AN INTELLIGENT CRF BASED CUTTLEFISH FEATURE SELECTION ALGORITHM FOR EFFECTIVE DIAGNOSIS

Gotam Singh Lalotra, R.S.Thakur

Department of Computer Applications, Maulana Azad National Institute of Technology, Bhopal, India-462003.

Email: singh.gotam@gmail.com

Received on 10-07-2016
Accepted on 25-08-2016

Abstract

In this paper, an Intelligent Condition Random Field (ICRF) based Cuttlefish Feature Selection Algorithm (ICRFCFA) for effective decision making over medical datasets has been proposed. This proposed feature selection algorithm helps to improve the prediction accuracy in less time. An ICRFCFA for feature selection is employed then the selected features are involved with cuttlefish optimization approach for finalizing the necessary features for effective decision making with less computation cost. Experiments are carried out for evaluating the efficiency of the proposed feature selection algorithm using Diabetes and Heart disease datasets.

Keywords: Feature Selection, Cuttlefish, Intelligent Conditional Random Field (ICRF), Cuttlefish Feature Selection (CFS).

1. Introduction

Selecting a feature subset keeping in view the target concepts, feature dimensionality reduction has become indispensable component and is an effective way for increasing prediction performance in terms of learning precision, simplicity and speed helping data visualization and result comprehensibility[1][2]. The problem of selecting the optimal number of features come across many fields like data mining, machine learning and pattern recognition, where a large number of features are involved [3][4]. It is observed that those feature conceptually noisy, redundant Perrier relevant to the undertaken problem requires a substantial amount of memory and often degrade the performance of the learning algorithms. So, integration conceptually redundant features and eliminating irrelevant features to reduce the feature dimensional in to a reduced manageable size with as less informational loss as possible is a fundamental target for any real world applications [7].
Many works have been proposed by various researchers in this direction in the past. Among them, Haider and Suresh [1] proposed a Hamming distance based binary PSO algorithm for feature selection and classification in gene expression data. Kapilet al. [2] proposed a CRF based feature selection for effective decision making. Ganapathy et al. [3] introduced an Intelligent CRF based feature selection method for effective decision making. Adel et al. [4] presented a new cuttlefish based optimization algorithm for feature selection used for effective decision making over KDD cup dataset. It selects the best quality of features from large volume of data and removes the irrelevant features. Jin et al [5] developed a new feature selection algorithm able to identify informative features that reveal intrinsic structures of a high dimensional space. Qi and Ivor [6] proposed a generalized sparse regularizer for feature selection, and also a unified feature selection framework for general loss functions. Despite the various feature selection algorithms proposed, it is hard to nominate a single best feature selection algorithm which can be termed as the “best” feature selection algorithm. In terms of time and performance more efforts are required in this direction. To overcome these gaps this paper proposes a new cuttlefish algorithm using ICRF for effective feature selection to improve the classification accuracy in medical diagnosis system.

Contributions of this paper are the use of Intelligent CRF and Enhanced Multiclass Support Vector Machine (EMSVM) for effective feature selection. The remainder of the paper is organized as follows: Section 2 explains in detail the proposed work. Section 3 provides the results and discussion. Finally, Section 4 gives conclusion and future works.

2. Proposed Work

In this paper, an Intelligent CRF [3] based Cuttlefish Feature Selection Algorithm according to [3][4][7] for effective feature selection over the Diabetes and Heart disease datasets is proposed. The best features are selected by applying the multistage evaluation process based on the removal of worst features from subsets in every stage. ICRFCFA reorders all the possible cases which are used in existing cuttlefish algorithm. The formulation for finding the new solution (NS) using reflection (R) and visibility (V) is described in Equation (1).

\[ NS = R + V \]  \hspace{1cm} (1)

ICRFCFA uses the two processes reflection and visibility to find a new solution. The cases work as a global search with the help of each point for finding a new area around the best solution with a specific interval. The creations of these processes are described in Equations (2) and (3), respectively.

\[ R_j = DR \times C_i[i].Points[j] \]  \hspace{1cm} (2)

\[ V_j = VD \times (B.Points[j] - C_i[i].Points[j]) \]  \hspace{1cm} (3)
where, \( C_1 \) is a group of cells, \( i \) is the \( i^{th} \) cell in \( C_1 \), \( \text{Points}[j] \) represents the \( j^{th} \) point of the \( i^{th} \) cell, best (B). Points represent the best solution points, \( \text{DR} \) represents the degree of reflection, and \( \text{VD} \) represents the visibility degree of the final view of the pattern. \( \text{DR} \) and \( \text{VD} \) are found as follows:

\[
\text{DR} = \text{ICRF}() \times (dr_1 - dr_2) + dr_2 \quad (4)
\]

\[
\text{VD} = \text{ICRF}() \times (vd_1 - vd_2) + vd_2 \quad (5)
\]

Where, the ICRF() function is used to generate random numbers between \((0,1)\) based on conditions and \( dr_1, dr_2, vd_1, vd_2 \) are constants defined by the user. For local search, ICRFCFA [5] is implemented to find the difference between the best solution and the current solution for producing an interval around the best solution as a new search area in this work. The formula for finding the reflection is as follows:

\[
R_j = \text{DR} \times \text{B.Point}[j] \quad (6)
\]

While the formulation for finding the visibility remains. The proposed algorithm also uses this case as a local search, but now the difference in average value and the Best (B) point is used for producing a small area around the best solution as a new area for search. The formulas to find the reflection and visibility in this case are as follows:

\[
R_j = \text{DR} \times \text{B.Points}[j] \quad (7)
\]

\[
V_j = \text{VD} \times (\text{B.Points}[j] - \text{AV}_B) \quad (8)
\]

Where, \( \text{AV}_B \) is the average value of the Best points. Finally, the ICRFCFA provides the selected features for effective decision making.

**Intelligent CRF based Cuttlefish Feature Selection Algorithm**

**Input:** Datasets

**Output:** Selected Features

**Step 1:** Initialize the population (features) with random subset.

**Step 2:** Evaluate fitness of the population using EMSVM.

**Step 3:** Store the best subset in B.

**Step 4:** Remove one feature from B using ICRF [3].

**Step 5:** Sort the original features in descending order based on the fitness value which is calculated according to [4].

**Step 6:** Randomly selected features is split into two and store into a set.
Step 7: Find the Reflection subset from randomly selected set using ICRF.

Step 8: Find the Visibility set for removing the elements of R using ICRF.

Step 9: New subset is created by using the features of visibility and reflection.

Step 10: Evaluate the new subset using EMSVM.

Step 11: If the new subset is better than the set B then the current new subset is considered as B.

ICRFCFA selects the best subset from resulted features which are finalized by the evaluation process using fitness function. Ascending the feature subsets and selecting features using Intelligent CRF in different form such as Reflection and Visibility. This algorithm is useful for selecting optimal features which are useful for making effective decision over the dataset.

3. Results and Discussion

This section provides the explanation about the evaluation criteria, dataset used and the experimental results.

Evaluation criteria: In this subsection the performance measures are discussed. The proposed model is evaluated based on three performance measures: The Detection Rate (DR), False Positive Rate (FPR), and Accuracy Rate (AR) [4].

\[ DR = \frac{S}{T} \times 100\% \] (9)

Where, S is number of diseases affected records that are correctly classified as disease and TA is the total number of disease affected records in the test dataset.

\[ FPR = \frac{L}{TN} \times 100\% \] (10)

Here, L is the number of normal that are incorrectly classified as disease and TN is the total number of normal in the test dataset.

\[ AR = \frac{C}{TI} \times 100\% \] (11)

Where, C is the number of correctly classified as instances and TI is the total number of instances in the test dataset.

Higher values of DR and AR, and lower values of FPR show better classification performance for medical diagnosis.

3.1 Dataset

Diabetes and Heart disease datasets are collected from the UCI Repository machine learning [9] for training and 10000 patient records for testing which are collected from medical experts and cancer institutes. In this work, Cleveland heart disease dataset is used, it contains 76 features but researchers identified only 14 features are important. This dataset...
contains 303 records. Diabetes files consist of four fields per record such as Date, Time, Code and Value [9]. This dataset contains 70 text files with full medical information of every patient.

### 3.1 Experimental Results

The experiments were carried out using Java on Core i3 CPU 1.90 GHz laptop with 4 GB RAM. Table 1 shows the performance of various stages in the proposed feature selection algorithm.

**Table 1: List of features selected in various stages of the ICRFCFA.**

<table>
<thead>
<tr>
<th>Feature Selection Stages</th>
<th>Heart Disease Dataset</th>
<th>Diabetes Dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full Features (Before Feature Selection)</td>
<td>Age, Sex, Cp, Trestbps, Chol, Fbs, Restecg, Thalach, Exang, Oldpeak, Slope, Ca, Thal, Class</td>
<td>33,34,35,48,57,58,59,60,61,62,63,64, 65,66,67,68,69,70,71,72</td>
</tr>
<tr>
<td>Stage 1</td>
<td>Age, Cp, Trestbps, Chol, Fbs, Restecg, Thalach, Exang, Slope, Thal</td>
<td>33,34,35,48,57,58,59,60,61,62,63,64</td>
</tr>
<tr>
<td>Stage 2</td>
<td>Age, Cp, Trestbps, Chol, Fbs, Restecg, Slope</td>
<td>33,34,58,59,60,61,62,63</td>
</tr>
<tr>
<td>Stage 3</td>
<td>Age, Cp, Restecg, Slope</td>
<td>33,34,58,59</td>
</tr>
<tr>
<td><strong>Selected Features</strong></td>
<td><strong>Age, Cp, Restecg, Slope</strong></td>
<td><strong>33,34,58,59</strong></td>
</tr>
</tbody>
</table>

Figure 1 shows the false positive rate analysis of the proposed ICRFCFA for the different number of features.

![False Positive Rate Analysis](image)

**Figure 1: False Positive Rate Analysis for various features selected by ICRFCFA.**

From figure 1, it can be observed that the performance of the less number of features used is better than other number of features selected by ICRFCFA. It is because of the fact that the less false positive rate is the selection accuracy.
Figure 2 shows the accuracy rate analysis for the different number of features selected by the proposed ICRFCFA.

**Figure 2: Accuracy Rate Analysis for various features selected by ICRFCFA.**

From figure 2, it can be observed that the accuracy rate is high when medium number of features is selected by ICRFCFA. This is due to the fact that the exact features selection.

Figure 3 shows the detection accuracy comparative analysis of the proposed ICRFCFA, CFA [4], and IASA with EMSVM [8].

**Figure 3: Performance Comparative Analysis for ICRFCFA, CFA and IASA+EMSVM.**

Figure 3 reveals that the detection rate of the proposed feature selection algorithm is better than the existing algorithms such as CFA [4] and IASA with EMSVM [8]. The reason for the significant performance improvement in this paper is the use of Intelligent CRF for the feature selection and removal.

### 4. Conclusion and Future Enhancements

An Intelligent CRF based cuttlefish algorithm (ICRFCFA) is proposed and implemented for effective feature selection in this paper to improve the performance of the medical expert system. The proposed algorithm is used to select the best and the worst features, and then with the application of cuttlefish optimization approach for finalizing the necessary features...
for effective decision making. Future works in this direction could be the introduction of new rules for effective feature selection.

References


Corresponding Author:
Gotam Singh Lalotra*
Email: singh.gotam@gmail.com