


ITC 1/50 Information Technology and Control Vol. 50 / No. 1 / 2021 pp. 102-122 DOI 10.5755/j01.itc.50.1.25349	An Efficient Technique for Disease Prediction by Using Enhanced Machine Learning Algorithms for Categorical Medical Dataset	
	Received 2020/02/22	Accepted after revision 2021/01/20
	 http://dx.doi.org/10.5755/j01.itc.50.1.25349	

HOW TO CITE: Veera Anusuya, V., Gomathi, V. (2021). An Efficient Technique for Disease Prediction by Using Enhanced Machine Learning Algorithms for Categorical Medical Dataset. *Information Technology and Control*, 50(1), 102-122. <https://doi.org/10.5755/j01.itc.50.1.25349>

An Efficient Technique for Disease Prediction by Using Enhanced Machine Learning Algorithms for Categorical Medical Dataset

V. Veera Anusuya

Assistant Professor, Dept. of CSE, Sri Vidya College of Engineering and Technology, Tamil Nadu – 626005, India; e-mail: veeraanushya@gmail.com

V. Gomathi

Head of the Dept., Dept. of CSE, National Engineering College, Tamil Nadu – 628503, India; e-mail: vgcse@nec.edu.in

Corresponding author: veeraanushya@gmail.com

In the 20th century, it is evident that there is a massive evolution of chronic diseases. The data mining approaches are beneficial in making some medicinal decisions for curing diseases. But medical data may consist of a large number of data, which makes the prediction process very difficult. Also, in the medical field, the database may involve both small and extensive datasets. This creates the study of a complex one for disease prediction mechanism. Hence, in this paper, we intend to improvise the machine learning approaches for disease prediction of both large and small datasets. Among the various machine learning procedures, classification and clustering methods play a significant role. Therefore, we are planned to improvise the machine learning algorithms. Then we introduced the enhanced machine learning algorithms for classification and clustering technique in this work for obtaining better accuracy results for disease prediction. In this proposed method, a process of preprocessing is involved, which follows by Eigen vector extraction, feature selection, and classification. Further, the most suitable features are selected with the use of Multi-Objective based Ant Colony Optimization (MO-ACO) from the extracted features for increasing the accuracy of classification and clustering. Here

we have shown the novelty in every stage of the implementation, such as feature selection, feature extraction, and the final prediction algorithm stage. The proposed method is compared with the existing technique on the measure of precision, Normalized Mutual Information, execution time, recall and accuracy. Here we conclude with the solution having more accuracy for all kind of categorical datasets which includes both small and large scale datasets.

KEYWORDS: Homogeneous clusters, Multi-Objective based Ant Colony Optimization, Categorical data analysis, large dataset, small dataset.

1. Introduction

Due to the advancements in the standard of living, the occurrence of chronic disease has increased significantly. It is essential to predict these diseases at an early stage, which further helps to reduce the risk of developing chronic diseases. So the concept of disease prediction becomes famous and propagates among the researchers. Since the medical data is of wide range, the process of disease prediction involves vast data. In recent years, the clustering and classification of the categorical data play a pivotal role in the field of data mining, especially in the medical field. Training, Extrapolation, and prediction [1] are the types of function learning from data. Learning from the categorical data [50] is a new effort; at the same time, the numbers of absolute values are conceivable and become exponential [33]. The learning models of the categorical medical datasets used so far are Deep Neural Network (DNN), Gaussian Process Latent Variable Model (GPLVM), Long-short term memory (LSTM) and Categorical Latent Gaussian Process (CLGP). The neural network has achieved the dimension reduction process [27]. Input vectors have constructed the conversion from high dimensional data to low dimensional data. Recent developments are derived based on the sample by CLGP. Nowadays, the Gaussian process established models of the latent variable have created a wide-range of interest in the community of machine learning concerning the capacity.

It reproduces the other confirmation around the process of latent on absorbing the non-parametric constituents. However, these prototypes might be persistent to confirm opinion specific technique and acquires different latent space for each distinct view. It is not that easy to improve the available Gaussian process based on the latent variable model [17]. This model was not only capable of holding a multi observational location but also give exterior information

with classification labels for acquiring improved latent space prediction. However, there were some issues like over fitting, computationally expensive learning's that lead to an increase in the consumption of time; the standard classification methods require excess time for processing this vast amount of data. It also requires a massive number of records for attaining a better outcome. For data set having a large number of categorical attributes [39] besides lower samples, further formation of absolute data values might not happen among the samples of training [46]. The latent variable models are taking the responsibility where in the annotations are created by making collaborations with the desired variables that are hidden, which in turn signifies the hidden commonalities over the explanations which are generally [40] smaller in size than the dimension of authentic data. The capability of clarifying the inconsistency of data with the lesser number of components and the ability for providing steady progression on the generation, which in turn creates the models of latent variables selection of real-world for the analysis of the high dimensional dataset.

On the other hand, the integration of previous acquaintance in latent standard variables is mystified. Also, the clustering techniques which are used for disease prediction are having issues like inadequate cluster descriptors, substantial degradation of the effectiveness in high dimensional data, highly sensitive in the initialization phase, outliers, and noise. Moreover, it cannot deal with the non-convex clusters, which have variations in density and size. Hence, to overcome these issues, a novel methodology is proposed. In the proposed work, from the UCI (University of California, Irvine) Repository, the categorical data is gathered by pre-processing the data for the removal of irrelevant or missing data. The extraction

features are performed based on the Eigen value, Eigen vector, and data.

Moreover, from this, the best features are selected with the use of a novel ant colony optimization technique. For choosing the elements, the best features are selected. By utilizing a novel kernel classifier, the process of classification is carried out for the classification of the best features that are chosen. Here the traditional classification techniques [21] such as RF (Random Forests), GBDT (Gradient Boosting Decision Tree), CART (Classification and Regression Trees), NB-M (Multinomial Naive Bayes), NB-DCM (Naive Bayes – Dirichlet Compound Multinomial), GC-LGM (Generative classification model based on Latent Gaussian model) are used for validating the performance of the proposed approach for small datasets. Similarly, the existing clustering techniques such as K-Mean, K-Modes, Manhattan techniques are used for confirming the performance of the same for large datasets. So here our motto is to get more accuracy than the existing methods for all types of datasets. In our method enhanced classification is used for the small dataset and improvised clustering is used for the large dataset as well as the small dataset.

The main intention of this work is as follows:

- To extract the features of the categorical data, the Latent Gaussian Process (LGP) based extraction technique is utilized.
- For selecting the most relevant features from the set of extracted features, a Multi-Objective based Ant Colony Optimization (MO-ACO) technique is employed, which can improve the performance of classification.
- To classify the categorical data with better accuracy, using Canonical based SVM Kernel Function classifier is used.
- To cluster the large scale data with the use of Integration of Manhattan Frequency K-Means with Cluster Center Initialization clustering approach.
- To prove that our classification mechanism is fit for small scale data.
- To prove that our clustering mechanism is suitable for both small scale and large scale data.

The contributions of this work are as follows: Generally, in the medical field, the database may involve

all the type of categorical data which may be small or extensive dataset. This dataset makes the analysis complicated for disease prediction. We aim to improve the machine learning approach to obtain better accuracy results in disease prediction of both the large and small datasets. Among the various machine learning procedures, classification methods play a significant role and achieve better accuracy results. Thus the classification technique is used for small data sets. Even though the classification techniques provide better results but still it lacks in prediction accuracy while using large datasets. For this reason, the clustering approach is also included in this work for obtaining the results with improved accuracy for large datasets.

The remaining portion of this paper is organized as follows: Section 2 provides the literature review of the various state-of-the-art techniques employed for the processing of categorical data. Section 3 objective and motivation towards the work was explained. Section 4 details the description of the proposed mechanism. Section 5 presents the performance analysis of the proposed method. Then, this paper is concluded in the final section.

2. Related Works

This section provides the literature review of the techniques and processes used for the prediction of disease in the medical area i.e., extraction, selection, classification, and clustering mechanism for analysis of diseases from the categorical data.

2.1. Classification

A generative classification approach [22] was suggested for the categorical data depending on latent Gaussian process. A categorical data appears in several applications like gene sequence analysis, natural language processing, and computer-aided diagnosis. For modeling, the categorical data process the latent Gaussian efficient using the Bayesian non-parametric approach. It can estimate the density. A generative classification model for labeling the categorical data, the estimation of class-conditional densities was done by the use of the latent Gaussian process. In the case of categorical data, this method will be a suitable one than the other methods. The major draw-

back of this work was that the time consumption and it takes time for computation as it reduces stability. A coupled attribute similarity learning on categorical data was described in this work [12]. Usually, the attributes were associated with each other utilizing a specific coupling relationship in the real-world data sources. However, the exploration of attribute coupling was done by introducing the co-occurrence of attribute values as they were capable of presenting the local-picture only for the analysis of categorical data. It was useful in capturing the global interactions and intrinsic among the attributes, especially in large scale categorical datasets. However, this approach failed to capture some connection degree attributes, which remains as the major drawback of this method. The semi-supervised learning method was offered in this work to imply & maximize relevance and minimize redundancy using Pearson's correlation coefficient [54]. This approach mainly concentrated on building the highly relevant feature subset. The significance of the features is estimated by the Incremental search model for computing features to coefficients and features to label coefficients, which is very simple to implement, and complexities are reduced. The proposed approach is evaluated with the binary and multi-category benchmark data sets and suitable feature subset is extracted for efficient learning mechanism. Sometimes, this approach increases the noise with labeled data, and it is tough to implement the SVM approach. An ant colony optimization was used for the mixed-variable optimization problems [35]. A new procedure is implemented for generating a mixed-variable benchmark function, artificially for tuning the parameters. This implementation in turn, increases the effectiveness and robustness of the mixed-variable optimization issues.

However, there were some limitations in this work suggested the unsupervised method to embed unlabeled categorical data keen on a continuous and low-dimensional space over and done with the Gaussian process. We can also use the Radial Basis Function with Automatic Relevance Determination as a kernel function of Gaussian processes. Categorical Latent Gaussian Process was implemented to estimate the class-conditional densities for learning hyper-parameters and subsequent likelihoods of latent continuous space. The proposed method is evaluated with splice-junction gene sequences data set and vot-

ing records data set. This approach mainly overcomes the sparsity problems of categorical data. This approach is straight forward for implementation and a very flexible one. There are no particular features in Gaussian processes that are not reproducible in other methodologies. There are no predictive variances and uncertainties which are not taken into account. Angelis et al. proposed the definite sequence of data mining with the use of a hybrid clustering technique [18]. In sequential data, the identification of various dynamics has become an essential factor in the field of life sciences like bioinformatics, marketing, social sciences, and finance. In this, the sequence of categorical data was altered utilizing extended Markov model to probabilistic space. After that, with the use of hierarchical clustering, the courses were clustered. Moreover, this method had some limitations.

A new design of the Hidden Parameter Markov Decision Process (HiP-MDP), a context for modeling relations of interrelated tasks was presented using low-dimensional latent embedding's [31]. Our new framework appropriately simulates the combined improbability in the latent constraints and the state space. Also, the original Gaussian Process-based model was replaced by a Bayesian Neural Network, allowing more accessible interpretation. Regardless of huge batches, every new occurrence still needs disintegrating the insecurity around the instance-specific factors to implement it well and quickly on the task. However, this technique may fail to address the problem of complex control issues that could be overcome. An efficient data-driven similarity learning algorithm was proposed for processing the categorical data, which includes the stages of frequency-based intra coupled similarity and inter coupled similarity [53]. Here, the similarity between the absolute values was estimated based on the relationship between the attributes.

Also, the dissimilarity metrics were defined based on specific requirements. Moreover, the Coupled Attribute Similarity for Objects (CASO) and Coupled Attribute Similarity for Values (CASV) measures were utilized to estimate the frequency distribution and attribute dependency aggregation. The advantage of this work was, it improved the accuracy and reduced the complexity by using inter and intra similarity measures. This work may contain the limitation of manipulating the approaches of attributes education

to adopt the enormously large data. In this work, a new framework was developed, namely, the variable-order Markov framework based on Weighted Conditional Probability Distribution (WCPD) for clustering the categorical sequences [48]. The first-order Markov model was utilized to represent the absolute sequences based on the weighted fuzzy indicator vector representation. Then, a decisive hierarchical algorithm and a two-tier statistical model were employed to cluster the categorical sequence with increased accuracy. The Probabilistic Suffix Tree (PST) was formed to select the memory sequences which were used for clustering and building the WCPD model. Also, a Model-based Categorical Sequence Clustering (MCSC) technique was employed to select one cluster and split into various sub-clusters repeatedly. The benefit of this work was, it provided the first initialization and increased efficiency by implementing the decisive algorithm.

2.2. Clustering

A probabilistic distance function was developed with a kernel density estimation method for clustering the categorical data [14]. In this system, the cluster scatter was defined for estimating the object to cluster distance in the categorical data. Based on the dispersion of categories, the categorical attributes were weighted, which improved the performance of clustering.

In this system, each categorical attribute was automatically assigned based on the correlation between the smoothed dispersion of the categories. Then, the number of certain clusters was estimated by defining the cluster validity index. The efficiency of this mechanism was validated by analyzing both the real world and synthetic datasets. The method of the overall kernel functions and explaining the technique on several kernels should be extended, which remains as a shortcoming of this work. In this work, a k-mode clustering technique was developed with a simple matching dissimilarity measure for processing the categorical objects [4]. The steps involved in this work were as follows: The cluster similarity term was computed by estimating the definite value of each attribute. The objects were partitioned by determining the membership value for all purposes in the cluster. This leads to reduced computational complexity. The cluster centers were updated for finding the modes of objects in

the same group. The attribute weights were computed by analyzing the whole dataset. During the performance evaluation, various datasets such as lung cancer, soybean, dermatology, heart disease data, letter recognition data, heart disease data, and mushroom data were utilized in this work. From the results, it was analyzed that the suggested mechanism offered better efficiency and scalability in clustering the categorical data. A kernel discriminant analysis and clustering with parsimonious Gaussian process modules were proposed [10]. In this method, the data is classified into useful data, categorical data. By combing various kernels, it was possible to sort mixed data. This methodology should be extended to the semi-supervised situation, which was the shortcoming of this work. In this section, various traditional approaches that are used for categorical data are discussed. Their working procedure, along with merits and demerits, are discussed in this paper. Alexandridis et al. proposed a novel learning approach was presented for the categorical data depending upon the Radial Basis Function networks (RBF) [2]. In this approach, the numerical values were referred to as RBF centers that were replaced with the categorical tuple centers. The initial step of RBF training was comprised of comprehensive center selection, which was accomplished by introducing a fast non-iterative absolute clustering algorithm in which the weights were assessed using linear regression. The result illustrated the presented approach offered better generalization. A new distance metric for processing the categorical data was utilized in this work by using an unsupervised learning technique [29]. Also, various distance metrics have been investigated in this work, which included hamming distance, modified value difference metric, Ahmad's distance metric, association based distance metric, and content-based distance metric.

In this system, the distance between the two values was estimated by determining the costs of frequency probabilities. Here, all the similarity measures were individually treated as categorical attributes. Also, the machine learning approach was utilized to obtain a proper distance metric for the given set of objects. Moreover, the weighting scheme was employed to assign the larger weights to the attributes for providing essential information. The significant benefits of this work were as follows: it offered better adjusting capability, highlighted the infrequent

items, and increased frequency. A different resounding subspace learning methods was suggested to untie the latent structures of three protuberant bilinear ways like Probit, Logit, and Tobit [44]. These were deliberated and exhausted entirely.

The determinant probability model that was being normalized through the substitute for the large scale categorical data is used. The Probit model assumes categorical data into quantized values of a positive analog-amplitude vector that resides in a direct low-dimensional subspace. Tobit is the model of high-quality censoring. The probabilistic Logit model simplifies logistic regression to the unsubstantiated instance. The rank regularization method is used for preprocessing the data. The disadvantage of this approach is reduced set representation, which provides only an approximation to the exact solution, and finding the expansions approach inevitably increases the complexity of the algorithm. In this paper, the author offered agnostic learning bounds for analyzing the risk of the Bayesian predictor [43]. It utilized the Regularized Cumulative-Loss Minimization (RCLM) for the posterior calculation against the best single predictor. The limits were implemented for various class of Bayesian models which was comprised of sparse Gaussian process (sGP), Generalized Linear Models (GLM), and Correlated Topic Models (CTM). In the case of CTM, the bound was precisely applied to the variation algorithm with distorted variation linked. For the instance of sGP and GLM, the bounds were implemented to the bounded variants of the log loss. The discrepancy amongst the loss was exposed with an alternative technique by using simple loss minimization. The empirical evaluation of CTM offered the ability to direct loss reduction. The approximation of collapsed variation has the benefits

of better functional performance as well as strong theoretical guarantees. But the major limitation was that the class was restricted to the point estimates in the results. A novel algorithm that utilizes a machine learning approach for the identification of longitudinal patterns on diagnosing disease was presented in the paper [52]. There were two types of technical uniqueness considered: one such form was to enable high learning specificity by a novel learning paradigm and the other one in the way of identifying risk driving diagnosis. Also, a series of investigates which exhibit the efficiency of the offered techniques were provided,

thereby revealing some novel perceptions concerning the most promising future research directions.

The reliability of the software is also finding out by using a machine learning technique [9]. Reliability is one of the major attributes of a software quality assurance system. In this paper, they have implemented the method called a recurrent neural network (RNN). In RNNs methodology output of the current task is dependent on the previous state and the current output is carried out as the input of the next computations. Here they are comparing the results, it shows the proposed RNN method will give more accuracy than the naïve bayes, decision tree and support vector machine.

The prediction of disease by using machine learning algorithm [15] over the big data from health-care communities was carried out. In most of the dataset, there must be some incomplete data entries are found. In this proposed method they use a latent factor model to reconstruct the incomplete data and they proposed a new convolutional neural network-based multimodal disease risk prediction (CNN-MDRP). Then they are comparing their result with CNN-based unimodel disease risk prediction (CNN-UDRP) algorithm. For performance evaluation, they are using F1-measure.

Due to the popularity increase of the social networks, it is difficult in identifying between centrality for large scale network [7]. Recently, few algorithms for identifying the most influential entities in the large-scale network in real-time applications were introduced [26, 32, 34] followed by the MapReduce-based incremental parallel algorithm for exploring the influential nodes based on betweenness centrality in a dynamic network where edges may be dynamically updated.

Small World Model [6], was implemented for a large scale community to find out the uncovering hidden communities in the social network. The main objective of this work is to improve the efficiency of algorithms, parallel programming framework like Map-Reduce for uncovering communities in the network. Here the nodes are mapped into communities, based on the random walk in the network. Small world network exhibits three important characteristics; they are short average path length, high clustering coefficient and exhaustive search using local information. Here they also applied Hadoop framework for solving the complex problem by distributing the computation in multiple nodes in the cluster. Metrics

used here in performance evaluation are Normalized Mutual Information (NMI), Modularity, F-Measure and Execution time.

Automatic knowledge extraction from electronic health records [51] has been reported with an extension of automatic detection technology of some chronic disease types. A novel method was stated in for the prediction of hospital admission kind was introduced as per the patient's medical history representation in the way of binary history vector [3]. The proposed technique was demonstrated with the use of real-time and massive scale dataset that was collected from the local hospitals. This technique provided a better rate of 91% of accuracy on the prediction of first future diagnosis.

We have to concentrate on feature selection part for selecting the best features; this can be helping the data mining technique to improve the percentage of the accuracy. Any data mining algorithms like classification, clustering and so on, we have to select the best feature first. Here we also made a survey on nature-based algorithms like a bee colony, firefly, genetic and ant colony and chosen the best among these algorithms. Artificial Bee Colony (ABC) algorithm [30] is a population-based stochastic optimization realizes the intelligent seeking behavior of honey bee swarms. It can be used for classification, clustering and optimization studies. ABC algorithm has three different groups. They have employed bees, onlooker bees and scout bees [38]. Here the total number of bees employed in the colony will equal the number of onlooker bees. The number of employed bees in the colony is also equated to the number of onlooker bees. The number of employed bees or onlooker bees equals the number of solutions in the population. An onlooker bee will wait in the dance area to make the food source selection decision. An onlooker bee is named employed bee once it goes to a food source. An employed bee that has consumed the food source turns into a scout bee, and its duty is to perform a random search to discover new resources. This is the complete process that happened in the ABC algorithm.

Firefly algorithm is a meta-heuristic algorithm that was inspired by the behavior of flashing lights of real fireflies [19]. The performance of this algorithm is based on the real behavior of fireflies that relies on the attraction between a firefly and another on the basis of their brightness. According to the algorithm, firefly is

a unisex, attractiveness is proportional to brightness and its brightness is determined by the landscape of the fitness function.

Genetic algorithm is one of the nature-based algorithms; this is a stochastic method for function optimization based on the mechanics of natural genetics and biological evolution. The complexity of the problem was reduced by applying a meta-heuristic approach [8]. Here they are identified the related communities in large scale social networks.

Ant colony optimization algorithm is one of the best algorithms for text categorization. This algorithm is inspired to be observation on real ants in their search for the shortest paths to food sources. In Aghdam et al, [37] they have shown the best result by using the ant colony for text categorization and they also comparing their result into the genetic algorithm.

In our proposed work, we are concentrated on each and every stage by giving the novelty in all the places. From the above surveys, the large and small data set has both advantages and disadvantages. The major drawbacks are as follows:

- Generalization capabilities have to be improved.
- Long training time and computational time.
- Reduced stability.
- It increases noise and hard to implement the SVM (Support Vector Machine) approach.

For overcoming these issues or drawbacks, our proposed system is implemented with high resolution.

3. Motivation and Objective

According to the World Health Organization (WHO), 56.9 million deaths occurred worldwide in 2016, among this 54% that is more than 30 million of the deaths happened by the diseases such as Ischaemic heart disease, stroke, chronic obstructive pulmonary disease, lower respiratory infections, Alzheimer disease, and other dementias, Trachea, bronchus, lung cancers, diabetes mellitus, diarrhoeal diseases, and tuberculosis [28]. The ultimate motivation behind the work is to early predict the disease accurately. At present, there are, so many ways to diagnose the disease but the accuracy will not up to the level. To eradicate this problem, the ultimate motivation towards

this work is, find out the new methodology that has to be predicted the diseases with more accuracy rate.

Many attempts were made by the researchers to overcome the issues and challenges faced by the various intermediate phases of data mining techniques such as missing data removal, data normalization, feature extraction, feature selection, and finally disease predictor algorithms. In that pathway, the major objectives of our proposed methodology are to concentrate on the following:

- a The feature extraction process is improved by multi-linear PCA with Categorical Latent Gaussian-based extraction.
- b Feature selection is done by the Multi-Objective ACO algorithm to select the best features.
- c Normal classification is replaced with the canonical based SVM Kernel Function-based classification algorithm.

- d The clustering method was overtaken by the Integration of Manhattan Frequency K-Means with Cluster Centre Initialization.

In this paper, for disease prediction, the categorical datasets alone are considered.

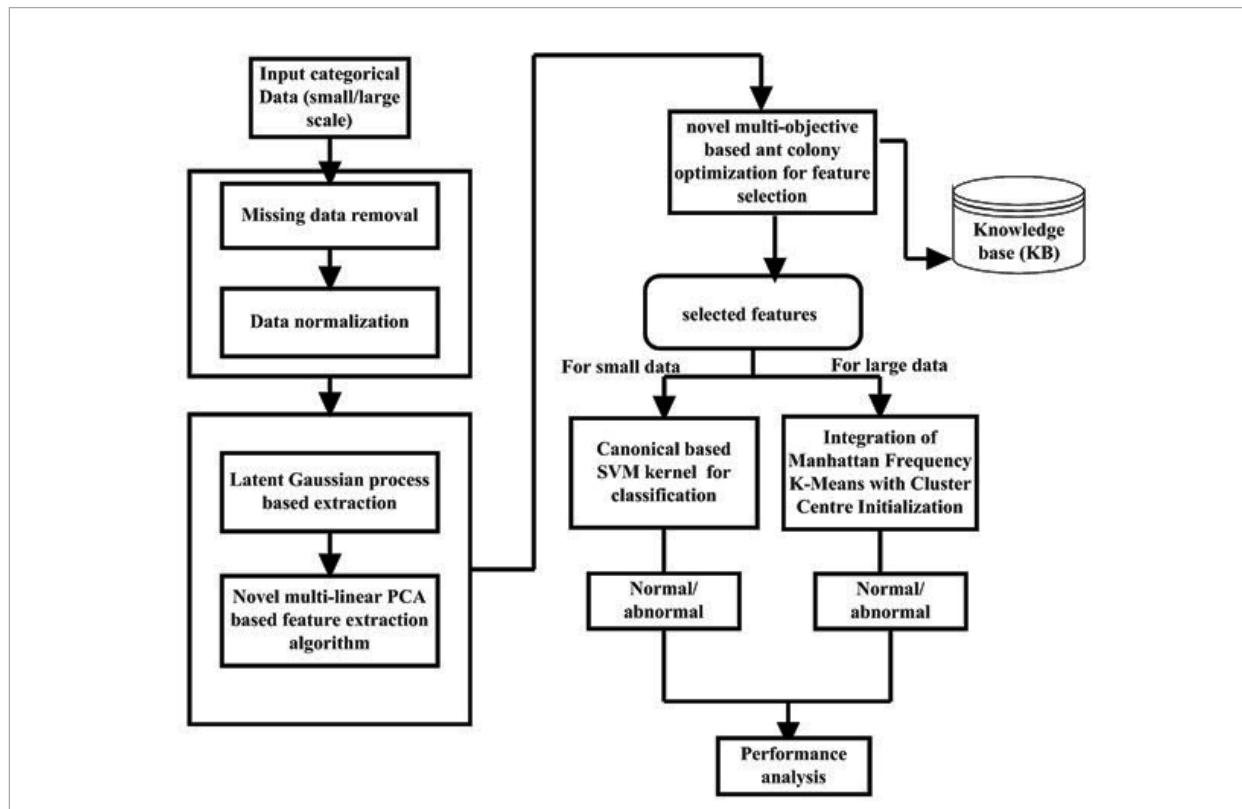
4. Proposed Work

This section provides a detailed description of the proposed mechanism. The overall flow of the proposed system is shown in Figure (1). At first, the categorical data is collected from the repository of UCI. For removing the missing or irrelevant data, the dataset is pre-processed, after which the feature extraction technique is performed, thus to extract the features from the data, Eigen Vectors, Eigenvalues.

Furthermore, a novel multi-linear principle component analysis depending on the feature extraction al-

Figure 1

Overall Flow of the proposed system



gorithm is then executed for the extraction of features from the data. After that, the extracted features are selected to select the best features on implementing a novel multi-objective based ant colony optimization algorithm. A technique of feature selection is utilized for selecting the best feature. From this selected best feature, a Canonical based kernel classifier algorithm is designed for classifying the best-selected features. Likewise, the clustering mechanism is performed with the help of the Integration of Manhattan Frequency with K-Means Cluster Centre Initialization. In the meantime, the trained features are stored in a knowledge database that could be made use in the testing phase. Here we can find the solution to predict the disease in both small and large scale. The datasets of more than 1000kB size are considered as large datasets, and the ones with size less than 1000kB are small datasets. For this, we can implement two different types of the algorithm, i.e., Canonical based kernel algorithm and Manhattan frequency with K-means cluster center initialization.

4.1. Pre-processing

The input is the categorical data that is the large and small scale data. The presence of noise or irrelevant data will harm the outcome. Hence, it is necessary to reduce or remove the noise, i.e., missing or unrelated data present in the input dataset. Initially, the dataset is pre-processed to minimize or eliminate the missing or irrelevant data [23]. The data normalization is also carried out in the pre-processing stage. The data pre-processing is carried out both in the training and testing phase.

4.2. Feature Extraction

In general, PCA [47] extracts the features based on linearly uncorrelated variables, not suitable for non-linear space data, and not performed by analyzing the relationship among data points. But in the Latent Gaussian Process-based multi-linear PCA (Principal Component Analysis) is a latent variable model in which the maximum likelihood solution for the hyper parameters is found through solving a non-linear kernel-based eigenvalue problem on the data's covariance matrix.

In the proposed algorithm, initially, input data is considered for pre-processing where missing values and null values are replaced. Then those pre-processed data is divided into dependent and

independent variables where categorical data is converted into numerical data.

Initialize the parameter value α and β for estimating the features matrix, here α is setting as 13, which used for consideration of minimal dimension of attributes for extracting correlated features. Next, compute the mean, standard deviation, softmax function, and covariance matrix for the processed data. The log-likelihood parameter is calculated using the estimated softmax function and a covariance matrix. Categorical Gaussian features are updated based on the kernel function, which is calculated by determining the maximum log-likelihood parameter and hyper parameter. Followed by, central components are computed by taking the covariance for the subtraction of average categorical Gaussian features. Eigenvalues and Eigenvectors are extracted from the primary elements. Finally, we obtain the principal component features by adding the absolute Gaussian features with Eigenvectors of the central components. Once the dataset is pre-processed, the features are extracted utilizing a latent Gaussian based feature extraction technique and a multi-linear PCA based feature extraction mechanism.

Algorithm 1: Multi-Linear PCA with Categorical Latent Gaussian Process-based Feature Extraction

Input: Dataset Dt

Output: $Feat_{c_{lgp}}$

Procedure:

Step 1: Let Dt be the input dataset

Step 2: By applying preprocessing techniques carry out missing values replacement and null value replacement

Step 3: Let the preprocessed dataset be Dt_p

Step 4: Now divide the Dt_p into dependent (Dt_y) and independent (Dt_x) variables

Step 5: After the separation of dependent and independent variables, the categorical data to numerical data conversion is done.

Step 6: Initializing $\alpha=13$ and $\beta=15$

Step 7: for $I=1$ to size (Dt_x)

Step 8: Compute the standard deviation (σ) of

$$\sigma = \sqrt{\frac{Dt_x(i) - Dt_x(j)}{\text{size}(Dt_x)}} * \sqrt{\frac{Dt_x(i) - Dt_x(j)}{\text{size}(Dt_x)}}$$

Step 9: Compute the mean (μ) of Dt_x , $\mu = \frac{\sum Dt_x}{size(Dt_x)}$

Step 10: Generate an identity matrix of σ , $I_{max} = \begin{bmatrix} \sigma & 0 \\ 0 & \sigma \end{bmatrix}$

Step 11: compute the softmax function by using the following equation

$$\text{Softmax}(Sf_x) = e^{\sigma_i} / \sum_{i=1}^{size(\sigma)} e^{\sigma_i}$$

Step 12: Generate the covariance matrix Cv_x

Step 13: Then the log-likelihood parameter is computed by using the below equations. Let x and y represent the size of Cv_x . Then let z be the size of the Cf_x vector

The generative model is represented by,

$$gm_y = \sum_{i=1}^x \sum_{j=1}^y \sum_{k=1}^z (Cv_{i,j} * Sf_x^k)$$

Step 14: Then from the above computations the hyperparameters are computed by using the below equations, $\theta_m = \sum_{x=1}^{size(\sigma)} (mean(Cv_x) * \sigma_m) + gm_y$,

Step 15: then the hyperparameter $K_x = \sum e^{-\frac{1}{2} * \theta_m}$

Step 16: Now compute the features using the generated parameters using

$$X_m = \sum Sf_x + 1 - \left\{ \left(\frac{\alpha}{\alpha + \beta} \right) - K_x * (\mu + I_{max}) \right\}, X_n = mean(Cv_x)$$

Step 17: compute the categorical Gaussian features by the equation $Cg_f = e^{-\frac{1}{2} * \sigma_m * (X_m - X_n) * (X_m - X_n)}$

Step 18: calculate the central components of the categorical features using,

$$\text{Cent} = \text{Covariance matrix (avg}(Cg_f) - \mu)$$

Step 19: Now extract the Eigenvalues and Eigenvectors of the central components to generate the multi-linear features

Step 20: The extracted features are represented by,

$$Feat_{c_{lg}_p} = Cg_f + (vec(Cent))^T$$

This technique is an effective method to obtain the best features present in the categorical input data. Gaussian Process (GP) benefits are as follows: it can directly capture the model improbability, once consuming GP, it is capable of improving preceding information and terms about the outline of the model by choosing altered kernel functions.

PCA approach has the following benefits like reduced complexity in images, combined with the usage of

PCA. Smaller database depiction as only the trainee images are kept in the system of their estimates on a condensed basis.

The decline of noise as the maximum variation source is selected, and thus the small differences in the background are disregarded automatically.

The list of features is extracted to check and filter out the best features in the next stage with the use of the feature selection mechanism.

An algorithm for multi-linear PCA with the categorical Latent Gaussian Process-based feature Extraction is shown. The input is the dataset. By applying the pre-processing technique, the missing or irrelevant data are removed and then divided as dependent and independent variables. Followed by this, categorical data are converted into numerical data. The standard deviations are estimated by initializing the alpha and beta values. After that mean, identity matrix, softmax function, covariance is generated. From the hyperparameters, the generation of features is made. From this produced feature, the computation of categorical Gaussian features is performed for which the eigenvalues, eigenvectors of the central components are generated to extract multi-linear features.

4.3. Feature Selection

Ant colony optimization algorithm (ACO) [36] generally used for the selection of optimal features based on the objective function; here, ACO is applied for solving the multi-objective problem.

Initialize the parameters of the ACO by setting the weight, minimum, and maximum velocity. Initialize the ant population as the extracted features, from which estimate the pheromone and speed of the communities. Update the local pheromone and velocity by comparing the features matrix. Estimate the fitness value based on the features matrix and computed pheromones. The above fitness value is used for predicting the relevant features. With the estimated fitness value is used for updating the global pheromone index by deriving the global objective function. Obtained global pheromone is processed for suppressing the redundant features. Finally, we generated the best feature matrix. From the extracted features, the best optimal features are selected with the use of a novel multi-objective based Ant-colony optimization (ACO) mechanism.

Algorithm 2: Multi-Objective ACO Algorithm for Best Feature Selection

Input: Extracted Features $Feat_{clgp}$

Output: Best selected Features ($Feat_{best}$)

Procedure:

Step 1: Initially set the weight, minimum, and maximum velocity

Step 2: As a next step, the input features are initialized into total features into a set of ant populations and also initialize the pheromone and velocity.

Step 3: Start the iteration by selecting an ant/feature, for each iteration, calculate the pheromone and velocity for cost computation by using the below equations

$$P_{ij} = X_{\min} + (X_{\max} - X_{\min}) * Feat_{ij}$$

and the velocity is given by $V_{ij} = V_{\min} + (V_{\max} - V_{\min}) * Feat_{ij}$

Where i and j represent the size of the features/ the number of the ants.

X_{\min}, X_{\max} represents the minimum and maximum initial values of the pheromone, and V_{\min}, V_{\max} are the minimum and maximum values of the velocity that is initially set.

Step 4: Update the values of the local pheromone and velocity values by using the below condition as,

$$P_{ij} = \begin{cases} P_{ij} \text{ if } (Feat_{ij} > \frac{1}{2}) \\ -P_{ij} \text{ otherwise} \end{cases} \& V_{ij} = \begin{cases} V_{ij} \text{ if } (Feat_{ij} > \frac{1}{2}) \\ -V_{ij} \text{ otherwise} \end{cases}$$

Step 5: Now the fitness value computation is done by using the objective function as,

$$G_{best}(P_{ij}) = \sqrt{\sum_{x=1}^l (P_{ij} - Y_j)^2}$$

Step 6: From the best fitness values, best path value is selected by using, the equation below,

$$Path_{sel} = \{P_{i-1,j} \text{ if } P_{i-1,j} > \max(G_{best}), \text{ for all } i \& j\}$$

Step 7: Now update global pheromone and the path using the below functions, Setting $\omega = P_0$

Step 8: Update the global best pheromone by using

$$P_{ij} = \sqrt{\omega_z * V_{ij} * (fit_j - P_{ij}) * (G_{best}(P_z) - P_{jz})}$$

and also update fitness values.

Step 9: Compute the best solution by choosing the best path as, $Feat_{best} = (G_{best} < \text{mean}(G_{best}))$

Step 10: Get the index of the best-selected features and generate the best feature matrix.

This ACO approach has the following advantages as it can search among the population in parallel, gave rapid discovery of the proper solutions, can quickly adapt to the changes like new distances, and has a guaranteed coverage. This technique is effective for a selection of

features where the selected best features are stored in a knowledge database (KB). In the testing phase, the best test features are selected, depending on the best indexes in KB. Also, the best selected features are kept in the machine that is trained. After the selection of the best features, the clustering mechanism is performed for the large scale data, and the classification technique is carried out for smaller-scale data.

A multi-objective ACO algorithm for the selection of the best feature is shown in which the input is an extracted feature. At first, the weight, maximum-minimum velocities are set, after which the input features are initialized into total features as a set of the ant population, thereby initializing the pheromone and velocity. Iterations start with the selection of elements or ant by the cost computation of velocity and pheromone. The values of local pheromone and velocity are updated. The fitness function values are computed from the objective function from which the path value is selected. The updating of global pheromone and the path is updated on computing the best solution for choosing the way that is best. The indexes of the best-selected features are attained by the generation of the best feature matrix. Finally, the best-selected features are obtained as an output.

4.4. Classification Mechanism

A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyperplane [25]. SVM usually performed with kernel function, which is used to solve any complex problem and provides strength during the training process, which improves the accuracy of the classifier. The selection of kernel function for hyper parameters is an essential task. In general, the Radial Basis Function (RBF) is used as a kernel function that uses the same weight parameters for every attribute and provided less accuracy. Here canonical based Kernel Function is used for updating the distributed weights.

Initially, training sets and testing sets are separated from the optimized features. Designed the custom kernel function for updating the hyper parameters using Jordon Canonical based Kernel Function also initialized the latent dimension and mesh matrix. The settings of variational and hyper parameters for all iteration are updated. Finally, the predicted results are being generated and are compared with the ground truth. Here we can apply the small dataset, and then the results are compared with the existing methods.

In the case of small scale data, the classifications of the best-selected features are classified by utilizing a Canonical based SVM Kernel function - based classification algorithm. SVM algorithms use a set of mathematical functions that are defined as the kernel. SVM has the following advantages: functions well with even formless and semi-structured documents like text, Images, and trees, the kernel pretend are the actual strong point of SVM. By the suitable kernel function, any problematic could be solved, unlike in neural networks, SVM is not resolved for local optima. This measures the high dimensional data quietly well, SVM models have an overview in practice, and the risk of over fitting is a smaller amount in SVM. The function of the kernel is to take data as input and transform it into the required form. Different SVM algorithms use different types of kernel functions such as linear, non-linear, polynomial, radial basis function (RBF), and sigmoid. The kernel functions return the inner product between two points in a suitable feature space.

Algorithm 3: Canonical based SVM Kernel Function-based Classification Algorithm

Input: Best selected features ($Feat_{best}$), classifier Labels (C_{lbls})

Output: classified output (C_{Ops})

Procedure:

Step 1: Let testing size =30% and training size be =70%

Step 2: From the selected features, separate the features into a training set ($T\gamma_{feat}$) and testing set (Ts_{feat})

Step 3: Instead of radial basis function the custom kernel classifier is designed in such a way that hyper-parameters for the classifier are pre-set in the feature extraction module, and the latent dimension are set, and also the size of the mesh matrix is initialized to 0.02

Step 4: The custom kernel is designed based on canonical form to utilize the function, $K = T\gamma_{feat} * (Mesh_{\beta} * C_{lbls}^T)$

Where, $Mesh_{\beta} = \frac{T\gamma_{feat}^2 - 7}{(T\gamma_{feat} - 2)(T\gamma_{feat} - 3)}$ // Jordan canonical expression

Step 5: Update the variational parameters and hyper-parameters for each iteration

Step 6: The predicted results (C_{ops}) are generated and compared with the ground truth (C_{lbls})

Also, the clustered data are classified, which are then trained in a trained machine to yield a classification output finally. Thus, the extracted features are classi-

fied effectively, which is then used to predict the disease in a medical area.

An Enhanced Kernel classifier Algorithm is presented in which input is the best-selected features and classifier labels. The testing size of 30 percent and training size of 70 percent is considered from which the selected elements are separated as training and testing set. Radial basis function enabling the design of a custom kernel classifier was made in a manner where the hyper parameters of the classifier are pre-set in the module of feature extraction. The size of the matrix mesh and the latent dimensions are initialized as 0.02. The parameters of variational and hyperparameters of every iteration are updated.

Finally, the predicted results are being generated and are compared with the ground truth. Here we can apply the small dataset, and then the results are compared with the existing methods. After that, we know that it was well suited for the small dataset and not for a large dataset, so we introduced the enhanced Manhattan frequency with K-means cluster center initialization for focusing the large dataset. This method provides the best results for the small dataset, and it was compared with the existing processes, but it's unfit for a broad set.

4.5. Clustering Technique

In general, cluster centroids are considered as randomly based on the ranges of given sample values. It leads to clustering the data points randomly, and it may provide misclassification and more deviation in similarity among classes.

Here, more efficient and scalable operations are being proposed towards the categorical clustering of the classified outcomes from the classifier. Besides, we incorporate the cluster center initialization in our method. For the large scale data, Integration of Manhattan Frequency K-Means [42] with Cluster Centre Initialization is carried out to cluster the vast amount of data and to classify the clusters concerning the best features. K-means approach has the following advantages: If variables are massive, then K-Means performs computationally quicker than hierarchical clustering if k is kept smalls, K-Means create tighter clusters than hierarchical clustering, mainly if the clusters are round.

The Manhattan Frequency k-Means (MFk-M) is a partitioned categorical clustering method based on

transforming the categorical data into numeric measures using the relative frequencies of the modalities in the attributes. This fact permits directly using traditional numeric clustering methods.

Algorithm 4: Integration of Manhattan Frequency K-Means with Cluster Centre Initialization

Input: Best selected features ($Feat_{best}$), Labels (C_{lbls})

Output: Clustered results (C_{Ops})

Procedure:

Step 1: Initially the number of clusters are set as K

Step 2: Then next step refers to the selection of initial cluster centers, it's done by the below steps,

Step 2a: Let D_{sel} be the set of data elements with the best-selected attributes, where n represents the number of best-selected attributes

Step 2b: For each best selected characteristics, compute the Mean (μ_A) and standard deviation of the best characteristics as (σ_A) in n and calculate the centile values for each attribute

Step 3: Now create the initial partition between the data elements and the attribute values $best_{sel}$

Step 4: Apply clustering in $best_{sel}$ characteristics and update the cluster labels and the initial cluster centers.

Step 5: Update distance by using hamming distance, $dist_{ham} = (x_{val} - y_{val})$

Step 6: Update the points of the cluster based on the data points

Step 7: For I in K, C_1 = cluster (data point in I)

Step 8: Compute mean scores in the group and the update the cluster centers

Step 9: Based on the cluster member, the cluster data points are extracted, and the clustered results are obtained.

At first, the numbers of clusters are set as K; the selection of initial cluster centers is referred to in the next step. In the set of data elements, the selected best attributes are there and n is the number of characteristics. The mean, standard deviation of the best features are computed for the best-selected attributes along with the evaluation of centile values for each attribute. The initial partition among the data elements and attribute values are then created. On applying the clustering technique, the cluster labels are updated with the initial cluster centers. Using a Hamming distance, the distance is updated. After the updating of cluster points,

the mean scores in the cluster are computed. The cluster data points are extracted depending on the cluster member by extracting the cluster output finally.

5. Performance Analysis

This section provides the performance analysis of the proposed mechanism on the medical dataset with categorical attributes. For both small and large datasets, 30% testing and 70% training are taken.

5.1. Dataset Description

5.1.1. SPECT Heart

The dataset defines and analyzes the cardiac Single Proton Emission Computed Tomography (SPECT) [49] images. Each of these patients is categorized into two groups: abnormal and normal. This dataset has two classes, wherein class 1 contains 15 samples, and class 2 contains 172 samples. The database of 267 SPECT image sets (patients) was used to extract features that condense the SPECT original images. Consequently, 44 continuous feature patterns were generated for the respective patient. The pattern was processed further to attain 22 binary feature patterns, and the size of this dataset is 9 kB.

5.1.2. Breast Cancer

This is one of the three fields delivered by the Institute of Oncology [11] that has frequently seemed in the literature of machine learning. (Lymph graphic and primary-tumor). This data set comprises 201 cases of one class and 85 illustrations of a different class. The instances are labeled by 9 attributes, nearly of which are linear, and some are insignificant. The size of the breast cancer dataset is 19 kB.

5.1.3. Gene Sequences

Samples (instances) are kept row-wise. Variables (attributes) of each model are RNA-Seq gene expression [24] stages restrained by the platform illumine HiSeq. Splice junctions, which are also referred to as Gene Sequences, are DNA sequence points at which 'superfluous' DNA is detached in the practice of protein formation in developed entities. The difficult modeled in this dataset is to identify a given categorization of DNA, the boundaries among introns (the parts of the DNA sequence that are spliced out), and exons (the parts of the DNA sequence retained after splicing).

The total size of this dataset is 385 kB. There are three different classes: 766 samples in class 1, 768 samples in class 2, and 1655 samples in class 3. It is comprised of 3190 data records and 60 feature patterns.

5.1.4. EEG Eye State

All data is from one constant EEG [20] extent with the Emotive EEG Neuro headset. The length of the extent was 117 seconds. The eye state was sensed through a camera in the course of the EEG extent and farther ahead automatically to the folder after examining the video frames. '1' designates the eye-closed and '0' the eye-open state. This dataset has 8257 samples in class 1 and 6723 samples in class 2. Its size is 1657 KB (1.6 MB). All standards are in consecutive order with the first restrained rate at the uppermost of the information. The EEG Eye state is comprised of 14980 data records and 14 feature patterns.

5.1.5. Epileptic Seizure Dataset

The original dataset [5] from the location contains five altered folders, all with 100 files, by every folder on behalf of a distinct issue/person. Every file is a record of brain action for 23.6 seconds. There are five different classes, each having 2300 samples. The consistent time-series is tested into 4097 data points. Every data point is the rate of the EEG record at an altered point in time. Also, there are 178 features present in this dataset, and the size is 7329 KB (7.3 MB). Therefore we must total 500 individuals by each has 4097 data points for 23.5 seconds.

5.2. Performance Metrics

5.2.1. Precision

Precision is well-defined as a quantity that is used to estimate the concert of the classification method.

$$Precision = TP / (TP + FP). \quad (1)$$

5.2.2. Recall

Recall processes ability of the estimated model to choose the illustration of an assured session as of a dataset. It is also named as sensitivity that is deliberated as

$$Recall = TP / (TP + FN). \quad (2)$$

5.2.3. F1-score

In the numerical analysis of binary segmentation, the F1 score is used to measure and test accuracy. The F1

score is constructed by the weighted average of precision and recall. The F1 measure reaches its best value at one or near one. The worst value is represented as zero or near to zero. F1 score is used to measure the test accuracy, and it includes both the precision and recall values. F1 score reaches the best value as one, and the worst value is zero.

$$F_1 = 2 \cdot \frac{1}{\frac{1}{recall} + \frac{1}{precision}} = 2 \cdot \frac{precision \cdot recall}{precision + recall}. \quad (3)$$

5.2.4. Accuracy

Accuracy denotes the closeness of a restrained value to a standard or known value. Accuracy is also stated to weighted arithmetic mean of Precision, and Inverse Precision besides weighted arithmetic mean of Recall and Inverse Recall

$$Acc = \frac{TP + TN}{P + N} \text{ or } \frac{TP + TN}{TP + TN + FP + FN}. \quad (4)$$

5.3. Performance and Comparative Analysis of Small Dataset (Classification)

The performance analysis and comparative analysis of the small dataset classification [13] process is shown in the following subsections.

5.3.1. Breast Cancer Dataset

The Table below shows the breast cancer dataset accuracy rate percentage for proposed and existing methods.

Table 1

Performance analysis of breast cancer dataset in terms of accuracy

Test accuracy rate (%)					
Dataset	Models	Mean	Best	Worst	SD
Breast cancer	RF	58.68	60.79	56.83	1.27
	GBDT	62.09	66.08	60.35	2.49
	CART	56.18	60.79	51.1	2.58
	NB-M	64.76	NA	NA	NA
	NB-DCM	55.95	NA	NA	NA
	GC-LGM	58.77	64.76	54.19	3.2
	Fengmao et al	71.45	77.97	60.79	4.81
	Proposed	80.86	82.2	78.32	0.01

In Table 1, the breast cancer dataset is taken, and the classification process is carried out for this small dataset. Various techniques are compared to the rate of accuracy. On comparing other traditional methods like RF (Random Forests), GBDT (Gradient Boosting Decision Tree), CART (Classification and Regression Trees), NB-M (Multinomial Naive Bayes), NB-DCM (Naive Bayes – Dirichlet Compound Multinomial), GC-LGM (Generative classification model based on Latent Gaussian model) in terms of mean, best, worst, and SD. From this analysis, the proposed method shows a better test accuracy rate of about 80.86 as mean, 82.2 for best, 80 for worst, and 0.01 SD.

Table 2

Performance analysis of breast cancer in terms of precision and recall

Models	Recall	Precision
RF	62.14	32.42
GBDT	53.37	33.47
CART	61.25	30.03
NB-M	55.36	36.05
NB-DCM	53.37	28.85
GC-LGM	62.68	32.65
Fengmao et al.	51.79	43.94
Proposed	72.16	75.06

The precision, recall values for the breast cancer dataset are shown in the Table 1 for both proposed and existing methods. From the table provided, the precision and recall values are high for the proposed method of comparing it with the existing methods. The values of the proposed method are shown in the last row.

5.3.2. Spect Heart

The Table 3 below shows the SPECT heart dataset accuracy rate percentage for proposed and existing methods.

In Table 3, the SPECT heart dataset is taken, and the classification process is carried out for this small dataset. Various techniques are compared to the rate of accuracy. On comparing other traditional techniques, the proposed method shows a better test accuracy rate

Table 3

Performance analysis of SPECT heart dataset in terms of accuracy

Dataset	Models	Test accuracy rate (%)			
		Mean	Best	Worst	SD
SPECT Heart	RF	82.85	81.82	80.75	0.39
	GBDT	81.82	NA	NA	NA
	CART	76.76	80.21	72.73	2.42
	NB-M	77.01	NA	NA	NA
	NB-DCM	81.28	NA	NA	NA
	GC-LGM	89.13	90.91	87.17	1.27
	Fengmao et al.	88.47	89.84	85.03	1.35
	Proposed	91.72	93.7	89.85	1.44

of about 91.72 for mean, 93.7 for best, 89.85 for worst, and 1.44 SD. The precision, recall values for the SPECT heart dataset are shown in the below table for both proposed and existing methods. Here the proposed value is taken from the mean of the proposed method.

In this, the mean value will be taken from ten times occurrence. These are the small datasets which are having less number of records in it. For this, we can go with our classification method itself.

Based on the results provided in Table 4, the precision and recall values are high for the proposed method while comparing with the existing ones.

Table 4

Performance analysis of SPECT heart in terms of precision and recall

Models	Recall	Precision
RF	86.67	30.19
GBDT	86.67	28.89
CART	86.67	24.04
NB-M	66.67	20.83
NB-DCM	66.67	25.64
GC-LGM	77.33	41.35
Fengmao et al.	80	38.95
Proposed	89.2	45.8

5.3.3. Gene Sequences

In Table 5, the gene sequence dataset was used for disease prediction and reported with accuracy rate for proposed and other existing methods.

Table 5

Performance analysis of gene sequence in terms of accuracy

Models	Test accuracy rate (%)			
	Mean	Best	Worst	SD
RF	93.16	93.71	92.6	0.29
GBDT	90.94	90.95	90.88	0.02
CART	90.46	NA	NA	NA
NB-M	92.74	NA	NA	NA
NB-DCM	90.39	NA	NA	NA
GC-LGM	94.12	95.92	91.71	1.47
Fengmao et al.	92.92	94.13	90.67	0.99
Proposed	93.404	95.8	91.31	1.805

In Table 5, the Gene sequence dataset is taken, and the classification process is carried out for this small dataset. Various techniques are compared to the rate of accuracy. On comparing other traditional techniques, the proposed method shows a better test accuracy rate of about 93.404 for mean, 95.8 for best, 91.31 for worst, and 1.805 SD. The precision, recall values for the Gene sequence dataset are shown in the below

Table 6

Performance analysis of gene sequence in terms of precision and recall

Models	Recall	Precision
RF	94.62	91.99
GBDT	91.94	90.09
CART	90.46	91
NB-M	91.55	93.76
NB-DCM	89.47	91.11
GC-LGM	95.96	93.34
Fengmao et al.	90.03	95.87
Proposed	94.86	98.88

table for both proposed and existing methods. It is inferred from the above table that the precision and recall values are high existing methods like RF, GBDT, CART, NB-M, NB-DCM, GC-LGM, Fengmao et al.

From this analysis, the proposed method shows a better result than the traditional techniques. In addition to the performance metrics such as percentage of accuracy, precision and recall, to validate the effectiveness of our proposed method, for the Breast Cancer, SPECT and Gene Sequence datasets, the F1 score values were also computed as detailed in Table 7.

Table 7

F1 score for small datasets

DATASET	F1_SCORE
Breast Cancer Dataset	60.52385
SPECT Heart Dataset	73.58144
Gene Sequence Dataset	96.82829

5.3.4. Performance and Comparative Analysis of Large Dataset (Clustering)

The performance analysis and comparative analysis of the large dataset clustering [41] process is shown in the above table shows. As reported in Table 8, the clustering technique of a large scale dataset provides a better result for the proposed method than the existing methods. The epileptic seizure [16] and EEG eye state [45] datasets are considered here as large datasets which provides an accuracy rate of about 0.828 and 0.9506. Similarly, the NMI rate of epileptic seizure and EEG eye state are 0.9739 and 0.9777. Likewise, the execution time is 0.015 and 0.235 for epileptic seizures and EEG eye state datasets.

Table 8

Performance analysis of large dataset in terms of accuracy, NMI, and execution time

Category	Dataset	Proposed clustering Algorithm		
		Accuracy	NMI	Execution time in sec
Big dataset	Epileptic Seizure	0.828	0.9739	0.015
	EEG Eye state	0.9506	0.9777	0.235

Table 9 presents the performance analysis of the large scale dataset for the epileptic seizure prediction. The conventional estimates in terms of rand score, V-measure score, silhouette score, and CH score, were compared with our proposed clustering technique for analyzing their better performance.

Table 9

Performance analysis of large dataset for Epileptic Seizure

For Comparing Epileptic Seizure Dataset				
Methods	Rand Score	V-measure Score	Silhouette Score	CH score
Fesim	0.45	0.6	0.16	230
Select K Best	0.3	0.5	0.1	150
Tree Classifier	0.3	0.45	0.13	140
RFE	0.2	0.35	0.05	100
Proposed	0.817	0.845	0.143	260

Based on the comparative analysis, it is found that our proposed approach is capable of providing better results for all the estimates and outperforms the existing methods.

In Table 10, the proposed clustering technique for handling a large scale dataset has been reported with better results than other similar techniques. The epileptic seizure and EEG eye state are taken, which provides an accuracy rate of about 0.828 and 0.9506. Similarly, the NMI rate of epileptic seizure and EEG eye state are 0.9739 and 0.9777. Likewise, the execution time is 0.015 and 0.235 for epileptic seizure and EEG eye state.

Table 10

Performance analysis of large dataset or EEG Eye state

Measures	K-Means	Sc	US-ELM K-means	Us-EF-ELM k-means	Proposed
Max Accuracy	55.11	55.11	59.8	58.68	95.06
Mean Accuracy	55.11	55.11	54.14 +/-2.20	58.68 +/-0.01	95.06 +/-1.5

Also, the proposed clustering technique is tested for the small dataset, as per Table 11; it shows that our proposed method yields better outcomes for small

datasets also. Despite the fact, the classification techniques provide better results; still, it lacks in prediction accuracy while using large datasets.

Table 11

Performance analysis of small dataset in terms of accuracy, NMI, and execution time

Category	Dataset	Proposed clustering Algorithm	
		Accuracy	NMI
Small Dataset	SPECT heart	0.979	0.989
	breast-cancer	0.851	0.876
	Gene Sequence	0.972	0.9738

Finally, it is observed that our proposed clustering method is well suited for both small and extensive dataset for the correct prediction of diseases in the medical field.

The sensitivity and specificity values of small datasets are listed in Table 12. Here sensitivity is the percentage of sick people correctly identified as having the condition. Likewise specificity is the percentages of healthy people are identified as not having the condition.

Table 12

Sensitivity and Specificity of small datasets

Dataset	Sensitivity	Specificity
Breast Cancer Dataset	82.222	86.743
SPECT Heart Dataset	93.721	94.532
Gene Sequence Dataset	95.843	95.674

Here we also extended the analysis part by made the cross-validation of all the given dataset by varying the percentage of training and testing data in the difference of 10%, from 10% to 90%. The result of this analysis is mentioned in Table 13 and it is displaying the best value for all the datasets.

Based on the comparative analysis of the proposed method with existing state-of-art approaches, it has been proven that our proposed method obtained more classification accuracy than other methods.

Table 13

Cross-validation of accuracy rate in small datasets

Data set	Test Size	SPECT Heart Dataset	Breast Cancer	Gene Sequence
Accuracy rate	10%	0.91304	0.75757	0.92231
	20%	0.92857	0.80645	0.93905
	30%	0.93721	0.82222	0.95843
	40%	0.96202	0.87394	0.96531
	50%	0.93877	0.89115	0.96930
	60%	0.94017	0.90909	0.98691
	70%	0.94814	0.87317	0.98973
	80%	0.96753	0.87982	0.99162
	90%	0.97109	0.88167	0.99578

6. Conclusion

In the present time, the clustering of categorical data was a challenging aspect in the medical field, which in turn handles the massive number of data. The prediction of disease in an early stage is of most importance for the quick treatment and curing them. In a large scale categorical data, the clustering technique was performed effectively in this proposed mechanism. Initially, small and large scale categorical data are given as input, which was pre-processed; feature extracted by using some novel multi-linear PCA based feature extraction technique. Then the selection of features is made by using a novel multi-objective based ant colony optimization approach. In the case of small scale data, the selected features are classified using canonical based SVM kernel classifier. Also, in large scale

data, the clustering was performed with the use of the Integration of Manhattan Frequency K-Means with Cluster Center Initialization clustering technique, after which the classification mechanism was carried. In our proposed method the classification technique gives more accuracy for small datasets.

For breast cancer dataset the previous existing highest accuracy is 77.97%, in our method accuracy is 82.2%. For SPECT heart accuracy of the existing method is 90.91%, in our method 93.7% is obtained. For gene sequence dataset the previous accuracy from the existing method was 95.92%, in our method 95.8% is obtained. In the proposed method, clustering is more suitable for large scale datasets and small datasets. For large datasets we are tested with two datasets EEG eye state and Epidemic seizure. For EEG eye state we got 95.06% accuracy and for Epidemic seizure, we got 82.8% accuracy.

From the study, it was evident that our classification method is suited for small scale datasets, and the clusters method was suitable for large scale datasets and also small scale dataset. From this, it was evident that the proposed mechanism would provide the results more accurate to predict the disease easily.

Then the result of proposed method is compared by using the performance metrics like precision, NMI, execution time, recall, and accuracy, and it is proved to be an out performing one.

Even this new method requires further improvement in implementing a unique iterative relocation based partitioning for clustering the big dataset in the future. The research will also extended by testing the numerical datasets. Also we have to do the analysis by changing any other nature-based algorithms instead of ACO algorithm. Here the main intension of the future work is to enhance this methodology to make adoptable for all kind of datasets.

References

1. Agresti, A. An Introduction to Categorical Data Analysis. Wiley-Interscience, 2018.
2. Alex, A., Eva, C., Nikolaos, G., Haralambos, S. A Fast and Efficient Method for Training Categorical Radial Basis Function Networks. IEEE Transactions on Neural Networks and Learning Systems, 2017, 28, 2831-2836. <https://doi.org/10.1109/TNNLS.2016.2598722>
3. Arandjelović, O. Discovering Hospital Admission Patterns Using Models Learnt from Electronic Hospital Records. Bioinformatics, 2015, 31, 3970-3976. <https://doi.org/10.1093/bioinformatics/btv508>
4. Bai, L., Jiye, L. The K-modes Type Clustering Plus Between-Cluster Information for Categorical Data. Neurocomputing, 2014, 133, 111-121. <https://doi.org/10.1016/j.neucom.2013.11.024>

5. Baumgartner, C. F., Konstantinos, K., Jacqueline, M., Tara, P. F., Sandra, S., Koch, M. L., Bernhard, K., Daniel R. SonoNet: Real-time Detection and Localisation of Fetal Standard Scan Planes in Freehand Ultrasound. *IEEE Transactions on Medical Imaging*, 2017, 36, 2204-2215. <https://doi.org/10.1109/TMI.2017.2712367>
6. Behera, R. K., Rath, S. K., Misra, S., Damasevicius, R., Maskeliunas, R. Large Scale Community Detection Using a Small-World Model. *Applied Sciences*, 2017, 7(11), 1173. <https://doi.org/10.3390/app7111173>
7. Behera, R. K., Naik, D., Ramesh, D., Rath, S. K. MR-IBC: Map Reduce-Based Incremental Betweenness Centrality in Large-Scale Complex Networks. *Social Network Analysis and Mining*, 2020, 10, 1-13. <https://doi.org/10.1007/s13278-020-00636-9>
8. Behera, R. K., Naik, D., Rath, S. K., Dharavath, R. Genetic Algorithm-Based Community Detection in Large-Scale Social Networks. *Neural Computing and Applications*, 2019, 1-17. <https://doi.org/10.1007/s00521-019-04487-0>
9. Behera, R. K., Shukla, S., Rath, S. K., Misra, S., Software Reliability Assessment Using Machine Learning Technique. In *International Conference on Computational Science and Its Applications*, Springer, Cham, 2018, 403-411. https://doi.org/10.1007/978-3-319-95174-4_32
10. Bouveyron, C., Fauvel, M., Girard, S. Kernel Discriminant Analysis and Clustering with Parsimonious Gaussian Process Models. *Statistics and Computing*, 2015, 25, 1143-1162. <https://doi.org/10.1007/s11222-014-9505-x>
11. Breastcancer. Accessed via <https://archive.ics.uci.edu/ml/datasets/breast+cancer>, 1988.
12. Cao, L. Coupling Learning of Complex Interactions. *Information Processing & Management*, 2015, 51, 167-186. <https://doi.org/10.1016/j.ipm.2014.08.007>
13. Cao, Y., Lu, Y., Pan X., Sun N. An Improved Global Best Guided Artificial Bee Colony Algorithm for Continuous Optimization Problems. *Cluster Computing*, 2018, 1-9. <https://doi.org/10.1007/s10586-018-1817-8>
14. Chen, L., Wang, S., Wang, K., Zhu, J. Soft Subspace Clustering of Categorical Data with Probabilistic Distance. *Pattern Recognition*, 2016, 51, 322-332. <https://doi.org/10.1016/j.patcog.2015.09.027>
15. Chen, M., Hao, Y., Hwang, K., Wang, L., Wang, L. Disease Prediction by Machine Learning Over Big Data from Healthcare Communities. *IEEE Access*, 2017, 5, 8869-8879. <https://doi.org/10.1109/ACCESS.2017.2694446>
16. Chen, Y., Li, H., Chen, M., Dai, Z., Li, H., Zhu, M. Enhancing Feature Selection with Density Cluster for Better Clustering. *Computational Methods in Systems and Software*, 2018, 138-150. https://doi.org/10.1007/978-3-030-00211-4_15
17. Connor, B. P., Crawford M. R., Holder, M. D. An Item Response Theory Analysis of the Subjective Happiness Scale. *Social Indicators Research*, 2015, 124, 249-258. <https://doi.org/10.1007/s11205-014-0773-9>
18. Deangelis, L., Dias, J.G. Mining Categorical Sequences from Data Using a Hybrid Clustering Method. *European Journal of Operational Research*, 2014, 234, 720-730. <https://doi.org/10.1016/j.ejor.2013.11.002>
19. Emad, M. M., Enas, M. F. E, Khaled. T. W., Akram, I. S. Feature Selection Approach Based on Firefly Algorithm and Chi-square. *International Journal of Electrical and Computer Engineering (IJECE)*, 2018, 8(4), 2338-2350. <https://doi.org/10.11591/ijece.v8i4.pp2338-2350>
20. Eyestate. Accessed via <https://archive.ics.uci.edu/ml/datasets/congressional+voting+records>, 2013.
21. Fengmao, L., Guowu, Y., William, Z., Chuan, L. Generative Classification Model for Categorical Data Based on Latent Gaussian Process. *Pattern Recognition Letters*, 2017, 92, 56-61. <https://doi.org/10.1016/j.patrec.2017.03.025>
22. Gal, Y., Chen, Y., Ghahramani, Z. Latent Gaussian Processes for Distribution Estimation of Multivariate Categorical Data. *International Conference on Machine Learning*, Lille, France, 2015.
23. García, S., Julián, L., Francisco, H. Tutorial on Practical Tips of the Most Influential Data Preprocessing Algorithms in Data Mining. *Knowledge-Based Systems*, 2016, 98, 1-29. <https://doi.org/10.1016/j.knsys.2015.12.006>
24. Gene. Accessed via [https://archive.ics.uci.edu/ml/datasets/Molecular+Biology+\(Splice-junction+Gene+Sequences\)](https://archive.ics.uci.edu/ml/datasets/Molecular+Biology+(Splice-junction+Gene+Sequences)), 1992.
25. Gokulnath, C. B., Shantharajah, S. P. An Optimized Feature Selection Based on Genetic Approach and Support Vector Machine for Heart Disease. *Cluster Computing*, 2019, 22, 14777-14787. <https://doi.org/10.1007/s10586-018-2416-4>
26. Green, O., McColl, R., Bader, D. A. A Fast Algorithm for Streaming Betweenness Centrality. *International Conference on Privacy, Security, Risk and Trust (PASAT) and International Conference on Social Computing (SocialCom)*, IEEE, 2012, 11-20. <https://doi.org/10.1109/SocialCom-PASSAT.2012.37>
27. Hinton, G. E., Salakhutdinov, R. R. Reducing the Dimensionality of Data with Neural Networks. *American*

- Association for the Advancement of Science, 2006, 313, 504-507. <https://doi.org/10.1126/science.1127647>
28. Information on top causes of death accessed via <https://www.who.int/en/news-room/fact-sheets/detail/the-top-10-causes-of-death>
29. Jia, H., Yiu-ming, C., Jiming, L. A New Distance Metric for Unsupervised Learning of Categorical Data. *IEEE Transactions on Neural Networks and Learning Systems*, 2016, 27, 1065-1079. <https://doi.org/10.1109/TNNLS.2015.2436432>
30. Karaboga, D., Ozturk, C. A Novel Clustering Approach: Artificial Bee Colony (ABC) Algorithm. *Applied Soft Computing Journal*, 2011, vol. 11, no. 1, 652-657. <https://doi.org/10.1016/j.asoc.2009.12.025>
31. Killian, T. W., Samuel, D., George, K., Finale, D. Robust and Efficient Transfer Learning with Hidden Parameter Markov Decision Processes. *Conference on Neural Information Processing Systems*, 2017, 6250-6261.
32. Kourtellis, N., Tharaka, A., Ramanuja, S., Adriana, L., Rahul, T. Identifying High Betweenness Centrality Nodes in Large Social Networks. *Social Network Analysis Mining*, 2013, 899-914. <https://doi.org/10.1007/s13278-012-0076-6>
33. Lam, D., Wei, M., Wunsch, D. Clustering Data of Mixed Categorical and Numerical Type with Unsupervised Feature Learning. *IEEE Access*, 2015, 3, 1605-1613. <https://doi.org/10.1109/ACCESS.2015.2477216>
34. Lee, M. J., Jungmin, L., Jaimie, Y. P., Ryan, H. C., Chin, C. Qube: A Quick Algorithm for Updating Betweenness Centrality. *International Conference on World Wide Web*, ACM, 2012, 21. <https://doi.org/10.1145/2187836.2187884>
35. Liao, T., Krzysztof, S., Marco, A. M., Thomas, S., Marco, D. Ant Colony Optimization for Mixed-Variable Optimization Problems. *IEEE Transactions on Evolutionary Computation*, 2014, 18, 503-518. <https://doi.org/10.1109/TEVC.2013.2281531>
36. Manoj, R. J., Praveena, A. M. D., Vijayakumar, K. An ACO-ANN Based Feature Selection Algorithm for Big Data. *Cluster Computing*, 2019, 22, 3953-3960. <https://doi.org/10.1007/s10586-018-2550-z>
37. Mehdi, H. A., Nasser, G., Mohammad, E. B. Text Feature Selection Using Ant Colony Optimization. *Expert Systems with Applications*, 2009, 36(3), Part 2, 6843-6853. <https://doi.org/10.1016/j.eswa.2008.08.022>
38. Mustafa, S. U., Yilmaz, N., Inan, O. Feature Selection Method Based on Artificial Bee Colony Algorithm and Support Vector Machines for Medical Datasets Classification. *The Scientific World Journal*, 2013. <https://doi.org/10.1155/2013/419187>
39. Park, I. K., Choi, G. S. Rough Set Approach for Clustering Categorical Data Using Information-Theoretic Dependency Measure. *Information Systems*, 2015, 48, 289-295. <https://doi.org/10.1016/j.is.2014.06.008>
40. Sajana, T., Rani C. M. S., Narayana, K. V. A Survey on Clustering Techniques for Big Data Mining. *Indian Journal of Science and Technology*, 2016, 9. <https://doi.org/10.17485/ijst/2016/v9i3/75971>
41. Salem, B. S., Naouali, S., Chtourou, Z. A Fast and Effective Partitional Clustering Algorithm for Large Categorical Datasets Using a K-means Based Approach. *Computers & Electrical Engineering*, 2018, 68, 463-483. <https://doi.org/10.1016/j.compeleceng.2018.04.023>
42. Sato, Y., Izui, K., Yamada, T., Nishiwaki, S. Data Mining Based On Clustering and Association Rule Analysis for Knowledge Discovery in Multiobjective Topology Optimization. *Expert Systems with Applications*, 2019, 119, 247-261. <https://doi.org/10.1016/j.eswa.2018.10.047>
43. Sheth, R., Khardon, R. Excess Risk Bounds for the Bayes Risk Using Variational Inference in Latent Gaussian Models. *Advances in Neural Information Processing Systems*, 2017, 5151-5161.
44. Shen, Y., Morteza, M., Georgios, B. G. Online Categorical Subspace Learning for Sketching Big Data with Misses. *IEEE Transactions on Signal Processing*, 2017, 65, 4004-4018. <https://doi.org/10.1109/TSP.2017.2701333>
45. Shifei, D., Zhang, N., Zhang, J., Xu, X., Shi, Z. Unsupervised Extreme Learning Machine with Representational Features. *International Journal of Machine Learning and Cybernetics*, 2015, 13042. <https://doi.org/10.1007/s13042-015-0351-8>
46. Silvestre, C., Cardoso, M., Figueiredo, M. A. T. Feature Selection for Clustering Categorical Data with an Embedded Modelling Approach. *Expert Systems*, 2015, 32, 444-453. <https://doi.org/10.1111/exsy.12082>
47. Singh, V., Verma, N. K., Cui, Y. Type-2 Fuzzy PCA Approach in Extracting Salient Features for Molecular Cancer Diagnostics and Prognostics. *IEEE Transactions on Nanobioscience*, 2019, 18, 482-489. <https://doi.org/10.1109/TNB.2019.2917814>
48. Society, T. X., Shengrui, W., Qingshan, J., Joshua Z. H. A Novel Variable-Order Markov Model for Clustering Categorical Sequences. *IEEE Transactions on Knowledge and Data Engineering*, 2014, 26, 2339-2353. <https://doi.org/10.1109/TKDE.2013.104>

49. SPECT. Accessed via <https://archive.ics.uci.edu/ml/datasets/spect+heart>, 2001.
50. Su, J., Chunjing, S. Clustering Categorical Data Based on Within-Cluster Relative Mean Difference. *Open Journal of Statistics*, 2017, 7, 173-181. <https://doi.org/10.4236/ojs.201772013>
51. Vasiljeva, I., Arandelovic, O. Automatic Knowledge Extraction from EHRs. *IJCAI Workshop on Knowledge Discovery in Healthcare Data*, 2016.
52. Vasiljeva, I., Arandjelović, O. Towards Sophisticated Learning from EHRs: Increasing Prediction Specificity and Accuracy Using Clinically Meaningful Risk Criteria. *International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2016, 38, 2452-2455. <https://doi.org/10.1109/EMBC.2016.7591226>
53. Wang, C., Dong, X., Zhou, F., Cao, L., Chi, C. Coupled Attribute Similarity Learning on Categorical Data. *IEEE Transactions on Neural Networks and Learning Systems*, 2015, 26, 781-797.
54. Xu, J., Tang, B., He, H., Man, H. Semisupervised Feature Selection Based on Relevance and Redundancy Criteria. *IEEE Transactions on Neural Networks and Learning Systems*, 2017, 28, 1974-1984. <https://doi.org/10.1109/TNNLS.2016.2562670>



This article is an Open Access article distributed under the terms and conditions of the Creative Commons Attribution 4.0 (CC BY 4.0) License (<http://creativecommons.org/licenses/by/4.0/>).