataset into subsets. The training dataset D is divided into N subsets in the size of D/N. "Bagging like strategies" is a strategy with combines bagging and cross-validation methods. Different methods are available to divide training data, such as small bags, no replication small bags, disjoint partition and disjoint bags [5]. Consider the following example as overall dataset in Fig. 1.

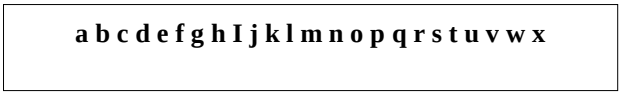

Figure.1 Original dataset

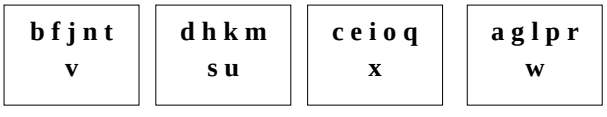

Figure.2 Disjoint partition

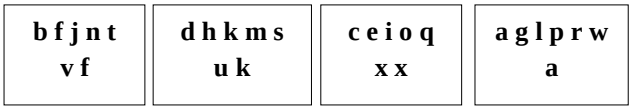

Figure.3 Disjoint bags

In disjoint partition technique, each data is selected randomly at only once from the original dataset and place it in any one of the subsets as Fig. 2. If we union all the sub-datasets, then we get original dataset.

Disjoint bags partition technique is same as disjoint partition, but any number of elements in a subset is randomly selected and replicated within the subset in the same number of times in all subsets as Fig. 3. Disjoint bags produces better output, comparatively.

3.1 Bagging algorithm

1. Get the overall training dataset D.
2. Divide the overall training dataset D into N subsets using "bagging like strategies".
3. Create one classification model for each training subsets using a particular classification algorithm that is N number of classification model.
4. Combine all the classification model and create a final classifier $H=H_N$ $(N=1\ldots n)$.
5. A test example object data d_i is classified in to a class label c_j .

The combined classifier is given as Eq. (1).

$$H(d_i, c_j) = \sum_{n=1}^{n} w_n H_n(d_i, c_j) \quad (1)$$

Where d_i is the test object data, c_j is class label of the test object data, H_n is each classification model, w_n is the weight of each classification model based on its accuracy level.

3.2 Decision tree

Decision tree classification is a well-known data mining strategy. It works well for noisy data and reduces the error, acceptably. So it is best for noisy and missing value dataset. It classifies the given dataset based on various attribute values. The classified attribute is called class label. Input for the algorithm is training dataset, attribute list and attribute selection method. The decision tree algorithm C4.5 is the improved version of ID3 version. The difference between them is how they do attribute selection. ID3 uses information gain to select best attribute set, which gives how much percentage an attribute will split the training dataset.. But the information gain has some bias problem; it selects an attribute which has a large number of distinct values. This bias is overcome by C4.5 algorithm successor of ID3.

$$\text{Entropy, S} = -\sum_{i=1}^{n} p_i \log_2 p_i \quad (2)$$

Where n is Total number of classes used in the learning model (1, 2, 3…n), p_i is Probability value that the data belongs to the class i.

The information gain is defined as the metric for analyzing the splitting criteria based on the entropy value as the Eq. (2). The information gain for each of the attribute is calculated as per the Eq. (3).

$$\text{Gain(A)=Entropy(S)} - \sum_{k=1}^{m} \frac{|S_k|}{|S|} \text{Entropy}(S_k) \quad (3)$$

Where Entropy(S) is entropy value for entire training dataset, Entropy(S_k) is the entropy value for the attribute 'k', k is the counter value for the attribute (1,2,3,...n). In decision tree C4.5 algorithm SplitinfoK is as Eq. (4).

$$\text{Splitinfo}_k = -\sum_{j=1}^{v} \frac{|S_J|}{|S|} \times \log_2 \left(\frac{|S_J|}{|S|}\right) \quad (4)$$

Where v is number of partition for the attribute k, S_J is number of objects in each partition of attribute k, s is total number of object in training data set.

$$\text{GainRatio(A)} = \frac{Gain(A)}{Splitinfo_k(s)} \quad (5)$$

The attribute which gets maximum GainRatio as Eq. (5) is selected as the splitting attribute.

3.3 Clustering

Clustering is a technique of grouping similar objects. K-means clustering technique is using partition algorithm and iterative approach to finding the clusters. Initially, an overall training dataset is partitioned into K number of arbitrary subsets and the initial centroid of each subset also arbitrarily selected, so it becomes NP-hard problem. The running time to find the final cluster is super polynomial with respect to input size [8].

The K-means++ algorithm overcomes above problems by new selection process of initial centroids. The first centroid is selected randomly from the given dataset. The remaining centroids are selected by probability proportional to squired distance from the rest of the data elements in the dataset [9].

Algorithm of K-means++

1. Get dataset and initialize number of cluster K and initialize cluster center set C as empty.
2. Select a centroid c_i randomly from the dataset D and add to cluster center set C.
3. While $|C| < K$ do
   i. Select data element x from the rest of the dataset with probability proportional to $D^2|x|$
   ii. $C \leftarrow C\{x\}$
4. The initial cluster centers are chosen; then continue with the k-means clustering algorithm procedure.

This seeding technique gives the considerable accuracy performance when compared to K-means algorithm.

4. Proposed work

Any authenticated cloud users may get the access of the cloud server. They can populate training dataset, and the new data object to be classified to the server. Then the user will get the class label of the new data object. This system uses the hybrid approach of decision tree C4.5, the k-means++ clustering algorithm, training data portioning technique and training dataset to obtain an overall result. This system includes five phases. This system includes five phases. In the phase one, the overall training dataset is split using disjoint bags technique. In the second phase, Decision tree model using C4.5 algorithm is constructed for each of the training sub-dataset (disjoint bags).In phase three, the weight of the each attribute is calculated from all the decision tree models. In phase four, these weights with attribute list used to construct the clusters using K-means++ algorithm and K centroids are generated using majority voting scheme. In final fifth phase, is a prediction phase, a new data object is given and its class label is determined. The class label is given to the user. The overall architecture is given in Fig. 4.

4.1 Weighting process of attributes and determining cluster centers

The weighting process is done to select list of attributes for the next process of clustering. This process is explained by diabetes dataset. Table 1 gives the attribute discretion of the dataset, which has 8 predictor attributes and one class label which tells whether the patient is diabetic or non-diabetic. Table 2 provides the sample diabetes dataset.

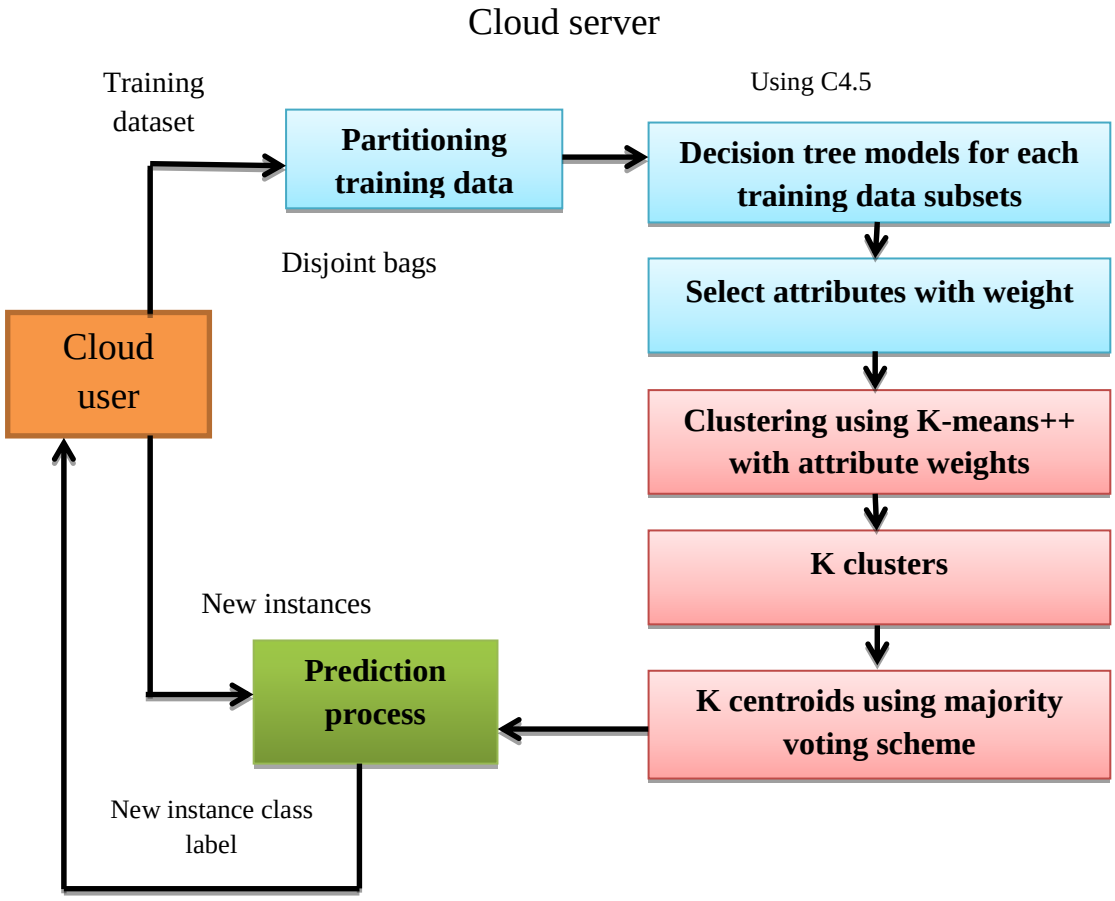

Figure. 4 Architecture of the system

Table 1. Description of Diabetes dataset

Attribute	Abbreviations	Specification
1	Pregnant	Number of times pregnant
2	Glucose	Plasma glucose concentration at 2 h. in an oral glucose tolerance test-(mg/dl)
3	DBP	Diastolic blood pressure-(mm Hg)
4	TSFT	Triceps skin fold thickness-(mm)
5	INS	2-Hour serum insulin-(mu U/ml)
6	BMI	Body mass index-(kg/m2)
7	DPF	Diabetes pedigree function
8	Age	Age
CLASS		*YES, tested* Positive:(diabetic)NO, tested Negative:(non- diabetic)

Table 2. Sample dataset of Diabetes

Input Attributes								Class Label
Pregnancy	Glucose	TSFT	BMI	Age	DBP	DBF	INS	Diabetic
5	175	22	15	20	60	0.01	30	0
5	190	40	50	55	90	0.08	65	1
14	190	35	35	45	90	0.07	80	1
17	150	15	15	20	50	0.02	25	0

Decision tree C4.5 is created for each partitioned training dataset. We get a different tree for each partition. That is different root nodes, and siblings depend on each sub training dataset. The weight of each attribute is differs based on the tree size and the position in which it appears in the tree model. Fig. 5 gives the sample tree created by one of the sub training dataset. Consider the Fig. 5, say first data partition decision tree model M, the height of the

tree is 3 and levels are 0,1,2,3. Weight of the Model M_i and attribute of A_k is as Eq. (6).

$$\text{Model\_weight}_{i,k} = (\text{Height\_M}_i - j + 1) / (\text{Height\_M}_i + 1) \quad (6)$$

Where $\text{Model\_weight}_{i,k}$ is the weight of attribute k on the created model i, Height_M_i is maximum levels of the model, j is level of attribute node k.

The weight of the attribute pregnancy frequency in level zero is (3-0+1) / (3+1) =1. Similarly, weight of plasma glucose and diastolic blood pressure in level one is (3-1+1) /(3+1)= 0.75, both have same weight because they are on same level, one [10]. If the same attribute is available in the next high level of the same tree, we can ignore to consider its weight. So in multiple occurrences of the same attribute in the same tree, the lower level weights only considered. After calculation of weight of each attribute from all the tree models, then the average weight of each attribute is calculated by Eq. (7).

The weight of the attribute pregnancy frequency in level zero is (3-0+1) / (3+1) =1. Similarly, weight of plasma glucose and diastolic blood pressure in level one is (3-1+1) /(3+1)= 0.75, both have same weight because they are on same level, one [10]. If the same attribute is available in the next high level of the same tree, we can ignore to consider its weight. So in multiple occurrences of the same attribute in the same tree, the lower level weights only considered. After calculation of weight of each attribute from all the tree models, then the average weight of each attribute is calculated by Eq. (7).

$$\text{Wavg}_k = (\text{Model\_weight}_{1,k} + \text{Model\_weight}_{2,k} + \ldots. \text{Model\_weight}_{n,k}) / \text{Total \# of models.} \quad (7)$$

Where Wavg_k is the average weight of an attribute k, Total # of models is the a total number of decision tree models.

If any of the attribute weight is zero, eliminate the attribute from the attribute list, for the next clustering process. So this weighting process [11] gives us the importance level of each attribute and it eliminates few attribute to be used in clustering. It decreases clustering time and improves the accuracy of clustering.

Algorithm to Weighting process of the attributes and determining cluster centres.

Input:
D: Training dataset.
C: Number of partitions to be done on overall training dataset(#disjoint bags).
K: Number of clusters to be created.
Output:
K: Cluster canters.

Algorithm

1. *Let TWA_k is total weight of a attribute in each model.*
2. *Divide the overall training dataset into C subsets using Disjoint bags technique.*
3. *Construct C4.5 decision tree models for each of the disjoint bags, Tree models are M_i , i= 1..C .*
4. *For I = 1 to C*
5. *Let $Model\_weight_k$ is the each attribute weight = 0*
6. *For each level j of each model M_i*
7. *If $Model\_weight_k$ = = 0*
8. *$Model\_weight_k = (Height\_M_i - j + 1) / (Height\_M_i + 1)$*
9. *End if*
10. *End for*
11. *$TWA_k = TWA_k + Model\_weight_k$*
12. *End for*
13. *Find the average of weight of each attribute form the attribute list by dividing number of models. $TWAavg_k = TWA_k / C$*
14. *If any attribute weight is zero then eliminate form the attribute list for further clustering process.*
15. *Calculate normalized weight of attribute list.*
16. *Use the weighted attribute list, training dataset and K-means ++ algorithm to find K clusters.*
17. *For each cluster centres assign class label using majority vote scheme.*
18. *Return C cluster centres.*

In the clustering process, weight of each attribute is used. The K-means++ algorithm generates K clusters. The voting scheme is used to assign class label to the centroid of each cluster. The voting scheme [12] is finding the class label which has maximum number of objects in the cluster that class label is assigned to the cluster centre.

5. Class label prediction of new objects

The new object, the class label to be predicted is received from cloud users and given to our system. It finds the class label of given data object and

returns the class label. This module finds the distance between new object and each cluster centers, chooses the nearest cluster center. That nearest cluster center's class label is assigned as the class label of the new object.

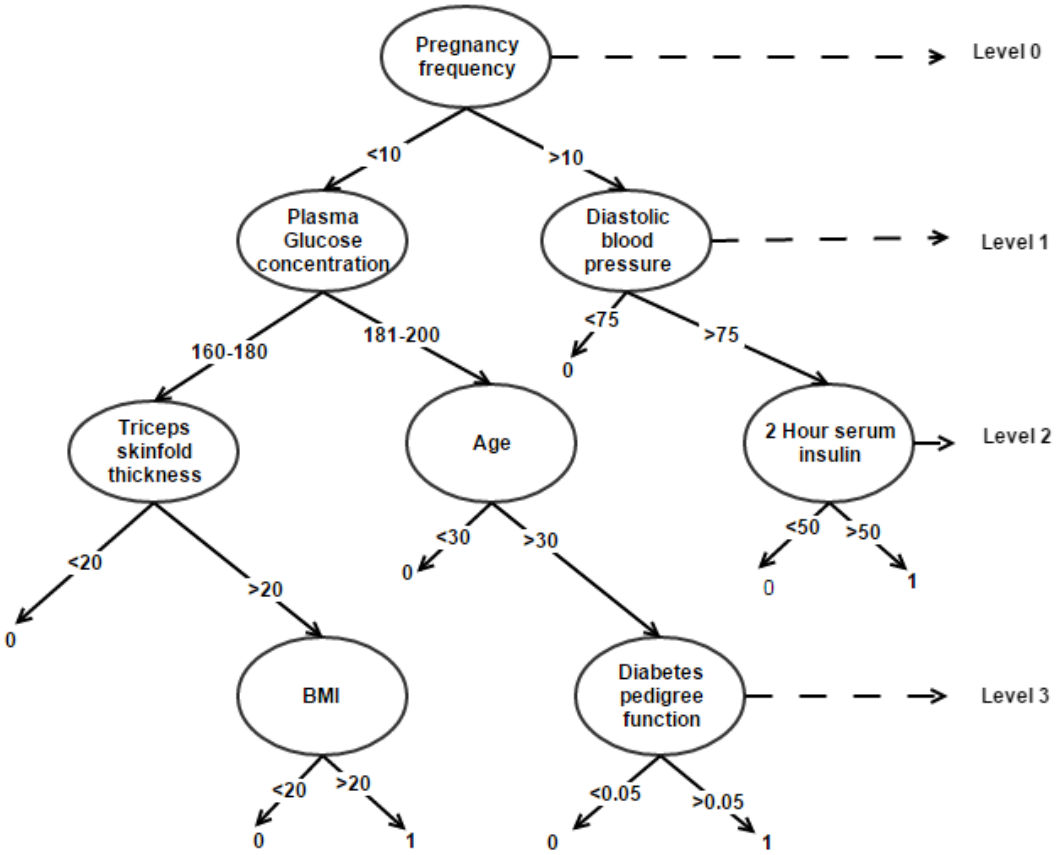

Figure.5 One Decision Tree model for a sub training dataset

Table.3 UCI dataset description used in the experiments

Attribute	Cardiotography Specification	Breast-cancerSpecification	Liver-Disorders Specification
1	FHR baseline (beats per minute)	Sample code number	Age of the patient
2	# of accelerations per second	Clump Thickness	Total Bilirubin
3	# of fetal movements per second	Uniformity of Cell Size	Direct Bilirubin
4	# of uterine contractions per second	Uniformity of Cell Shape	Alkaline Phosphotase
5	# of light decelerations per second	Marginal Adhesion	Alamine Aminotransferase
6	# of severe decelerations per second	Single Epithelial Cell Size	Aspartate Aminotransferase
7	# of prolongued decelerations per second	Bare Nuclei	Total Protiens
8	percentage of time with abnormal short term ariability	Bland Chromatin	Albumin
9	mean value of short term variability	Normal Nucleoli	Albumin and Globulin Ratio
10	width of FHR histogram	Mitoses	Class label
11	minimum of FHR histogram	Class label:benignmalignant	
12	Maximum of FHR histogram		
13	# of histogram peaks		
14	# of histogram zeros		
15	histogram mode		
16	histogram median		
17	histogram variance		
18	Histogram tendency		
19	Class label:fetal state class code (N=normal; S=suspect; P=pathologic)		

6. Experimental results and discussion

To evaluate our proposed system (HCK++) of Hybrid data classification using decision tree C4.5, Bagging and K-means++ algorithm. We used four data sets such as PIMA, breast cancer, liver disorders and cardiotocography from the standard knowledge discovery data repository, UCI. PIMA diabetes data description is given in the table1, and the remaining dataset description is given in the table 3.

Using the decision tree and bagging technique, we do attribute selection for the best performance of K-means++ algorithm. The number of decision tree model is determined by the training dataset partition technique that is n fold cross validation and disjoint bags. So the number of decision tree models created is depends on the number of objects in the training dataset.

In some clustering practice, there are methods to estimate number of clusters. In K-means algorithms, the value of K is the number of clusters to be created. The good method to find [13] number of clusters is $\sqrt{}(n/2)$ where n is the number of data objects in the training dataset and the approximate number of objects in each clusters is $\sqrt{}(2n)$.

The table 4 shows the number of clusters created for each dataset used. The accuracy analysis is the most promising metric to judge any classifiers. Especially in the health care systems, accurate diagnostics the first requirement in any health-related techniques. Accuracy of any classifier is measured by the fraction of the number of correctly classified instances, and the total number of classification completed successfully in the testing phase of the classifiers as Eq. (8).

$$\text{Accuracy of a classification} = \frac{CT}{TT} \quad (8)$$

Where *CT* is number of correctly classified instances and *TT* is a total number of classifications completed successfully.

Analyzing the performance is the prominent task in any classifier design, and the accuracy of any classifier is established based on the algorithm and also depending on the number of training and testing objects also. But a classifier algorithm should work with a variety of datasets. The table 5 gives the measure of the accuracy of various algorithms and our proposed approach (HCK++). The graph in Fig. 6, shows the pictorial representation of the accuracy of decision tree C4.5, K-means clustering algorithm, K-means++ clustering algorithm and our proposed hybrid approach (HCK++) for each standard dataset. From the graph, it is evident that our system gives comparatively better accuracy results than other approaches. Existing most of the classification algorithms are using only information gain technique to select the best attribute for classification [16,17]. But our HCK++ approach uses decision tree bagging to get weighted attribute and the weighted attributes applied in creation of clustering. Based on the clusters new instances are classified. Our proposed method is better than existing techniques because, in the method of Appling the full training data set to find weighted attribute, some of the impotent attribute weights are overridden by forthcoming training data. But by using bagging method, important attribute weights are preserved to the current bagged data set. The advantage of careful seeding of k-means++ algorithm improves the accuracy and speed of the clustering. Since our approach integrates the benefits of bagging in decision tree and careful seeding of clustering, and it improved the accuracy. Any authenticated cloud users can populate their dataset to the system, and then they can use the developed classifier number of time for their prediction of the unknown data object.

Table. 4. Estimation of number of clusters for each dataset

Dataset	Instances	Number of clusters
PIMA	768	20
Brest cancer	29418	120
Liver-Disorders	8930	66
Cardiotography	2126	33

Table 5. Prediction accuracy of various algorithm

Dataset	Accuracy in %			
	C4.5 Decision Tree	K-means	K-means++	HCK++
PIMA	72.36	66.72	78.45	91.01
Breast cancer	71.24	65.81	76.62	90.42
Liver-Disorders	71.32	65.94	75.32	88.79
Cardiotography	70.9	62.65	74.55	89.02

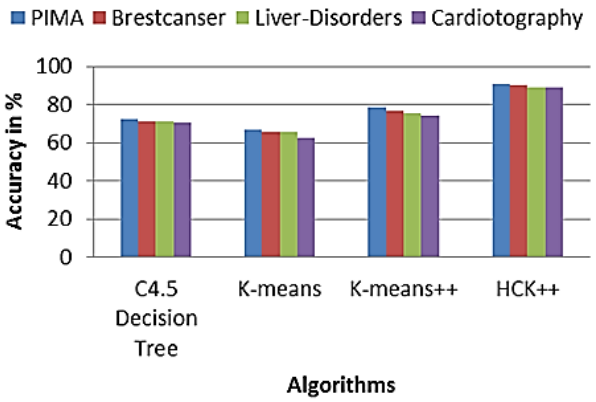

Figure. 6 Accuracy comparisons of algorithms

7. Conclusion

The hybrid approach for data classification in e-health cloud using decision tree, bagging technique and K-means++ helps the e-health cloud users to analyze their health data for personal benefits or organizational benefits. Our proposed system is analyzed and obtained its accuracy measures. The result of the research can be summarized in the following manner; our proposed HCK++ system produced high prediction accuracy when compared with other decision tree classification and clustering techniques. This system also gives great prediction accuracy in different medical data set such as PIMA, Breast cancer, Liver-disorder, and cardiotography. It outperformed. In future, this system can be improved using scalable clustering algorithm and to handle fuzzy dataset. Any authenticated cloud users can access the framework and get their classified label for their unknown data element.

References

[1] E. Şaykol, "On the Economical Impacts of Cloud Computing in Information Technology Industry", In: *Proc. of 5th International Conference on Eurasian Economies,* Skopje, Macedonia. pp.1-7, 2014.

[2] Z. Dong , N. Liu , R. R. Cessa, "Greedy scheduling of tasks with time constraints for energy-efficient cloud-computing data centers" ,*Journal of Cloud Computing*, Vol. 4, No.1 ,pp. 1-14, 2015.

[3] S.J. Miah, J. Hasan, G. Gammack, "On-Cloud Healthcare Clinic: An e-health consultancy approach for remote communities in a developing country", *Telematics and Informatics*, Vol. 34, No. 1, pp. 311-322, 2017.

[4] H. Chen, B.C. Cheng, G.T. Liao, T.C. Kuo, "Hybrid classification engine for cardiac arrhythmia cloud service in elderly healthcare management", *Journal of Visual Languages & Computing,* Vol. 25, No. 6 , pp.745-753, 2014.

[5] K. Machová, F. Barcak, P. Bednár, "A bagging method using decision trees in the role of base classifiers ", *Acta Polytechnica Hungarica*, Vol. 3, No. 2, pp. 121– 132, 2006.

[6] W. Martinez, J.B. Gray, "Noise peeling methods to improve boosting algorithms" , *Computational Statistics & Data Analysis* , Vol. 93, No.1, pp. 483-497, 2016.

[7] M.M. Fernándezs, I. Serrano, G. Bueno, O.Déniz, "Bagging Tree Classifier and Texture Features for Tumor Identification in Histological Images", *Procedia Computer Science*, Vol. 90, No.1, pp .99-106, 2016.

[8] Z. Friggstad , M. Rezapour , M. R. Salavatipour, "Local search yields a PTAS for k-means in doubling metrics", In: *Foundations of Computer Science, IEEE 57th Annual Symposium*, New Brunswick, New Jersey, pp. 365-374, 2016.

[9] D. Wei, "A constant-factor bi-criteria approximation guarantee for k-means++ ", In: *Advances in Neural Information Processing Systems*, Barcelona, Spain, pp.604-612, 2016.

[10] C. Kaewchinporn, N. Vongsuchoto, A. Srisawat, "A combination of decision tree learning and clustering for data classification", In: *Computer Science and Software Engineering, Eighth International Joint Conference*, Nakhon pathom, Thailand, pp. 363-367, 2011.

[11] C.H. Lee, "A gradient approach for value weighted classification learning in naive Bayes", *Knowledge-Based Systems*, Vol. 85, No.1, pp. 71-79, 2015.

[12] G.M. Dakhel, M. Mahdavi, "A new collaborative filtering algorithm using k-means clustering and neighbours' voting", In: *Hybrid Intelligent Systems (HIS),11th International Conference on IEEE*, Malacca, Malaysia, pp.179-184, 2011.

[13] J.Han, and M.Kamber, *Data Mining: Concepts and Techniques*, 3rd Edn., Morgan Kaufmann, San Francisco, 2011.

[14] P.D. Kaur, I. Chana, "Cloud based intelligent system for delivering health care as a service" ,*Computer methods and programs in biomedicine*, Vol. 113, No. 1 ,pp.346-359, 2014.

[15] K. Wan, V. Alagar, "Characteristics and classification of big data in health care sector", In: *Natural Computation, Fuzzy Systems and Knowledge Discovery 12th International*

Conference on IEEE, Changsha, China, pp. 1439-1446, 2016.

[16] V.V. Kamadi,A.R. Allam, S.M. Thummala, "A computational intelligence technique for the effective diagnosis of diabetic patients using principal component analysis (PCA) and modified fuzzy SLIQ decision tree approach", *Applied Soft Computing*, Vol. 49, No.1, pp.137-145, 2016

[17] T.R. Baitharu, S.K. Pani, "Analysis of Data Mining Techniques for Healthcare Decision Support System Using Liver Disorder Dataset" , *Procedia Computer Science*, Vol. 85,No.1, pp. 862-870, 2016.