Available online at **www.ijmrhs.com**

ISSN No: 2319-5886

International Journal of Medical Research & Health Sciences, 2018, 7(8): 130-140

Prediction of Accuracy for Hepatocellular Carcinoma Patients using Cluster based Feature Ranking

Preetam Pal, Birmohan Singh* and Manpreet Kaur

Department of Computer Science and Engineering, Sant Longowal Institute of Engineering and Technology, Punjab, India

**Corresponding e-mail: birmohans@gmail.com*

ABSTRACT

In 2014, hepatocellular carcinoma (HCC) cancer ranks second in the list of reasons for cancer-oriented deaths over the world. The incidences of hepatocellular cancer have approximately doubled, in the last two decades and the mortality rate due to HCC has also increased. There will be approximately 30,200 liver cancer deaths in 2018. The clinician provides treatments to HCC patients by using evidence-based medicine, which may not effectively resolve the problem of each patient. For assisting the decision making of the clinician, research works have been done in this field to extract information from these clinical data using computational methods. In this paper, a methodology has been proposed for the prediction of hepatocellular carcinoma patient data. A two-phase cluster based feature ranking procedure has been proposed and applied to the pre-processed data. Markov Blanket-based clustering method has been proposed, in which, the redundancy among the features is computed to rank the features. Total 6 different classifiers namely C4.5, ENSEMBLE, ANN, kNN, Naive Bayes, and SVM have been used for evaluation of the proposed methodology in terms of classification accuracy on HCC data by comparing it with some other most common feature selection methods (ReliefF, mRMR, MIM, and FCBF). The better the classification accuracy of the proposed methodology shows its effectiveness for prediction of HCC data.

Keywords: Hepatocellular carcinoma, Markov Blanket, Feature ranking, Classifier

INTRODUCTION

Hepatocellular carcinoma (HCC) is one of the most frequently found primary liver cancer [1]. According to reports of International Agency for Research on Cancer, World Health Organization (IARC-WHO), published over cancer in 2014, hepatocellular cancers are the second most frequent reason for cancer-related deaths worldwide. As per this report, from the last two decades, incidences of the HCC have doubled, and mortality rate due to HCC has been increased [2]. HCC is created in the chronic liver inflammation setting and most commonly associated with the infection of viral hepatitis (hepatitis B or C) or vulnerable to toxins like aflatoxin or alcohol [1]. Some of the possible reasons are inadequate prevention strategies for hepatitis C virus (HCV) and hepatitis B virus (HBV) infection. A rapid increase in lifestyle-related diseases like alcoholic liver disease and non-alcoholic fatty liver disease (NAFLD) are also some of the other main causes of HCC [2].

According to statics of WHO, 8.2 million cases of deaths and 14.1 million cases of cancer are noticed in the year 2012 [3]. There are an estimated 1.4 million deaths cases found per year related to hepatitis liver cancer and cirrhosis. Out of these nearly 47% are related to hepatitis B virus, 48% to hepatitis C virus. Viral hepatitis is also an increasing cause of mortality among HIV infected people. Nearly 2.6 millions of people who are infected with HIV are co-infected by hepatitis B and 2.9 with hepatitis C virus. Approximately 240 millions of people are infected with chronic hepatitis B virus, 130 to 150 million with chronic hepatitis C virus [4]. More than 90% of primary liver cancers cases are covered by hepatocellular carcinoma (HCC). As per the report of Marinho, et al., admissions of HCC related incidence has been tripled from 1993 to 2005, with a proportional rise in the overall costs of admission [5]. According to the report of Portuguese Society of Hepatology (PSH) number of the liver, cases has increased by approximately 70% from the year 2010 to 2015, seeking a high-level national awareness for liver diseases [3]. In Taiwan, HCC is a leading cause of cancer-related deaths since 1984, results in 7700 annual cancer deaths (with a mortality rate of 25.77 liver cancers

in per thousand peoples) [1]. As per the report of American Cancer Society, there will be 42,220 new cases of liver cancer during 2018 in the US, and about three-fourths of which will be the cases of HCC. Liver cancer incidents are increasing about 3% per year and the death rate has increased 2.5% per year from 2006 to 2015.

According to Population-Based Cancer Registries (PBCRs) report of the National cancer registry program of the Indian Council of Medical Research (ICMR) available on the website (www. ncrpindia.org), provides the information regarding various cancers from the year 2012 to 2014. Delhi (19746), Thiruvananthapuram District (15640), Mumbai (13357), Chennai (11659) and Kollam (11012) are top five PBCRs registering for a maximum number of cancer cases [2]. The ratio of male: female patient of HCC in India is 4:1.2. According to a survey report of a verbal autopsy has been conducted in 1.1 million homes throughout the country, 6.8/100,000 and 5.1/100,000 are the standard mortality rate of HCC for men and women in India. Various unpublished records from the different tertiary care centers provide strong evidence for a rapid increase of HCC incidences in India [2].

Liver transplantation, local ablation therapies, e.g., percutaneous ethanol injection therapy (PEIT), surgical resection, transcatheter arterial chemoembolization (TACE), microwave coagulation therapy (MCT) and target therapy are some examples of frequently following modalities for treatment of HCC [1]. Hepatic resection is another effective treatment and standard modality used for HCC protection. However, even with improvements in diagnosis and treatment, the overall mortality in the patients of HCC is higher than in the other types of cancer patients [6].

Different researchers who have worked on hepatocellular carcinoma are presented here. In the year 2013, a comparison was made over predictive models (logistic regression and artificial neural network) of mortality for HCC patients undergoing resection. And evaluation of the performances of logistic regression and artificial neural network models with different survival year was made. Better performance results were achieved with ANN at one, three, and five-year models [6]. In the same year, Atupelage, et al., proposed a feature descriptor for observing the characteristics of the histopathological textures in a discriminative manner. It used the fractal geometric analysis methodology with four multifractal measures for the making of the eight-dimensional feature space. It also used a bag-of-feature-based classification model for discrimination of multiple HCC images into 5 groups by using Edmondson and Steiner's grading system. They used 3 feature selection methods for searching the most discriminant feature vectors in order to obtain higher accuracy in the classifier. Different experiments were performed for evaluations are:

- Classification of non-neoplastic tissues and tumors, and
- Grading the Hepatocellular Carcinoma images into five different classes [7].

In the next year, recurrence predictive models were developed for the patients of HCC who were taking treatment of radiofrequency ablation (RFA). The authors used different feature selection methods (simulated annealing, genetic algorithm, random forests and hybrid methods SA+RF and GA+RF) for the selection of an important subset of features from 16 clinical features. Predictive models were made with support vector machine. Better results were achieved with hybrid methods [1]. In the year 2015, Santos, et al., proposed an oversampling approach, based on the cluster, for accounting the heterogeneity of the patients surviving with hepatocellular carcinoma. Pre-processing was done by using data imputation and appropriate distance metrics for handling the heterogeneous and missing data. The approach was applied for reducing the effect of reduced sizes on survival prediction of underlying patient profiles. They used logistic regression and neural networks for classification [3].

In the present work on the classification of HCC data, a methodology has been proposed that consists of data pre-processing, feature selection and classification. The presence of missing values in the data induces the need for filling these missing values. A cluster-based feature ranking method has been proposed that works in 2 phases, in the first phase, irrelevant features have been removed. In the second phase, cluster formation and picking the features to rank has been presented in this paper. The 5 commonly used classifiers for medical diagnostic problems have been applied for checking the performance. The proposed feature ranking method:

- Is capable of determining the numbers of cluster automatically, and
- Works efficiently for the higher dimensional data also, since the irrelevant features are removed in the first phase.

Definitions and Framework

In this section, some basic concepts related to the proposed feature selection methodology like, information theory, Markov Blanket, feature relevance, and redundancy are discussed.

Shannon's information theory provides strong criteria for quantizing the information about random variables using a probabilistic framework by defining entropy and mutual information that is used for determining the relevance between variables in many feature selection algorithms [8].

In the year 1994, the authors defined 3 levels of feature relevancy by categorizing them into 3 disjoint sets; strongly relevant features, weakly relevant features and irrelevant features based on a probabilistic framework. Where strongly relevant features are the most important feature as these contain unique information about target class, irrelevant features are unnecessary features as these contain no information about target class, and weakly relevant features are partially important as these contain some (not unique) information about target class [9].

Redundancy among feature is considered by using a correlation measure between features like the Pearson correlation coefficient (linear correlation measure) and symmetric uncertainty (a non-linear correlation measure) [10]. In the year 1996, Koller and Sahami proposed a cross-entropy framework called Markov Blanket for identifying and removing redundant and irrelevant features [11]. In the year 2004, the authors proposed an approximation framework named approximate Markov Blanket for determination and elimination of redundant features via explicitly considering the feature redundancy in a faster way [10].

If a discrete random variable X has alphabets χ, and probability density function is p (x)=Pr {X=x}, x ϵ χ. The entropy of X, H(X) is defined as [12]:

$$H(X) = \sum_{x \in \chi} p(x).\log \mathrm{p}(x) \tag{1}$$

For two random variables X and Y, the mutual information I (X;Y) existing between them is defined as follows [12]:

$$I(X;\mathrm{Y}) = \sum_{x \in \chi} \sum_{y \in \chi} \frac{p(x,y)}{p(x).p(y)} \tag{2}$$

The non-linear correlation between two variables (X, Y) is calculated by using symmetric uncertainty as follows [10]:

$$SU(X,\mathrm{Y}) = 2\frac{I(X,Y)}{H(X)+H(Y)} \tag{3}$$

Definition 1 (Markov Blanket): In a data set X with feature set F, for a given feature $F_i \in F$, let M_i □ F ($F_i \notin M_i$), M_i is said to be a Markov Blanket of F_i iff, P (F-M_i-{F_i}, C | F_i, M_i)=P (F-M_i-{F_i}, C | M_i)

Definition 2 (Approximate Markov Blanket): For two relevant features F_i and F_m ($i \neq m$), F_m is considered as an approximate Markov Blanket, feature for F_i iff, SU (F_m, C) ≥ SU (F_i, C) and SU (F_m, F_i) ≥ SU (F_i, C).

Various information theoretic feature selection algorithms like MIFS, MRMR, etc. make consideration of redundancy among candidate feature with already selected features in the process of incremental search via cumulative sum measure of correlation between already selected features and candidate feature using mutual information [13,14]. While many of the algorithms make consideration of redundancy implicitly with relevance, in the process of incremental search for the next best candidate feature for including in an already selected feature subset by using the cumulative sum measure. But these methods are more time complex as these apply incremental search, also feature redundancy consideration is not strong because the redundancy of a feature with a subset is calculated using the average of cumulative sum measure of redundancy of that feature with each feature of that subset.

In the proposed feature ranking method, redundancy among the features is handled explicitly by using approximate Markov Blanket concept for making clusters of weakly relevant redundant features with respect to strongly relevant features. It is a fast and efficient way of dealing with feature redundancy. Approximate Markov Blanket-based clustering method has been used by many researchers for considering redundancy among features like, Song, et al., used it for making clusters of redundant features in their proposed feature selection method and Wang, et al., used it for making clusters of redundant features in their proposed feature selection method named SRFS [15,16]. Markov Blanket-based clustering method has been used in the proposed methodology for computing the redundancy among features in a faster and efficient way to rank the features.

Proposed Cluster-Based Feature Ranking

The features which are strongly relevant and contain more relevant (important) information relative to class prediction; weakly contain partially relevant (not unique) information relative to class prediction whereas irrelevant features contain no information about the class. So, all strongly relevant features, some of the weakly relevant but non-redundant features and none of the irrelevant features should be included in the final optimal feature subset.

Thus, a feature selection method should consider a process of selecting all strongly relevant features, removing irrelevant features and selecting most necessary weakly relevant features. To fulfill these objectives, we have proposed a novel feature ranking method, which ranks relevant feature using a clustering framework by considering:

- The relevance of strongly relevant features with target class, and
- Redundancy score of weakly relevant features in clusters with respect to their strongly relevant representative features.

For the proposed feature ranking method, in the first phase, removal of irrelevant features has been carried out. In the second phase, cluster construction has been done. Thereafter, using clusters thus formed we rank strongly relevant and weakly relevant features from the feature set by using class relevance of strongly relevant features and redundancy score of weakly relevant features present in clusters. Redundancy score calculation of a feature 'f' in a cluster with respect to its representative feature 'r' is Redundancy score (f,r)=I (f; L|r), where L is a class label. The feature ranking method used in the proposed methodology is:

Input: Dataset X (N: instances, D: features), Label L, relevance threshold α

Output: Final ranked features

//Removal of Irrelevant Feature

Fintial = $\{f_1, f_2, f_3, \ldots\ldots, f_D\}$ // F_{intial}: initial set of features

S = Null, $F_{relevant}$ = Null // $F_{relevant}$: set of relevant features

for i = 1: D do

if SU (f_i , L) ≥ α then

$F_{relevant} \leftarrow F_{relevant} \cup f_i$ // Add 'f_i' to the relevant feature set

endif

endfor

$F_{final} \leftarrow$ Sorted $F_{relevant}$ //Sorted in decreasing order

($F_{final} = \{f_1, f_2, f_3, \ldots\ldots\ldots, f_m\}$) // m: number of features in F_{final}

//Clusters construction

//a: Cluster creation (element selection)

$F_{remain} \leftarrow F_{final}$, $k \leftarrow 0$, $N_c \leftarrow 0$ //F_{remain}: features to be clustered, N_c: no. of clusters

$F_{Rep} \leftarrow$ NULL, $F_{Sec} \leftarrow$ NULL // F_{Rep}: cluster representatives features, C: set of clusters

while $F_{remain} \neq$ Null

$k \leftarrow k+1$, $N_c \leftarrow N_c+1$ // k: index of currently created cluster

$f_i \leftarrow$ first element in F_{remain} // most relevant feature yet not clustered

$i \leftarrow$ index of f_i in F_{final}

$C_{Rep} \leftarrow$ seqAdd(C_{Rep} , f_i) // seqAdd(A,B) : concatenate A and B sequentially

for j = i+1 to m

if $SU(f_i, L) \geq SU(f_j, L)$ & $SU(f_i, f_j) \geq SU(f_j, L)$ then

$C_k \leftarrow C_k \cup f_j$ // Add f_j in kth-cluster, C_k: k^{th} Cluster

$F_{Sec} \leftarrow F_{Sec} \cup f_j$ // Add f_j to F_{Sec}

$F_{remain} \leftarrow F_{remain} - f_j$ // Remove f_j from F_{remain}

endif

endfor

endwhile

//Ranking of relevant features

FinalRankedFeatures ← C_{Rep} // initializing FinalRankedFeatures by strongly relevant

cluster representative features

n_{sec} ← number of features in F_{Sec}

RedundancyScore[1: n_{sec}] ← 0 // Initialize redundancy score of each secondary feature by 0

for i = 1 to n_{sec}

$f_i \leftarrow i^{th}$ feature in F_{Sec}

for j = 1 to N_c

if $f_i \in C_j$

frep ← Representative feature of C_j

RedundancyScore(f_i) ← RedundancyScore(f_i) + FUNred(f_i, frep, L)

endif

endfor

endfor

$SortedF_{Sec}$ ← decreasingly sort features in F_{Sec} by RedundancyScore value

FinalRankedFeatures ← seqAdd(FinalRankedFeatures, $SortedF_{Sec}$)

Suppose, a dataset of 15 features, F_{intial}=[f_1, f_2, f_3, f_4, f_5, f_6, f_7, f_8, f_9, f_{10}, f_{11}, f_{12}, f_{13}, f_{14}, f_{15}] and class label is L. For a relevance threshold α, suppose we get a set of relevant features $F_{relevant}$=[f_1, f_3, f_4, f_7, f_9, f_{10}, f_{11}, f_{15}] and a set of irrelevant features [f_2, f_5, f_6, f_8, f_{12}, f_{13}, f_{14}]. After arranging in the decreasing order of SU with label 'L', F_{final}=[f_3, f_7, f_9, f_1, f_{10}, f_4, f_{15}, f_{11}]. Now in the next phase, cluster construction is carried out. Figure 1 gives an example to show 2 clusters formed from all the relevant features.

Figure 1 Clusters formed using relevant features

In this way, it results in 2 representative features (f_3 and f_{15}) and 2 sets of secondary features ({f_{10}, f_9, f_7, f_{11}} and {f_1, f_{11}, f_4}). Now to formulate the ranking, redundancy score of each secondary feature is computed, which requires their redundancy score with respect to their representative feature in clusters. The final ranking of features is achieved by making the concatenation of representative features, i.e. [f_3, $f1_5$] and decreasingly sorted secondary features by their redundancy score value, i.e. [f_{11}, f_4, f_7, f_9, f_{10}, f_1].

MATERIALS AND METHODS

The proposed methodology for prediction of the hepatocellular carcinoma patient data includes 3 steps: a) data pre-processing, b) feature selection and c) classification. Where data pre-processing is performed to fill the missing values present in the data along with discretizing it for applying different information theoretic frameworks used in the proposed methodology, then feature selection has been applied to features ranked according to their importance in class prediction, then classification is performed for evaluating the efficiency of the proposed methodology.

Data preprocessing: HCC survival dataset for experimentation purpose is taken from the UCI machine learning repository, and it has the following properties [3]:

It is a multivariate dataset; it has 165 numbers of instances with 49 attributes and a binary class label. There are attribute missing values in the dataset. It contains 3 types of attributes: 26 nominal or ordinal attributes, 4 integers valued attributes and 19 continuous attributes. In order to apply the dataset for evaluation of the proposed methodology, we have first pre-processed HCC survival dataset. The steps for preprocessing of the dataset are:

- Filling the missing values using cubic spline data interpolation method.
- Rounding of filled missing value in attribute columns with nearest non-missing values available to filled missing value in attributes column.
- Making discretization of continuous and integer type attributes using minimum description length (MDL) discretization method.

Spline interpolation is preferred over polynomial interpolation most of the times because interpolation error is small in spline interpolation even in case of lower degree polynomials. Spline interpolation handles the problem of Runge's phenomenon very efficiently, in which oscillation occurs between points on using polynomials of higher degree [17]. We have used the spline data interpolation method for filling missing values in the dataset. The spline is a kind of the interpolation technique which uses a piecewise polynomial as interpolant.

The original dataset contains nominal, ordinal and integer type attributes. They filled the missing values in a column of these attributes of the dataset are floating point values. To maintain the uniformity in these attributes, we have rounded off these missing value with a nearest non-missing value available in the column of those attributes.

The original dataset has some continuous-valued attributes, and the proposed feature ranking method is based on an information theoretic framework for declaring correlation between features, which requires calculation of mutual information and entropies. These calculations are difficult for continuous data, minimum description length (MDL) has been used for discretization of continuous attributes, where we have discretized continuous and integer type attributes of missing value-filled and rounded dataset [18].

Feature selection: The features in this work are selected from the proposed feature ranking method. The performance is checked for all these features, and the set of features giving better performance is selected and fed to a classifier for classification.

Classification: To evaluate the proposed methodology, 6 classifiers which are commonly used in the classification problems have been used. These 6 classifiers include tree-based C4.5, multiple learning algorithms based ensemble, artificial neural network, K-nearest neighbor, the probability-based Naïve Bayes, Support vector machine [19-24].

C4.5 is an extended form of ID3 decision tree algorithm that accepts both continuous and discrete features, handle missing values, solve overfitting problems by pruning, etc., type features for making of a decision tree using training instances that can be used for classification of new instances. In the year 2015, Kohestani, et al., have used the C4.5 decision tree algorithm for predicting the seismic liquefaction capability of the soil by earthquake based on the cone penetration test data [25]. In the year 2017, Ngoc, et al., used the C4.5 algorithm for classification of English documents into semantics (positive, negative, and neutral) [26].

Ensemble learning combines the several models to improve machine learning results. In the year 2010, Kuncheva, et al., used a random subspace ensemble model for classification of brain images, obtained with the help of functional magnetic resonance imaging (fMRI) [27]. In the year 2013, Zhang, et al., proposed a microscopic biopsy image classification model using cascade random subspace ensembles method including reject options in order to enhance

the reliability of the classification [28]. In the year 2016, Karasu and Baskan used ensemble subspace kNN classifier for classification of power quality disturbances [29].

Artificial neural networks are the particular type of computational models made by taking inspiration from biological neural networks and used for solving complex problems like classification, rapid information processing, learning and adaptation, pattern recognition and modeling, speech, vision, and control systems. In the year 2011, Turnip and Hong, proposed an adaptive neural network classifier of 6 different mental tasks from EEG-based P300 signals [30]. In the year 2016, Khadse, et al., proposed an artificial neural network model based on conjugate gradient back-propagation for real-time power quality assessment [31]. In the year 2017, Khadse, et al., proposed an electromagnetic compatibility estimator model using the artificial neural network with scaled conjugate gradient algorithm [32].

The k-Nearest Neighbor (kNN) is used for both classification and regression purposes [33]. It searches for k-nearest instances into training instances using some distance measures and refers class which is most common in them. In the year 2012, Ramteke and Monali used kNN method for classification of CT brain images into normal and abnormal classes [34]. In the year 2014, Babu, et al., used a k-NN classifier for off-line handwritten digit recognition using structural features [35]. In the year 2016, Adeniyi, et al., used kNN for automated web usage data mining and recommendation system making for classification of online and in real-time to identify clients/visitors into particular user groups [36].

Naive Bayes uses probabilistic frameworks for classification by using Bayes theorem and considering independence among the features. In the year 2013, Soelistio, et al., proposed a model for analyzing digital newspaper sentiment polarity by using a Naive Bayes classifier algorithm [37]. In the year 2014, Mohamad, et al., proposed an automatic bacteria identification framework for classification of 3 famous classes of bacteria namely Cocci, Bacilli and Vibrio from microscopic morphology using the Naive Bayes classifier [38]. In the same year, Saleh, et al., used different models of Naive Bayes classifiers for authorship attribution in Arabic like simple Naive Bayes (NB), multinomial Naïve Bayes (MNB), multi-variant Bernoulli Naive Bayes (MBNB) and multi-variant Poisson Naive Bayes (MPNB). They achieved the best result on MBNB [39]. In the year 2017, Krishnan utilized a Naive Bayes classifier for emotion recognition from tweets [40].

Support vector machine (SVM) is used for both classification and regression problems. It uses training instances for finding the best hyperplane that categorizes the dataset into two classes; then the model is used for classifying new instances. In the year 2014, Singh, et al., proposed a land use/land cover (LULC) estimation model by using a support vector machine (SVM) classifier [41]. In the year 2015, Harris used clustered SVM classifier for credit scorecard development [42]. In the same year, Demidova, et al., used SVM in combination with fuzzy clustering algorithm for developing an approach for object's classification [43].

In this work, we have used kernel distribution for the Naive Bayes classifier. In kNN, we have taken k equals 1 and Euclidean distance as an input parameter of the classification model. In ANN, we have taken 10 layered scaled conjugate gradient backpropagation neural network models for classification. In C4.5, we have taken 100 numbers of maximum splits with Gini's diversity index as split criteria in the classification model. In SVM, we have used Quadratic kernel function for mapping of the training data into kernel space. We have used subspace Aggregation Method and nearest neighbors learner with 30 number of learners and 25 subspace dimension for in an ensemble classification model. Implementation of all these classifiers has been done by using MATLAB 2017.

RESULTS AND DISCUSSION

The proposed methodology for prediction of HCC patient data has been evaluated using 6 classifiers. After pre-processing, the input data is applied to a proposed two-phase cluster-based feature ranking method. For comparison of the proposed methodology, a combination of 4 different feature selection algorithms and 6 classifiers are used. The features selection algorithms used include an individual feature evaluation based algorithm ReliefF [44], a feature subset evaluation based algorithm mRMR [14], mutual information based ranking method MIM, and a fast correlation base algorithm FCBF [10,45].

Relevance threshold in the proposed method is set to the symmetric uncertainty of $(\lfloor n/\log n \rfloor)^{th}$ ranked feature (where n is the number of features of the dataset) like used in Yu, et al., in 2004 and Wang, et al., in 2017. A comparison of the average classification accuracy has been presented the Table 1. The results have been compiled on 10 fold cross-

validation procedure using an average of 20 iterations. SVM with quadratic kernel function gives better results as compared to the other combinations.

Table 1 A comparison of the classification accuracy

Methods	C4.5	Ensemble	ANN	kNN	Naive Bayes	SVM
ReliefF	66.30 ± 0.04	67.07 ± 0.04	65.81 ± 0.01	63.01 ± 0.04	67.65 ± 0.01	67.21 ± 0.05
mRMR	67.50 ± 0.03	70.26 ± 0.04	66.62 ± 0.02	66.48 ± 0.01	69.53 ± 0.04	69.80 ± 0.03
MIM	65.36 ± 0.03	69.19 ± 0.04	67.31 ± 0.02	66.79 ± 0.01	68.31 ± 0.05	68.50 ± 0.06
FCBF	65.67 ± 0.03	68.82 ± 0.04	63.86 ± 0.03	65.53 ± 0.01	73.17 ± 0.01	69.16 ± 0.03
Proposed	66.58 ± 0.03	71.49 ± 0.04	71.88 ± 0.02	72.10 ± 0.02	73.95 ±0.04	76.25 ± 0.02

Figure 2 shows a comparison of the accuracy of different methods on 6 different classifiers. From the comparative analysis in this figure, it has been deduced that the proposed method achieves better results for HCC data on most of the classifiers.

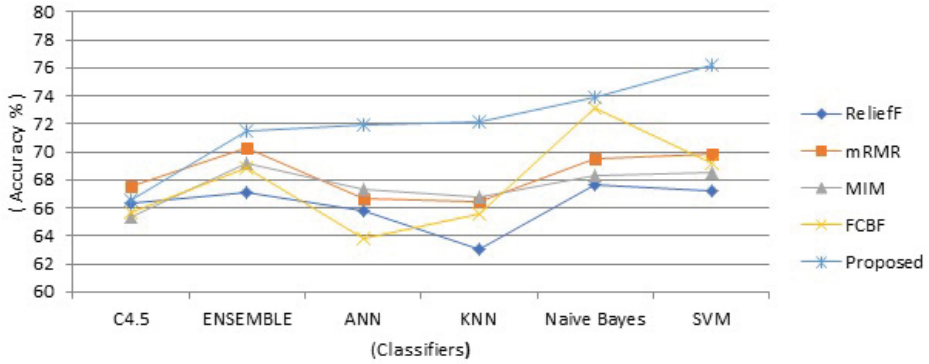

Figure 2 Accuracy comparisons of feature selection methods on different classifiers

In Table 2, five performance metrics accuracy, sensitivity, specificity, AUC (area under the receiver operating characteristic curve) and F-measure (a measure of test's accuracy) have been used to evaluate the performance of different feature selection method on SVM classifier. The proposed method has a high mean accuracy of 76.25% with a standard deviation of 0.02%, whereas FCBF has a second highest mean accuracy of 69.16% with a standard deviation of 0.03%. The sensitivity of the proposed methodology is 0.79 with a standard deviation of 0.02 and specificity of 0.72 with a standard deviation of 0.05. AUC and F-measure of the proposed method also achieve the best rank with mean value and a standard deviation of 0.81 ± 0.02 and 0.80 ± 0.02 respectively. The proposed method with SVM classifier gives better performance results, which show its effectiveness for classification of HCC data.

Table 2 Average performance of different methods on SVM

Algorithm	Accuracy (%)	Sensitivity	Specificity	AUC	F-measure
	Mean ± Std	Mean ± Std	Mean ± Std	Mean ± Std	Mean ± Std
ReliefF	67.21 ± 0.05	0.78 ± 0.03	0.49 ± 0.10	0.74 ± 0.04	0.75 ± 0.01
mRMR	69.80 ± 0.03	0.77 ± 0.03	0.56 ± 0.03	0.80 ± 0.02	0.78 ± 0.02
MIM	68.50 ± 0.06	0.75 ± 0.07	0.57 ± 0.05	0.74 ± 0.07	0.74 ± 0.06
FCBF	69.16 ± 0.03	0.76 ± 0.03	0.57 ± 0.04	0.77 ± 0.02	0.75 ± 0.01
Proposed	76.25 ± 0.02	0.79 ± 0.02	0.72 ± 0.05	0.81 ± 0.02	0.80 ± 0.02

In the year 2015, Santos, et al., proposed a cluster-based oversampling approach and they achieved an accuracy of 75.2% with 0.011 standard deviations for augmented sets. The results achieved in the proposed work are 76.25% with 0.02 standard deviation.

CONCLUSION

The present study proposes a methodology for prediction of mortality for hepatocellular carcinoma (HCC) patients. This includes a 2-phase cluster based feature ranking method the output of which is examined with 6 classifiers.

The results have also been compared with the results of the other researcher who have used the same dataset. The best output has been achieved with support vector machines for the data collected from the UCI machine learning repository. It has been observed that the SVM with the two-phase cluster based feature ranking improves the prediction accuracy significantly and decrease the miss-classification error. Also, this proposed methodology consisting of cluster-based feature ranking and support vector machine has shown better performance in prediction as compared to the other feature selection methods used in combinations with the other classifiers. The proposed methodology has great potential and can be used to support in decision making and prediction in HCC patient data.

DECLARATIONS

Conflict of Interest

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

REFERENCES

[1] Liang, Ja-Der, et al. "Recurrence predictive models for patients with hepatocellular carcinoma after radiofrequency ablation using support vector machines with feature selection methods." *Computer methods and programs in biomedicine,* Vol.117, No.3, 2014, pp. 425-34.

[2] S. K. Acharya and S. B. Paul, "Hepatocellular Cancer (HCC): Screening and Management," 2014.

[3] Santos, Miriam Seoane, et al. "A new cluster-based oversampling method for improving survival prediction of hepatocellular carcinoma patients." *Journal of biomedical informatics,* Vol. 58, 2015, pp. 49-59.

[4] World Health Organization. "Global health sector strategy on viral hepatitis 2016-2021. Towards ending viral hepatitis." 2016.

[5] Marinho, Rui Tato, José Giria, and Miguel Carneiro Moura. "Rising costs and hospital admissions for hepatocellular carcinoma in Portugal (1993-2005)." *World Journal of Gastroenterology,* Vol. 13, No. 10, 2007, p.1522.

[6] Chiu, Herng-Chia, et al. "Mortality predicted accuracy for hepatocellular carcinoma patients with hepatic resection using artificial neural network." *The Scientific World Journal,* 2013.

[7] Atupelage, Chamidu, et al. "Computational grading of hepatocellular carcinoma using multifractal feature description." *Computerized Medical Imaging and Graphics,* Vol. 37, No. 1, 2013, pp. 61-71.

[8] Yu, Sung-Nien, and Ming-Yuan Lee. "Conditional mutual information-based feature selection for congestive heart failure recognition using heart rate variability." *Computer Methods and Programs in Biomedicine,* Vol. 108, No. 1, 2012, pp. 299-309.

[9] John, George H., Ron Kohavi, and Karl Pfleger. "Irrelevant features and the subset selection problem." *Machine Learning Proceedings,* 1994, pp. 121-29.

[10] Yu, Lei, and Huan Liu. "Efficient feature selection via analysis of relevance and redundancy." *Journal of Machine Learning Research*, 2004, pp. 1205-24.

[11] Koller, Daphne, and Mehran Sahami. *Toward optimal feature selection*. Stanford InfoLab, 1996.

[12] Meyer, Patrick Emmanuel, Colas Schretter, and Gianluca Bontempi. "Information-theoretic feature selection in microarray data using variable complementarity." *IEEE Journal of Selected Topics in Signal Processing,* Vol. 2, No. 3, 2008, pp. 261-74.

[13] Battiti, Roberto. "Using mutual information for selecting features in supervised neural net learning." *IEEE Transactions on Neural Networks,* Vol. 5, No. 4, 1994, pp. 537-50.

[14] Peng, Hanchuan, Fuhui Long, and Chris Ding. "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy." *IEEE Transactions on Pattern Analysis and Machine Intelligence,* Vol. 27, No. 8, 2005, pp.1226-38.

[15] Song, Qinbao, Jingjie Ni, and Guangtao Wang. "A fast clustering-based feature subset selection algorithm for high-dimensional data." *IEEE Transactions on Knowledge and Data Engineering,* Vol. 25, No. 1, 2013, pp. 1-14.

[16] Wang, Yintong, et al. "An efficient semi-supervised representative's feature selection algorithm based on information theory." *Pattern Recognition,* Vol. 61, 2017, pp. 511-23.

[17] "Spline interpolation," Wikipedia, https://en.wikipedia.org/wiki/Spline_interpolation. Accessed: 20 May. 2018.

[18] Fayyad, Usama, and Keki Irani. "Multi-interval discretization of continuous-valued attributes for classification learning." 1993.

[19] Quinlan, J. Ross. *C4. 5: programs for machine learning*. Elsevier, 2014.

[20] Gul, Asma, et al. "Ensemble of a subset of kNN classifiers." *Advances in Data Analysis and Classification,* 2016, pp. 1-14.

[21] Charalambous, Christakis. "Conjugate gradient algorithm for efficient training of artificial neural networks." *IEE Proceedings G (Circuits, Devices and Systems),* Vol. 139, No. 3, 1992, pp. 301-10.

[22] Altman, Naomi S. "An introduction to kernel and nearest-neighbor nonparametric regression." *The American Statistician,* Vol. 46, No. 3, 1992, pp. 175-85.

[23] Rish, Irina. "An empirical study of the naive Bayes classifier." *IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence,* Vol. 3, No. 22, New York: IBM, 2001.

[24] Cortes, Corinna, and Vladimir Vapnik. "Support-vector networks." *Machine Learning,* Vol. 20, No. 3, 1995, pp. 273-97.

[25] Ardakani, A., and V. R. Kohestani. "Evaluation of liquefaction potential based on CPT results using C4. 5 decision tree." *Journal of AI and Data Mining,* Vol. 3, No. 1, 2015, pp. 85-92.

[26] Ngoc, Phu Vo, et al. "A C4. 5 algorithms for English emotional classification." 2017, pp. 1-27.

[27] Kuncheva, Ludmila I., et al. "Random subspace ensembles for fMRI classification." *IEEE Transactions on Medical Imaging,* Vol. 29, No. 2, 2010, pp. 531-42.

[28] Zhang, Yungang, et al. "Breast cancer diagnosis from biopsy images with highly reliable random subspace classifier ensembles." *Machine Vision and Applications,* Vol. 24, No. 7, 2013, pp. 1405-20.

[29] Karasu, Seçkin, and Saim Başkan. "Classification of power quality disturbances by using ensemble technique." *Signal Processing and Communication Application Conference (SIU), 2016 24th*, IEEE, 2016.

[30] Turnip, Arjon, and Keum-Shik Hong. "Adaptive neural network classifier for EEG signals of six mental tasks." *Control, Automation and Systems (ICCAS), 2011, 11th International Conference on*. IEEE, 2011.

[31] Khadse, Chetan B., Madhuri A. Chaudhari, and Vijay B. Borghate. "Conjugate gradient back-propagation based artificial neural network for real-time power quality assessment." *International Journal of Electrical Power and Energy Systems,* Vol. 82, 2016, pp. 197-206.

[32] Khadse, Chetan B., Madhuri A. Chaudhari, and Vijay B. Borghate. "Electromagnetic compatibility estimator using scaled conjugate gradient backpropagation based artificial neural network." *IEEE Transactions on Industrial Informatics,* Vol. 13, No. 3, 2017, pp. 1036-45.

[33] Cover, Thomas, and Peter Hart. "Nearest neighbor pattern classification." *IEEE Transactions on Information Theory,* Vol. 13, No. 1, 1967, pp. 21-27.

[34] Ramteke, R. J., and Y. Khachane Monali. "Automatic medical image classification and abnormality detection using K-Nearest Neighbour." *International Journal of Advanced Computer Research,* Vol. 2, No. 4, 2012, pp. 190-96.

[35] Babu, U. Ravi, Y. Venkateswarlu, and Aneel Kumar Chintha. "Handwritten digit recognition using K-nearest neighbor classifier." *Computing and Communication Technologies (WCCCT), 2014 World Congress on* IEEE, 2014.

[36] Adeniyi, D. A., Z. Wei, and Y. Yongquan. "Automated web usage data mining and recommendation system using K-Nearest Neighbor (KNN) classification method." *Applied Computing and Informatics,* Vol. 12, No. 1, 2016, pp. 90-108.

[37] Soelistio, Yustinus Eko, and Martinus Raditia Sigit Surendra. *"Simple text mining for sentiment analysis of political figure using naive bayes classifier method*." 2015.

[38] Mohamad, Noor Amaleena, et al. "Bacteria identification from microscopic morphology using naive bayes." *International Journal of Computer Science, Engineering and Information Technology,* Vol. 4, No. 1, 2014.

[39] Altheneyan, Alaa Saleh, and Mohamed El Bachir Menai. "Naïve Bayes classifiers for authorship attribution of Arabic texts." *Journal of King Saud University-Computer and Information Sciences,* Vol. 26, No. 4, 2014, pp. 473-84.

[40] Krishnan, Hema, M. Sudheep Elayidom, and T. Santhanakrishnan. "Emotion Detection of Tweets using Naïve Bayes Classifier." *Emotion*, 2017.

[41] Singh, Sudhir Kumar, et al. "Appraisal of land use/land cover of mangrove forest ecosystem using support vector machine." *Environmental Earth Sciences,* Vol. 71, No. 5, 2014, pp. 2245-55.

[42] Harris, Terry. "Credit scoring using the clustered support vector machine." *Expert Systems with Applications,* Vol. 42, No. 2, 2015, pp. 741-50.

[43] Demidova, Liliya, Yulia Sokolova, and Evgeny Nikulchev. "Use of fuzzy clustering algorithms ensemble for SVM classifier development." *International Review on Modelling and Simulations,* Vol. 8, No. 4, 2015, pp. 446-57.

[44] Robnik-Šikonja, Marko, and Igor Kononenko. "Theoretical and empirical analysis of ReliefF and RReliefF." *Machine Learning,* Vol. 53, No. 1-2, 2003, pp. 23-69.

[45] Lewis, David D. "Feature selection and feature extraction for text categorization." *Proceedings of the workshop on Speech and Natural Language*. Association for Computational Linguistics, 1992.