A Case Study of Medical Data Classification Using Hybrid Adaboost KNN along with Krill Herd Algorithm (KHA)

Dudekula Mahammad Rafi1*, Chettiar Ramachandra Bharathi2

1 Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Avadi, Chennai, Tamil; Vivekananda Institute of Engineering & Technology, JNTU University, Hyderabad, India
2 Department of ECE, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Avadi, Chennai, Tamil Nadu, India

Corresponding Author Email: rafiveltech@gmail.com

Received: 22 November 2018
Accepted: 16 January 2019

Keywords: case study investigation, medical classification, krill herd algorithm, hybrid Adaboost k-nearest neighbor, accuracy, sensitivity and specificity

1. INTRODUCTION

The target of information mining is to distinguish substantial, novel, possibly valuable, and justifiable relationships and examples in existing information. Finding valuable examples in information is known by names (e.g., Knowledge extraction, data revelation, data reapring, and information design preparing). The expression "information mining" is basically utilized by analysts, database specialists, and the business networks. The term Knowledge Discovery in Databases (KDD) alludes to the general procedure of finding valuable information from information. Information mining is data extraction from database. Information mining systems are utilized to get right restorative analysis.

Arrangement of information mining frameworks should be possible as indicated by the sort of information sources mined, database included, the sort of learning found, and mining strategies utilized. Order is the association of information in given classes. Arrangement approaches ordinarily utilize a preparation set where every one of the articles are as of now connected with realized class marks. The grouping calculation gains from the preparation set and manufactures a model. The model is utilized to characterize new items.

When a Classification show is constructed dependent on a preparation set, the class mark of an article can be predicted dependent on the trait estimations of the item and property estimations of the classes. Grouping is like Classification, bunching is the association of information in classes. In any case, the test increments as the enthusiasm for information mining develops quickly. As information digging systems for therapeutic information grouping has not been completely examined, there is an incredible potential for further work and intriguing bearings for research.

Krill Nearest Neighbor (KNN) strategy has been utilized in applications, for example, information mining, measurable example acknowledgment, picture preparing, acknowledgment of pennmanship, Electrocardiography (ECG) infecion order. The Krill Nearest Neighbor (KNN) strategy is an occurrence based learning technique that stores every single accessible datum set and orders new informational index dependent on similitude measure.

Half and half Adaboost Krill Nearest Neighbor (KNN) is a classifier. The Hybrid Adaboost Krill Nearest Neighbor (KNN) arrangement limiting the preparation set blunder and boosting the edge so as to accomplish the best speculation capacity. Because of its profitable nature, Hybrid Adaboost Krill Nearest Neighbor (KNN) has been connected to a wide scope of errands. Specifically, Hybrid Adaboost Krill Nearest Neighbor (KNN) has been to perform great on numerous restorative analysis undertakings. Nonetheless, there is as yet a requirement for improving the Hybrid Adaboost Krill Nearest Neighbor (KNN) classifier's execution. Anyway the exchange of basically sick patients requires great coordination to give the symptomatic apparatuses and the most suitable treatment for their conditions [17]. Hence it is very important to classify the disease on time and thereby providing treatments to the patients for risk avoidance. Sometimes same properties with the same symptoms of diseases required different treatments. Therefore accurate decisions and classification of data is required.

Medical data are mounting at hospitals, health centers and clinics. Eventually the quantity of information in the medical domain is increasing in recent decades. Restorative information contain data about research center test outcomes, tolerant socioeconomics, drug store data, radiology reports and pictures, pathology reports, emergency clinic affirmation, release and exchange dates, release rundown and
advancement notes. An important concern of the healthcare industry is to provide individualized patient care and not to collect or maintain data. However, medical data is necessary for decision making, drug administration, diagnosis, statistical analysis and evidence. Hence, efficient and effective techniques are required to extract useful information from this data.

The goal of data classification is to assign input data to one of a finite number of classes. The process of designing a data classifier design starts with data collection for any real-world problem and then undergoes a series of steps as given in Figure 1.

![Figure 1. Data classifier design](image)

An information classifier initially requires choice of features for every issue space. Great order execution requires choice of successful features and determination of a classifier that can make restricted preparing information, memory and figuring power.

Researchers attempted to use contrasting methods to hint at progress accuracy of data request. It is central to perform danger data course of action with a particular ultimate objective to perceive the disease. In this way to crush those issues our proposed procedure is used. Here at first the pre-dealing will be associated with expel accommodating data and to change over sensible model from rough helpful datasets. In the wake of preprocessing for perfect assurance of features ACO based SVM is used. This system portrays the tumor data as run of the mill and sporadic. From this time forward our proposed system decisively arranges the tumor data using perfect features.

From this time forward to crush the issue a couple of new methods make sense of how to dissect infection in a totally data driven way, using multivariate course of action or backslide to perceive the disease. In this way to crush those issues our proposed procedure is used. Here at first the pre-dealing will be associated with expel accommodating data and to change over sensible model from rough helpful datasets. In the wake of preprocessing for perfect assurance of features ACO based SVM is used. This system portrays the tumor data as run of the mill and sporadic. From this time forward our proposed system decisively arranges the tumor data using perfect features.

2. CASE STUDY DESIGN

The main intention of the study is to classify the medical data with high accuracy. The primary objective of this Case Study work is to develop computer-aided systems to assist clinicians in decision making. To achieve the objective of study, we present a conceptual framework for designing case study. The data we study is collected from patients to breast cancer disease. The overall process of Case Study can be divided into four stages,

- **Stage 1: Preprocessing**
- **Stage 2: Optimal Feature selection using KHA**
- **Stage 3: Classification using HAKNN**
- **Stage 4: Case study Comparison**

The detailed process of the A Case Study approach using Hybrid Adaboost KNN is shown in Figure 2.

![Figure 2. A case study approach using hybrid adaboost KNN](image)

Preprocessing

Preprocessing is most important step in data mining, which collect the pre-processing will be applied to extract useful data and to convert suitable sample from raw medical datasets. In preprocessing, the non-numerical data and missing values are removed from the input dataset and obtained the numerical dataset. The preprocessed output is fed to the further process.

3. OPTIMAL FEATURE SELECTION USING KRILL HERD ALGORITHM (KHA)

After preprocessing optimal features are selected by using...
krill Herd Algorithm (KHA). KHA is a streamlining technique to tackling enhancement issues that depends on krill herd swarms of organic and ecological forms.

### 3.1 Krill Herd Algorithm (KHA), initialization

The main parameters of the Krill Herd Algorithm(KHA) are the total evolution number, the population size, \( D_{\text{max}}^{\text{max}} \) (Maximum induced motion), \( S_f \) (Sensing factor), \( r_t \) (Random number ) and \( R_{D}^{\text{max}} \) (Maximum diffusion speed). In this technique, Krill Herb represents the features value.

### 3.2 Fitness calculation

Evaluate the fitness utility depends on the maximum accuracy and moreover select the finest result. Fitness = Max(accuracy)

Setting of each individual krill are,

(a) krill individuals Development,
(b) Foraging action,
(c) Random dispersion.

### 3.3 Development initiated by other Krill individuals

Speed of every krill individual is impacted by the development of the other Krill's to keep up a high thickness. Three impacts namely neighborhood impact (x), target impact (y) and repulsive impact (z) are used to evaluate the direction of motion induced (\( \xi_m \)). Krill individual motion may be formulated as

\[
D_{\text{new}} = \xi_m D_{\text{max}} + \chi_t D_{\text{old}}
\]

The sensing distance \( S_d \) between the individual krills and the neighbors formulated by,

\[
S_d = \frac{1}{2N} \sum_{m}^{N} \left| F_m - F_m \right|
\]

### 3.4 Foraging action

The foraging velocity of \( m \)th krill individual can be expressed by

\[
F_{\text{new}} = S_d F_m + \chi_t F_{\text{old}}
\]

### 3.5 Random dispersion

To enhance the population diversity random diffusion process is mainly considered and it is expressed by

\[
R_{\text{new}} = \beta \times R_{D}^{\text{max}}
\]

### 3.6 Updating the krill position

In updating the position, the individual krill changes its present positions also, moves to better positions in light of induction movement, foraging movement and random dispersion movement. In this three movements, the upgraded position of the \( m \)th krill individuals during the interval of \( t \) and \( \Delta t \) might be communicated by

\[
P_m(t + \Delta t) = P_m(t) + \Delta t \frac{dP_m}{dt}
\]

Based on the above procedure, we select the optimal features and then the selected features are fed to the classification process.

### 3.7 Classification using Hybrid Adaboost KNN algorithm

Finally, the optimal features are furnished to Hybrid Adaboost KNN classifier for the purpose of classification.

### 3.8 Hybrid Adaboost KNN algorithm

Hybrid Adaboost technique combine multiple “weak classifiers” into a single “strong classifier”. Each weak classifier should be trained on a random subset of the total training set. Adaboost assigns a “weight” to each training sample, which determines the probability that each sample should appear in the training set. After training a classifier, Adaboost increases the weight of the classifier. All the learners are simple and weak and must have error less than 0.5. Otherwise, the process is stopped since its continuation makes the learning become difficult for the next classifier. Also, the initial probability of selecting sample is considered to be uniform. In fact, the weight of sample shows the importance of the sample. The final hypothesis is obtained through weighted voting of \( T \) number of weak hypotheses. The steps involved in this algorithm are shown below.

**Step 1:** Initialization of weight \( W \)

**Step 2:** In the case, \( t \leq T \text{and} \text{err} < 0.5 \), then normalize weight \( W^t \), so that, \( \sum_{i=1}^{N} W_i^t = 1 \)

**Step 3:** Call KNN, providing with the weight \( W^t \), get hypothesis \( h^t : X \rightarrow \{-1,1\} \)

### 3.9 Krill Nearest Neighbor (KNN) classifier

The KNN calculation is a strategy which is utilized for characterization of any items or different components dependent on the nearest preparing information which are accessible in the element space. The qualities closest to the Krill esteem will be picked for the characterization results. When the grouping of closest neighbor is done, convert it into vector esteem with fixed length by using the Euclidean separation work in KNN which is given in beneath articulation,

\[
E_d = \left( \sum_{k=1}^{N} (x_k - y_k)^2 \right)^{1/2}
\]

where \( x, y \) are feature values;

The premise of the KNN classifier is the little neighborhood in the comparable highlights. These procedures will give better exactness in arranging the outcomes.

**Step 4:** Compute \( \text{err}^t = \sum_{i=1}^{N} W_i^t e_i^t \). Where \( e_i = 1, \text{if } h^t(x_i) \neq y_i \), and 0 otherwise

**Step 5:** Set \( \alpha^t = 0.5 \text{log}(1 - \text{err}^t) / \text{err}^t \)

**Step 6:** Update the weights to be as follows,

\[
W_i^{t+1} = W_i^t \exp(2\alpha^t e_i^t)
\]

**Step 7:** Put \( T = t + 1 \) and the process repeats until \( \text{err} = 0 \).

After every classifier is prepared, the classifier's weight is determined dependent on its exactness. Increasingly precise
classifiers are given more weight. At long last we characterize
the medicinal information with high precision esteem.

4. RESULTS AND DISCUSSION

The Case Study of Medical classification is implemented
using Python software and the experiment is done using i5
processor with 3GB RAM.

4.1 Dataset description

4.1.1 Mammographic mass data set

These datasets are taken from the UCI machine learning
repository. The database comprises of around 2,620 cases.
For each case, dual images of every breast, inter related
patient information, like age, period of the tumor, subtlety
rating for varieties from the standard, American College of
Radiology (ACR) breast thickness rating are considered. The
Mammograms are digitized by various scanners depending
upon the wellspring of the data.

4.2 Evaluation metric

The evaluation metrics used here contains True Positive,
True Negative, False Positive and False Negative, Sensitivity,
Specificity and Accuracy.

\[
\text{Sensitivity} = \frac{TP}{TP + FN} \\
\text{Specificity} = \frac{TN}{FN + TN} \\
\text{Accuracy} = \frac{TP + TN}{TP + FN + FP + TN}
\]

4.3 Case study comparative analysis

We can build up that our Case Study achieves great
exactness for the medical data classification. Hybrid Adaboost
Algorithm together with KNN algorithm is utilized for
medical data classification in our investigation strategy.
And furthermore we can set up this forecast precision result
by contrasting different classifiers. Case Study classifier is
contrasted with the Hybrid Adboost KNN along with Krill
Herd Algorithm (KHA), existing Ant Colony Optimization
(AOC) and Support Vector Mechanism (SVM) classifiers. The
Comparison results are introduced in the Table 1.

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Proposed Method (KHA+HAKNN)</th>
<th>Existing Method (AOC+SVM)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>Sensitivity</td>
<td>Specificity</td>
</tr>
<tr>
<td>Mammographic Mass data</td>
<td>95.86%</td>
<td>98.5%</td>
</tr>
</tbody>
</table>

Accuracy:

![Figure 3. The comparison outcomes of the Accuracy measure](image)

Sensitivity:

![Figure 4. The comparison outcomes of the Sensitivity measure](image)

Specificity:

![Figure 5. The comparison outcomes of the Specificity measure](image)

4.4 Performance analysis

The results of case study assist in analyzing the
effectiveness of the prediction method. The results of the
breast cancer datasets are provided in below table 2

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Accuracy</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mammographic Mass data</td>
<td>95.86%</td>
<td>98.5%</td>
<td>96.5%</td>
</tr>
</tbody>
</table>

From Table 1, it is clear that the accuracy value for
mammographic mass data obtained using the proposed
method is 95.86 %, similarly the sensitivity and specificity
value obtained is 98.5 % and 97.6 % respectively.
5. CONCLUSION

Breast cancer is danger disease among ladies around the world. Case Study investigation technique of Breast cancer data classification can be done with help of optimal feature selection. Krill Herd Algorithm (KHA) is used to selecting optimal features. The optimal features are classified with Hybrid Adaboost KNN classification algorithm. Experimentation data sets are collected from UCI machine learning public database. Case Study performance evaluated using accuracy, sensitivity and specificity. The Study investigation Hybrid Adaboost KNN technique achieve the maximum accuracy, sensitivity and specificity were 95.86 %, 98.5 % and 96.5 %. Case Study investigation implemented using python software.

REFERENCES