# Chapter 5

## AN EFFICIENT APPROACH FOR PREDICTION OF NUCLEAR RECEPTORS AND THEIR SUBFAMILIES

Nuclear receptors, a class of protein found within the cells are responsible for sensing steroid, thyroid hormones and other molecules. Nuclear receptor acts as a ligand activated transcription factor and have the ability to regulate gene expression by integrating with DNA sequences to control the development, metabolism, reproduction and homeostasis thus nuclear receptors are a potential drug target for the various diseases such as diabetes, cancer, osteoporosis and inflammatory diseases (Robinson *et al.*, 2003; Moore *et al.*, 2006).The functions of the nuclear receptor are closely related to these diseases, so for analyzing the cellular mechanism related to physiology and pathophysiology, it is necessary to know about the function and structure of the nuclear receptor. So the main objective is to develop a novel approach for the prediction of families and subfamilies of nuclear receptor to develop a new drug and therapeutic mechanism for the diseases.

There are various computational methods are available in literature to predict the nuclear receptors and their subfamilies. All these proposed methods have their merits and demerits but play a significant role in the development of the prediction for nuclear receptor and their subfamilies. But the main drawback is that all the proposed methods use very small protein samples to predict the nuclear receptors and their subfamilies. Here, the dataset is constructed from latest version of NucleaRDB (Vroling *et al.*, 2011) and more features are extracted from a protein samples. In this chapter, the fuzzy k-nearest neighbor is proposed to use with minimum redundancy maximum relevance (MRMR) (Peng *et al.*, 2005) feature selection algorithm for selecting the optimal features to enhance the performance of the classifier for the prediction of nuclear receptors and their subfamilies.

## 5.1. Background

Here, the sequence of nuclear receptors with their properties and the proposed methods and model that are used to predict nuclear receptors and their subfamilies are presented.

### 5.1.1. Material and methods:

In this chapter, the protein sequences of nuclear receptors are collected from the latest version of NucleaRDB. The description of datasets collected for all eight subfamilies: NR1: Thyroid hormone like (TR, RAR, ROR, PPAR, VDR), NR2: HNF4-like (HNF4, RXR, TLL, COUP, USP), NR3: Estrogen like (ER, ERR, GR, MR, PR, AR), NR4: Nerve Growth factor IB-like (NGFIB, NURR), NR5: Fushi tarazu-F1 like (SF1, FTF, FTZ-F1), NR6: Germ cell nuclear factor like (GCNF1), NR0A: Knirps like (KNI, KNRL, EGON, ODR7), and NR0B: DAX like (DAX, SHP) are shown in Table 5.1.

**Table 5.1: Number of sequences belonging to each nuclear receptor and their subfamilies**

| Family | Subfamilies of NR | Number of sequence |
|---|---|---|
| NR | NR1 | 807 |
| | NR2 | 719 |
| | NR3 | 669 |
| | NR4 | 102 |
| | NR5 | 136 |
| | NR6 | 38 |
| | NR0A | 30 |
| | NR0B | 42 |
| Non-NR | N/A | 1198 |

### 5.1.2. Features extraction of protein sequences

In this chapter, to fully characterize protein sequence seven feature vectors are extracted from PROFEAT server (Rao *et al.*, 2011) to represent the protein sample, including amino acid composition, dipeptide composition, correlation, composition, transition, distribution of physiochemical properties and pseudo amino acid composition with total of 857 number of features being calculated for the prediction of nuclear receptors and their eight subfamilies.

The description of total 857 number of features derived from sequences of nuclear receptors is shown in Table 5.2.

**Table 5.2: Description of sequence derived features of nuclear receptors**

| S. No. | Features of protein sequences | Total No. of features | Description |
|---|---|---|---|
| 1 | $X_1$ to $X_{20}$ | 20 | Amino acid composition |
| 2 | $X_{21}$ to $X_{420}$ | 400 | Dipeptide composition |
| 3 | $X_{421}$ to $X_{660}$ | 240 | Correlation factors |
| 4 | $X_{661}$ to $X_{681}$ | 21 | Composition |
| 5 | $X_{682}$ to $X_{702}$ | 21 | Transition |
| 6 | $X_{703}$ to $X_{807}$ | 105 | Distribution of physiochemical properties |
| 7 | $X_{808}$ to $X_{857}$ | 50 | Pseudo amino acid composition |

## 5.2. Proposed method and model

Here, a fuzzy k-nearest neighbor classifier is proposed to use with minimum redundancy maximum relevance (MRMR) feature selection algorithm. The minimum redundancy maximum relevance is proposed for selecting the optimal features to enhance the performance of the classifier for the prediction of nuclear receptors and their subfamilies.

### 5.2.1. Feature subset selection

Fuzzy k nearest neighbor is a powerful machine learning method but it cannot perform automatic feature selection. To reduce this limitation various filter and wrapper based feature selection methods are used to select feature subset. Filter methods are computationally simple and easy to handle high-dimensional dataset but they ignore the interaction between selected feature and classifier while wrapper methods include the interaction and may also take correlation between features, but they have a higher risk of over fitting than filter methods. In this chapter, the minimum redundancy maximum relevance (MRMR) (Peng *et al.*, 2005) a filter method is used to select a feature subset. The brief description of MRMR feature selection is as follows:

The minimum redundancy maximum relevance feature selection method select a feature subset in which each subset of feature has the minimal redundancy with other features and maximal relevance with target class. In this method the subset of features is obtained by calculating the mutual information between the features themselves and between the features and the class variables. For binary classification the class variable $c_k$ is 1 or 2. The mutual information *MI(x,y)* of two features x and y is calculated as

$$MI(x,y) = \sum_{i,j \in N} p(x_i \, , y_j) log \frac{p(x_i, y_j)}{p(x_i) \, p(y_j)} \tag{5.1}$$

where $p(x_i)$ and $p(y_j)$ is the marginal probability density and $p(x_i, y_j)$ is the joint probability. Similarly, the mutual information *MI(x, c)* of between class variable *c* and feature *x* is also calculated as

$$MI(x,c) = \sum_{i,k \in N} p(x_i \, , c_k) log \frac{p(x_i, c_k)}{p(x_i) \, p(c_k)} \tag{5.2}$$

The minimum redundancy condition is to minimize the total redundancy of all features selected by Min (Redundancy) where

$$Redundancy = \frac{1}{|S|^2} \sum_{x,y \in S} MI(x,y) \tag{5.3}$$

where *S* is the feature subset and |*S*| is the number of feature in *S*.

The maximum relevance condition is to maximize the total relevance between all features in *S* and class variable. It is calculated as Max (Relevance) where

$$Relevance = \frac{1}{|S|}\sum_{x \in S} MI(x, c) \tag{5.4}$$

Here, first feature having the highest *MI(x, c)* is selected according to equation (5.4) and the rest of the features are selected in incremental way where earlier selected features are remains in the features set. The optimal subset of features is selected by optimizing the equations (5.3) and (5.4) simultaneously through mutual information difference criterion.

$$Max\ (Relevance - Redundancy) \tag{5.5}$$

### 5.2.2. Classification of nuclear receptors and their subfamilies

For the classification of nuclear receptors and their subfamilies, here a fuzzy-k-nearest neighbor based classifier is proposed to be used. The k-nearest neighbor classifier (Cover and Hart, 1967) gives equal importance to all feature vectors including class representative in assigning class membership. It is based on finding the *k* nearest neighbor, and taking a majority vote among the classes of these *k* neighbors, to assign a class for the test sample but there is no indication about the strength of membership. Therefore, in this chapter a generalized nearest neighbor i.e. Fuzzy k-nearest neighbor method (Keller, 1985) is used. The fuzzy k-nearest neighbor classifier calculate the membership values that indicate how much degree the test sample belongs to the particular class, instead of roughly assigning the label based on a voting from the k-nearest neighbors. Fuzzy principle is very useful because it is impossible for any features to contain the complete information.

Let $X_1, X_2, \ldots., X_N$ be a set of features representing *N* proteins in the training set which has been classified into *M* classes: $C_1, C_2, \ldots., C_M$ , where $C_i$ denotes the i$^{th}$ class ,*Y* is a test sample protein, *k* is the number of the nearest neighbors and m greater than 1 is the fuzzy coefficient.

**Fuzzy-k-NN (X, C, Y, k)**

```
{
    1. $k$ ← get nearest neighbor($Y$, $k$)
    2. Select the value of fuzzy coefficient '$m$' for the fuzzy k-nearest
       neighbor
    3. for i=1 to size (test data)
      {
    4. for j=1 to size (train data)
            {
                i. Calculate $d(Y, X_j)$ is the Euclidean distance between the
                   test protein sample '$Y$ ' and its $j^{th}$ nearest protein $X_j$ in the
                   training dataset.
               ii. Calculate the weight $d(Y, X_j)^{-2/(m-1)}$
              iii. Calculate the sum of weight for k nearest neighbor
                   $\sum_{j=1}^{k} d(Y, X_j)^{-2/(m-1)}$
               iv. Calculate the fuzzy membership value of test sample
                   protein $Y$ for the $i^{th}$ class $\mu_i(Y)$
                   $\mu_i(Y) = \frac{\sum_{j=1}^{k} \mu_i(X_j) d(Y, X_j)^{-2/(m-1)}}{\sum_{j=1}^{k} d(Y, X_j)^{-2/(m-1)}}$
                   where $\mu_i(X_j)$ is the fuzzy membership value of the
                   protein $X_j$ to the $i^{th}$ class it is set to 1 if $X_j \in C_i$ , otherwise,
                   0.
                v. After calculating all the memberships for a test protein, it
                   is assigned to the class with which it has the highest
                   membership value so the predicted class for the test
                   protein '$Y$' as
                        Predicted class = argmax $_{i \in C}$ $(\mu_i(Y))$.
            }
      }
}
```

## 5.3. Result and performance analysis

Here, the fuzzy k-nearest neighbor is proposed to use with minimum redundancy maximum relevance feature selection algorithm for selecting the optimal features to enhance the performance of the classifier for the prediction of nuclear receptors and their subfamilies.

### 5.3.1. Performance measures

In this chapter, 10-fold cross validation is used to measure the performance of fuzzy k-NN classifier. In *K*-fold cross validation the dataset of all proteins is partitioned into *K* subsets where one subset is used for validation and remaining *K-1* subsets is used for training. This process is repeated for *K* times so that every subset is used once as a test data. In this chapter, accuracy (ACC) and Matthew's correlation coefficient (MCC) are used to measure the performance of the proposed method.

### 5.3.2. Result analysis

Before measuring the performance of fuzzy k-NN classifier the value of two parameter k and m are determined by using all 857 sequence derived features for the prediction of nuclear receptors and their subfamilies. In this chapter,10-fold cross validation is used to select the optimal value of k and m and it is found that the highest accuracy and MCC is obtained at k=7 and m= 1.25. The performance analysis of classifier with different values of k and m is shown in Figure 5.1 and 5.2.

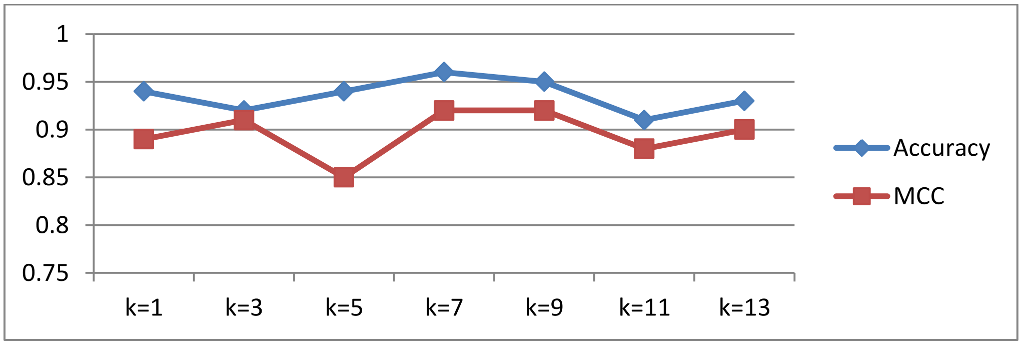

**Figure 5.1: Accuracy and MCC for different values of k**

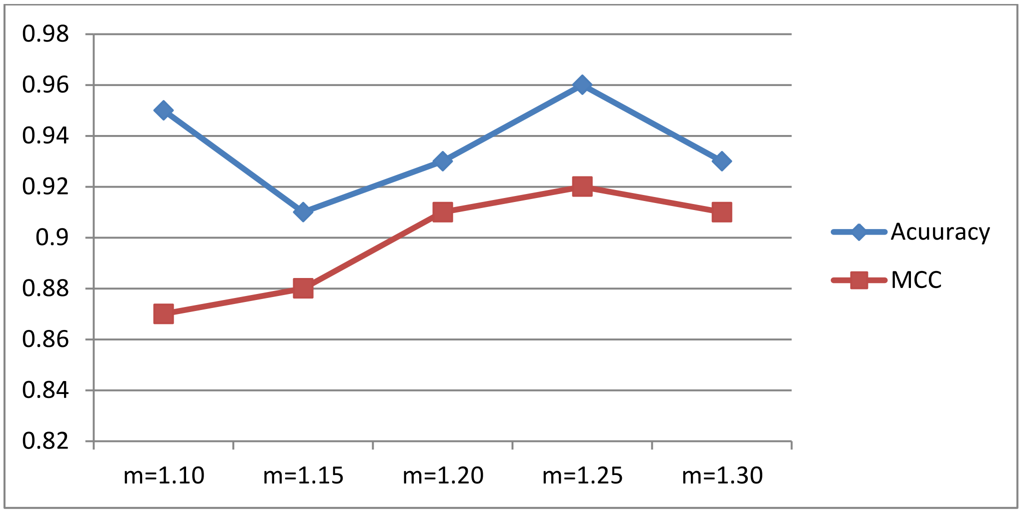

**Figure 5.2: Accuracy and MCC for different values of m**

After determining the values of k and m the performance of the classifier is evaluated with different combination of feature vectors and it is observed that the performance of the classifier is continuously improved by using the amino acid composition, amino acid with dipeptide composition and amino acid with dipeptide composition and correlation feature respectively. Composition, transition, distribution and pseudo amino acid composition feature vector are not affecting the performance for the prediction of nuclear receptors but have an importance for the prediction of nuclear receptors subfamilies and also in the selection of optimal number of features by MRMR feature selection algorithms. The analysis of result is shown in Table 5.3 and 5.4.

**Table 5.3: Result analysis for classification of nuclear and non-nuclear receptors with different data sets**

| Family | AAC | | AAC+DC | | AAC+DC+CF | | AAC+DC+CF+ CTD | | AAC+DC+CF+ CTD+PAAC | |
|---|---|---|---|---|---|---|---|---|---|---|
| | ACC | MCC | ACC | MCC | ACC | MCC | ACC | MCC | ACC | MCC |
| NR | 99.54 | 0.96 | 100.0 | 0.96 | 100.0 | 0.98 | 100.00 | 0.97 | 100.0 | 0.97 |
| Non-NR | 95.42 | 0.96 | 95.08 | 0.96 | 97.04 | 0.98 | 96.46 | 0.97 | 96.46 | 0.97 |
| Overall | 97.48 | 0.96 | 97.54 | 0.96 | 98.52 | 0.98 | 98.23 | 0.97 | 98.23 | 0.97 |

**Table 5.4: Results analysis for the classification of nuclear receptors subfamilies with different datasets**

| Subfamily | AAC | | AAC+DC | | AAC+DC+CF | | AAC+DC+CF +CTD | | AAC+DC+CF +CTD+PAAC | |
|---|---|---|---|---|---|---|---|---|---|---|
| | ACC | MCC | ACC | MCC | ACC | MCC | ACC | MCC | ACC | MCC |
| NR1 | 87.80 | 0.88 | 92.31 | 0.89 | 93.06 | 0.92 | 93.35 | 0.91 | 96.39 | 0.94 |
| NR2 | 75.00 | 0.78 | 93.24 | 0.93 | 93.67 | 0.91 | 94.74 | 0.89 | 96.43 | 0.87 |
| NR3 | 86.36 | 0.74 | 90.63 | 0.86 | 94.52 | 0.89 | 92.21 | 0.88 | 89.74 | 0.85 |
| NR4 | 81.25 | 0.86 | 85.72 | 0.92 | 88.89 | 0.91 | 100 | 0.84 | 100 | 0.84 |
| NR5 | 72.73 | 0.75 | 87.50 | 0.93 | 95.24 | 0.97 | 90 | 0.78 | 92.86 | 0.89 |
| NR6 | 100 | 0.70 | 100 | 0.81 | 100 | 0.86 | 100 | 1.00 | 100 | 1.00 |
| NR0A | 80 | 0.89 | 100 | 1.00 | 100 | 1.00 | 100 | 1.00 | 100 | 1.00 |
| NR0B | 100 | 0.80 | 100 | 0.92 | 100 | 1.00 | 100 | 1.00 | 100 | 1.00 |
| Overall | 85.39 | 0.80 | 93.68 | 0.91 | 95.67 | 0.93 | 96.29 | 0.91 | 96.93 | 0.92 |

After evaluating the datasets minimum redundancy maximum relevance (MRMR) algorithm is used for selecting the optimal features for the classifier to predict the nuclear receptors and their subfamilies. The accuracy and MCC are evaluated for different number of features and it is observed that classifier performance is better for 400 features for prediction of families and subfamilies of nuclear receptor. The analysis of result is shown in Table 5.5 and 5.6.

**Table 5.5: Result analysis for selecting the optimal number of features among all the features for classification of nuclear and non-nuclear receptors**

| Family | NF=250 | | NF=300 | | NF=350 | | NF=400 | |
|---|---|---|---|---|---|---|---|---|
| | ACC | MCC | ACC | MCC | ACC | MCC | ACC | MCC |
| NR | 99.16 | 0.92 | 100 | 0.97 | 100 | 0.97 | 100 | 0.97 |
| Non-NR | 91.18 | 0.92 | 95.61 | 0.97 | 95.87 | 0.97 | 96.18 | 0.97 |
| Overall | 95.17 | 0.92 | 97.80 | 0.97 | 97.93 | 0.97 | 98.09 | 0.97 |

**Table 5.6: Result analysis of optimal feature selection for the classification of subfamilies of nuclear receptors**

| **Sub-families** | | **NR1** | **NR2** | **NR3** | **NR4** | **NR5** | **NR6** | **NR0A** | **NR0B** | **Overall** |
|---|---|---|---|---|---|---|---|---|---|---|
| NF=50 | **ACC** | 86.08 | 86.15 | 91.07 | 84.42 | 81.25 | 100 | 100 | 83.33 | 89.04 |
| | **MCC** | 0.83 | 0.85 | 0.86 | 0.78 | 0.8 | 0.7 | 1 | 0.91 | 0.84 |
| NF=100 | **ACC** | 88.68 | 98.39 | 88.89 | 83.33 | 92.56 | 100 | 100 | 80 | 91.48 |
| | **MCC** | 0.84 | 0.91 | 0.82 | 0.77 | 0.56 | 1 | 1 | 0.89 | 0.85 |
| NF=150 | **ACC** | 94.74 | 96.92 | 90.63 | 100 | 90 | 100 | 100 | 100 | 96.54 |
| | **MCC** | 0.93 | 0.89 | 0.86 | 0.88 | 0.81 | 1 | 1 | 0.92 | 0.91 |
| NF=200 | **ACC** | 90 | 93.94 | 92.75 | 90 | 88.89 | 100 | 100 | 100 | 94.45 |
| | **MCC** | 0.85 | 0.83 | 0.87 | 0.75 | 0.84 | 1 | 1 | 0.91 | 0.88 |
| NF=250 | **ACC** | 87.91 | 93.22 | 92.73 | 90.91 | 88.24 | 100 | 100 | 100 | 94.13 |
| | **MCC** | 0.88 | 0.89 | 0.88 | 0.77 | 0.9 | 0.5 | 1 | 0.44 | 0.78 |
| NF=300 | **ACC** | 88.75 | 98.39 | 87.88 | 91.67 | 100 | 100 | 100 | 100 | 95.84 |
| | **MCC** | 0.85 | 0.91 | 0.84 | 0.91 | 1 | 1 | 1 | 0.89 | 0.93 |
| NF=350 | **ACC** | 88.61 | 93.22 | 88.14 | 83.33 | 93.75 | 100 | 100 | 100 | 93.38 |
| | **MCC** | 0.87 | 0.89 | 0.86 | 0.73 | 0.93 | 0.7 | 1 | 0.89 | 0.86 |
| NF=400 | **ACC** | **94.12** | **98.36** | **97.01** | **100** | **93.33** | **100** | **100** | **100** | **97.85** |
| | **MCC** | **0.93** | **0.9** | **0.93** | **0.95** | **0.9** | **0.7** | **1** | **0.91** | **0.9** |

### 5.3.3. Comparative Analysis

In this chapter, the performance of the classifier is evaluated at 400 number of optimal features obtained from MRMR feature selection algorithms and the result is compared with the previous three approaches for the prediction of nuclear receptors and their seven subfamilies. After that we compared our result with recently proposed approach NRpred-FS (Wang *et al.*, 2014) that are used eight subfamilies in their datasets. It is observed that the method proposed in this chapter improve the performance for the prediction of subfamilies of nuclear receptors. The comparative analysis is shown in Table 5.7, 5.8 and 5.9.

**Table 5.7: Result comparison for the prediction of NR and non-NR among existing methods and proposed method**

| Family | NRpred-FS (Wang *et al.*, 2014) | | iNR-Physchem (Xiao. *et al.*, 2012) | | NR-2L (Wang *et al.*, 2011) | | Proposed Method | |
|---|---|---|---|---|---|---|---|---|
| | ACC | MCC | ACC | MCC | ACC | MCC | ACC | MCC |
| NR | 96.23 | 0.97 | 96.23 | 0.95 | 98.11 | 0.83 | 100 | 0.97 |
| Non-NR | 99.60 | 0.97 | 98.80 | 0.95 | 90.80 | 0.83 | 96.18 | 0.97 |
| Overall | 98.79 | 0.97 | 98.18 | 0.96 | 92.56 | 0.85 | 98.09 | 0.97 |

**Table 5.8: Result comparison for the prediction of seven subfamilies among existing methods and proposed method**

| Subfamily | NRpred-FS (Wang *et al.*, 2014) | | iNR-Physchem (Xiao. *et al.*, 2012) | | NR-2L (Wang *et al.*, 2011) | | Proposed Method | |
|---|---|---|---|---|---|---|---|---|
| | ACC | MCC | ACC | MCC | ACC | MCC | ACC | MCC |
| NR1 | 96.34 | 0.90 | 94.00 | 0.87 | 86.00 | 0.88 | 94.12 | 0.93 |
| NR2 | 92.65 | 0.91 | 97.22 | 0.93 | 86.11 | 0.85 | 98.36 | 0.90 |
| NR3 | 84.85 | 0.84 | 100 | 0.95 | 100 | 0.86 | 97.01 | 0.93 |
| NR4 | 90.91 | 0.95 | 71.43 | 0.84 | 85.71 | 0.7 | 100 | 0.95 |
| NR5 | 80.00 | 0.89 | 83.33 | 0.91 | 83.33 | 0.86 | 93.33 | 0.90 |
| NR6 | 80.00 | 0.89 | 100 | 1.00 | 100 | 1.00 | 100 | 0.70 |
| NR0 | 100 | 0.96 | 66.67 | 0.81 | 75.00 | 0.86 | 100 | 1.00 |
| Overall | 90.59 | 0.92 | 88.13 | 0.90 | 88.10 | 0.85 | 97.85 | 0.90 |

**Table 5.9: Result comparison for the prediction of eight subfamilies NRpred-FS and proposed method**

| Subfamily | NRpred-FS (Wang et al., 2014) | | Proposed Method | |
|---|---|---|---|---|
| | ACC | MCC | ACC | MCC |
| NR1 | 96.34 | 0.9.0 | 94.12 | 0.93 |
| NR2 | 92.65 | 0.91 | 98.36 | 0.90 |
| NR3 | 84.85 | 0.84 | 97.01 | 0.93 |
| NR4 | 90.91 | 0.95 | 100.00 | 0.95 |
| NR5 | 80.00 | 0.89 | 93.33 | 0.90 |
| NR6 | 80.00 | 0.89 | 100.00 | 0.70 |
| NR0A | 100 | 0.96 | 100.00 | 1.00 |
| NR0B | 100 | 1.00 | 100.00 | 0.91 |
| Overall | 90.59 | 0.92 | 97.85 | 0.90 |

## 5.4. Conclusion

In this chapter, a fuzzy k- nearest neighbor classifier with minimum redundancy maximum relevance have introduced for the classification of nuclear receptors and their eight subfamilies. Seven feature vectors are used to represent the protein sample, including amino acid composition, dipeptide composition, correlation, composition, transition, and pseudo amino acid composition. To improve the prediction performance the minimum redundancy maximum relevance algorithm has used to select the optimal feature subset. The values of two parameter k and m have been determined to predict nuclear receptors and their subfamilies and it is observed that the highest accuracy and MCC is obtained at k=7 and m= 1.25. The overall accuracies of 10-fold cross validation with optimal number of features; k and m are 98.09% and 97.85% and MCC values of 0.97 and 0.90 for the prediction of nuclear receptors and their subfamilies respectively. From the obtained results and analysis it was observed that the performance of the proposed approach for the classification of nuclear receptors and their eight subfamilies is performing better than some other standard methods available in literature.