A new algorithm relief hybrid (HRelief) for biological motifs selection

Faouzi Mhamdi and Hanen Mhamdi

***Abstract*— Feature selection plays a crucial role in the automatic learning field, since the non relevant and /or redundant ones can influence the strength of discrimination of a learning algorithm. In fact, select a minimum set of informative and relevant features can increase the performance of algorithms and the precision of prediction, minimize the time of data treatment, facilitates their visualization as well as their analysis. In this paper, we present a series of adaptations of algorithms for the motifs selection of Relief filtering algorithm. In the first two adaptation ways (HRelief1 and HRelief2) we transformed Relief in hybrid algorithms by using a classifier to evaluate the subset of the features generated. The third way of adaptation (HRelief3) helps in treating the problem of redundancy of features. Based on the experimentations done so far, these improvements resulted in an interesting outcome that encourages us to go into the depth of this orientation field.**

I. INTRODUCTION

During the last decade, the motivation for dimension reduction has shifted from being an illustrative example to becoming a real prerequisite for model building. As the dimensionality of the data increases, many types of data analysis and classification problems become significantly harder. The bioinformatics application domain[1] is one in particular where datasets often have a very large number of descriptors. Compounding the problem, these datasets often have relatively few examples,therefore, selecting the most informative features[2] is a decisive preprocessing step in the knowledge discovery in biological databases process. The objective of motif selection is to reduce the number of motifs in the dataset such that the selected motifs incorporate as much information from the entire dataset as possible. Reducing the number of motifs can have numerous positive implications, such as eliminating redundant or irrelevant motifs, decreasing development time and improving the performance of learning models. In this paper we focus on the problem of motif selection for protein classification[3]. The main objective is to optimize the performance of the classifier on test data and avoid over fitting.

There are three dimensions to categorize selection methods: Search strategy (Complete, Heuristic and Random), evaluation criterion (Filter, Wrapper and Hybrid) and data mining techniques (classification, clustering, prediction...). There are four basic steps in any motifs selection method: *(i)*Generation: to select candidate motif subset. *(ii)* Evaluation: to evaluate the generated candidate motif subset and output a relevancy value. *(iii)* Stopping criteria will determine whether it is the defined optimal motif subset and *(iv)* the last step is validation.

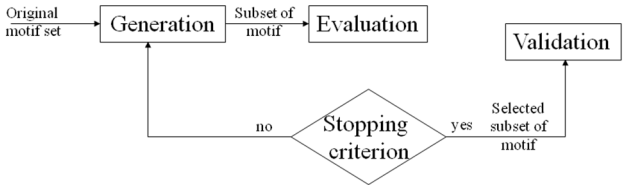

Fig. 1. Four main steps in a motif selection method

Generally speaking, motif selection techniques are often classified in three categories depending on how they combine the motif selection search with the construction of the learning model: filters, wrappers and embedded methods.

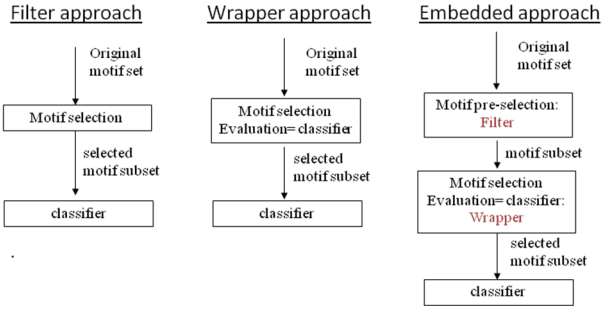

Fig. 2. Motifs Selection categories for Classification.

Among the features selection methods that are mostly used recently we find the Feature Ranking Search (FRS) methods[4]. After calculating the individual utility of features, we select the best ones that give a utility outreaching a certain value degree. Among the methods of features selection that use this principle we find the benchmark of the features weight "weighting method". In this realm, we talk about a problem of estimation of features for which several techniques have been proposed in the recent years. Relief algorithm is one of the most common due to its simplicity and effectiveness [5] to measure the feature relevance. This represents a technique of estimation which detects the features mostly significant correlated to the family predicted.

II. OVERVIEW ABOUT USED BIOLOGICAL DATA

The processing of knowledge discovery for protein classification is achieved in two steps: motif extraction and motif selection. The first step is the extraction of features from the original description in order to build a motif-value table which is useful for ulterior data mining techniques.

This work was not supported by any organization

F. Mhamdi is with the Research Laboratory of Technologies of Information and Communication & Electrical Engineering, ESSTT University ,1008 Tunis, TUNISIA Faouzi.mhamdi@ensi.rnu.tn

H. Mhamdi is with the Research Laboratory of Technologies of Information and Communication & Electrical Engineering,, ESSTT University ,1008 Tunis, TUNISIA mhamdi.hanen@gmail.com

978-1-4799-3163-7/13/$31.00 ©2013 IEEE

However, the use of data in their native profile is not possible, the methods of data mining cannot be applied on biological sequences under their prior profile. So It is necessary to transform this unstructured data in a table individuals/motifs,T with a line i represents a structure of proteins, a column j represents an extract motif, the case T *[i,j]* represents thus the discretion of j motifs in the i former structure,this value indicates the importance of motif in a sequence. The last column represents the families to which the different sequences belong.

Inspired from text mining we have used n-grams(a sequence of n characters)extraction technique in order to produce descriptors[6]. However, unlike the text classification, there is no "natural" separation in the character sequences it is not possible to extract "words" for which we can easily attach semantics properties, thus n-grams technique seemed to be a good compromise in our case. Previous works [7] showed that n=3(3-grams) is a good compromise to produce a minimal motif subset with high discrimination of protein sequence capacity.

In fact, several kids of values can be attributed to a motif, regarding the fact that many studies have been realized in the automatic classification of texts we have tested four kinds of data representation which are as follows :

- Boolean: indicates if an extracted motif is present or not within a sequence, if T [i,j]=1 so the j motif is present in the i sequence otherwise it is absent.
- Occurrences: number of occurrence of a motif within a sequence, T[i,j]=k, k $k \geq 0$, thus the j motif is present k times in the i sequence.
- Frequency: relative frequency of a motif with regard to a number of motifs that composed a sequence.
- TF*IDF: corrects the frequency of motifs with regard to its function and frequency in the file.

The Boolean data representation can give the impression that it's rather rough but several studies in the text mining domain showed their effectiveness.

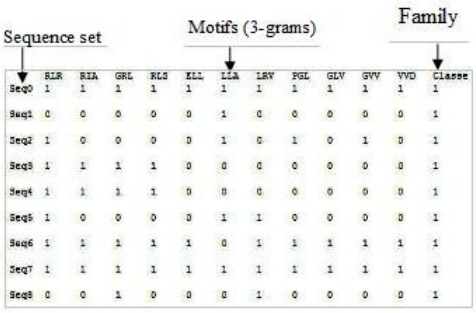

Fig. 3. Learning boolean file.

The used protein families in this work are randomly extracted from the data bank SCOP [8]. A protein sequence is a series of amino acids that have a specific order. There exist 20 amino acids that allow the description of a protein by a sequence of characters. Besides, proteins are grouped into several families according to the functions they perform, and all proteins contained in the same family have similar structures. Thus, by giving a set of proteins with known properties, we have to look for inducing classification rules that associate motifs to protein families (classes).

TABLE I
PROTEIN PRIMARY STRUCTURE FAMILY

	Sequences number	Average length	Total length
f_1	40	884	35382
f_2	47	667	31363
f_3	54	601	32480
f_4	81	505	40907
f_5	68	462	31473

III. RELIEF ALGORITHM

Among existing feature selection techniques, the Relief algorithm[5] is one of the most successful filer techniques, it used as a pre-processor to remove irrelevant features before the learning stage. The success of this algorithm is due to the fact that it's fast, easy to understand and implement and accurate even with dependent motifs and noisy data. The algorithm is based on a simple principle it estimates the quality of motifs according to how well their values distinguish between sequences that are near to each other. For this purpose, we assume two family classification problems given a randomly selected sequence that is represented by a vector composed of n motif values $X_i(x_1,x_2\ ,...,x_j)$ where x_j denote the value of motif m_j of X_i. Relief searches for its two nearest neighbours: one from the same family, called nearest hit H, and the other from a different family, called nearest miss M. It then updates the quality estimate for all the motifs, depending on the values for X_i, M, and H. This process is repeated p times where p is a parameter predefined by the user. Relief use the heuristic generation strategy to select motifs subset candidate and the distance as evaluation function. Relief uses a function *"diff"* to calculate the difference between the values of motifs between these two sequences. The weight W_i is of the quality of the motif m_i. The function that calculates the difference between each couple of instance used by Relief is given by the following equation:

$$diff(m_j,X_1,X_2) = \begin{cases} |X_1 - X_2| \text{ if } m_j \text{ is numerical} \\ 0 \text{ if } m_j \text{ is nominal and } X_1 = X_2 \\ 1 \text{ if } m_j \text{ is nominal and } X_1 \neq X_2 \end{cases} \quad (1)$$

The relevance level of motifs m_j defined by the following equation:

$$W[m_j] = W[m_j] - \frac{diff(m_j,X_i,H)}{p} + \frac{diff(m_j,X_i,M)}{p} \quad (2)$$

Algorithm 1 Relief Algorithm

Require: δ, m, τ
Ensure: ϕ
  ϕ=∅,W(0,0,...,0) //Initialization of selected subset of features and of weight vector
  for $i = 1,...,m$ **do**
    Select randomly an instance R_i
    H=closer-same-class(R,δ)
    M= closer-different class (R,δ)
    for $j = 1,...,n$ **do**
      W_j=W_j-diff$(x_j, H_j)^2$+diff$(x_j, M_j)^2$
    end for
  end for
  for $j = 1,...,n$ **do**
    if $W_j \geq \tau$ **then**
      $\phi \leftarrow \phi \cup a_j$
    end if
  end for

The Relief algorithm is of complexity $O(m * n)$ within calculation time. It can select the relevant features in linear time in function of a number of features n and the number of instances m. Comparative study[7] between diverse filers algorithms enables us to conclude that the algorithms CFS achieved a good balance between the classification rate and the number of selected motifs for the whole data. However, due to the results provided by the algorithm Relief and not neglecting the characteristics of biological data, a big number of features and a weak number of instances. It becomes apparent to access techniques that examine directly the correlation of each feature for which several techniques have been proposed in the recent years, among which we cite the most frequent algorithm Relief.

IV. The different adaptation ways of Relief algorithm

After having presented the principle basics of Relief algorithm, we go back to our contribution which consists of elaborating new algorithms of features selection based on Relief algorithm by adopting a hybrid approach. It is about combining the approach wrapping and the approach filter. The evaluation criteria of wrapping methods are about the rate of classification error. This approach allows selecting the sub space of features in function of classification results. However, what has varied for an algorithm of classification is not necessarily for the other,because the performance obtained by the wrapping methods depending on the algorithm of classification used . On the other hand, the use of such method of evaluation necessitates the classification of each sub space candidates, which is expensive in time calculation, particularly, when the dimension of initial space of feature has increased [9]. In this case, it becomes better to use filter approaches, which are less cheap in time calculation. To respond to the dilemma "classification rate / calculation time" and, consequently, to combine the advantages of these two approaches,we use a hybrid approach. By adopting this type of approach, a function of evaluation is first of all used to preselect the sub-space of features that are most discriminating. Then, the classification error rates of obtained prototypes by considering each sub space discriminated selected are compared in order to determine the final sub space [10]. The hybrid approaches constitute a good alternative to achieve the compromise between classification quality in the space kept and the calculation time necessary for features selection.

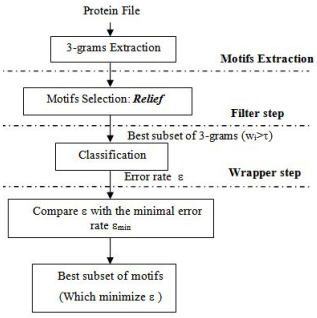

Fig. 4. Hybrid Relief process.

A. first adaptation : HRelief1

If $\Psi_k(f_1, f_2..., f_n)$ a set of features (sub-chains extracted) from chain families f_1, f_2,..., f_n. our first algorithm of features section is a hybrid algorithm in a way that in each iteration the set of selected features is evaluated according to two criteria : the first is an independent criterion (filter) and the second one is a dependant criterion (wrapper). By adopting this algorithm, we operate in the following way: In the first step, we calculate for each sequence the vector weight which has as length the number of features included in the set Ψ_k. After, we select the features in a declining way. According to their dependency degree of the discriminated family fi in order to filter the features and to select those having a positive weight. In the second step, we utilize the wrapper approach by using a classifier in order to calculate the classification error rate ε of the subset of received features during the first stage. The set ϕ_{min} received represents the set of the selected features which possess a minimal error rate ε_{min} regarding the other subset. Our first algorithm of motifs selection (sub-chains extracted) receives the set of features (3-grams extracted), Ψ_k (f_1, f_2,.., f_n), the table of data T and return the optimal subset of selected motifs ϕ_{min}. The classifier used during this wrapper step of our algorithm receive the subset of motifs having a positive weight, represented by the sub-table $T_{\phi p}$, and returns the classification error rate associated ε, defined by:

$$\varepsilon = \left(\frac{\text{total Number of sequence wrong classified}}{\text{total number of sequence}} \right) \quad (3)$$

This rate helps in evaluating the pertinence of the subset of the current features. In fact, we consider that the subset of

features is the best subset, in case the classification error rate associated ε is minimal. The Relief1 algorithm uses the following variables:Ψ_n:Pile representing the set of extracted features and ϕ_p representing the subset of current features.

Algorithm 2 HRelief1

Require: Ψ_n,T
Ensure: ϕ_{min}
  $\varepsilon_{min} = 1$,W(0,0,...,0) //Initialization of weight vector and $\varepsilon_{min}(0 \leq \varepsilon_{min} \leq 1)$
  for $i = 1,...,m$ **do**
    for $j = 1,...,m-1$ **do**
      Calculate for each instance R_i: H et M
      for $s = 1,...,n$ **do**
        $W_s = W_s\text{-}diff(x_i,H_j)^2 + diff(x_i,M_j)^2$
      end for
      QuickSort (Ψ_n, T)
      Classifier ($T_{\phi p}$,ε)//ϕp subset of motifs with positive Weights.
      if $\varepsilon \leq \varepsilon_{min}$ **then**
        $\varepsilon_{min} \leftarrow \varepsilon$
        $\phi_{min} \leftarrow \phi_p$
      end if
    end for
  end for

The complexity of this algorithm is $\max(O(m^2 * O_{class}), O(m^2 * n)))$, where O_{class} is of complexity of calculation time of the used classifier.

B. Second adaptation : HRelief2

Our second algorithm of motifs selection, HRelief2 , receives the subset of features ψ_{nmin} and the data table $T_{\phi min}$ that represents the returned results by the algorithm HRelief1 , and sends back the subset of selected features, ϕ_n represents by the pile ϕ_{min1}. On the other hand, this algorithm calls a classifier. This classifier receives the subset of features, represented by the sub table $T_{\phi n}$,and returns the error rate of the associated classifier ε. This algorithm realizes a features selection, according to a research direction in front (best-first search): we initialize features set to be selected ϕ_n to the empty set. After, during each iteration, we select the feature that possesses the weight the most high, we delete it from ψ_{nmin}, we insert it in ϕ_n and we calculate the new classification error rate. We repeat the process until ψ_{nmin} becomes empty. The set $\phi min1$ obtained consequently, represents the subset of the selected features. We use the following functions:

- insert : function to insert in the pile.
- delete : function to delete a pile.

Algorithm 3 HRelief2

Require: ψ_{nmin},$T_{\phi min}$
Ensure: ϕ_{min1}
  $\varepsilon_{min} = 1$,$\phi_n = \varnothing$
  repeat
    Delete (t,ψ_{nmin})
    Insert(t,ϕ_n)
    Classifier ($T_{\phi n}$,ε)
    if $\varepsilon \leq \varepsilon_{min}$ **then**
      $\varepsilon_{min} \leftarrow \varepsilon$
      $\phi_{min1} \leftarrow \phi_n$
    end if
  until $\psi_{nmin} = \varnothing$

Relief Hybride2 algorithm is of complexity $O(\mid \psi_{nmin} \mid * O_{class})$ with regard to calculation time, where O_{class} is the complexity of the used classifier.

C. Third Adaptation HRelief3: study of features redundancy

The previous algorithms suffer from weakness: it does not take into consideration the redundancy of features. Therefore it is necessary to introduce a new constraint in features selection. For this reason, in our third algorithm we tried to treat the problem of features redundancy, by considering their relevance. This constraint consists of two steps. At first, since we treat the features represented by the 3-grams so we look for if a 3-gram m_i possesses two characters in the three possible identical to m_j. If this is the case we continue to the second step which consists of comparing the weight $w[m_i]$ and $w[m_j]$ of both features m_i and m_j checking the first condition. Consequently, the set of features ψ_n dispensed of the feature that possesses the most weak weight compared to the other feature, is evaluated via a classifier in order to calculate its classification error rate ε. This rate is compared with ε_{min} received with the set of features ϕ_{min1} selected by the algorithm HRelief2. If $\varepsilon$is inferior to ε_{min} so the feature possessing the weight that is mostly weak will be eliminated. Our algorithm receive the subset of features ϕ_{min1} selected by the algorithm HRelief2 with its error rate of classification ε_{min} and the data table $T_{\phi n}$: sub-table of table T of which the columns represent the subset of features ϕ_{min1}. The exit of this algorithm becomes the subset of features ϕ_n without redundant features which minimize the error rate of classification.

Algorithm 4 HRelief3

Require: Φ_{min1},ε_{min},$T_{\phi n}$
Ensure: ϕ_n
  $\phi_n \leftarrow \$\Phi_{min1}, \tau = \frac{2}{3}$
  for $i = 1,...,n$ **do**
    for $j = i+1,...,n-1$ **do**
      if $Compare_motifs\ (m_i, m_j, \tau)$=True **then**
        if $W[m_i] \geq W[m_j]$ **then**
          Classifier ($T_{\phi n}$-m_j,ε)
          if $\varepsilon \leq \varepsilon_{min}$ **then**
            $\varepsilon_{min} \leftarrow \varepsilon$
            $\phi_n \leftarrow \phi_n$-m_j
          end if
        else
          Classifier ($T_{\phi n}$-m_i,ε)
          if $\varepsilon \leq \varepsilon_{min}$ **then**
            $\varepsilon_{min} \leftarrow \varepsilon$
            $\phi_n \leftarrow \phi_n$-m_i
          end if
        end if
      end if
    end for
  end for

The complexity of this algorithm is $n^2 * O_{class}$ with regard to calculation time.

V. Experimental study

To evaluate these algorithms, five protein families have been extracted at random from the data bank *SCOP*. We have chosen a discrimination strategy between the protein families: these families will be discriminated two by two, learning algorithm was applied (k=1-nearest neighbour)[11] after and before selection to estimate the error rate of classification for each couple of protein family. We chose a 5 X 2 cross validation[11]. The results illustrated in tab.2

TABLE II
Comparison between the results received with our algorithms of features selection with those obtained using the filter Relief algorithm.

Proteins	Motif number				Error rate			
families	RF	RH1	RH2	Rh3	RF	RH1	RH2	RH3
$f_1 \cup f_2$	986	181	130	98	0.0860	0.0767	0.0535	0.0512
$f_1 \cup f_3$	1506	183	180	114	0.1447	0.1128	0.1085	0.1085
$f_1 \cup f_4$	1044	349	320	195	0.0817	0.0483	0.0450	0.0417
$f_1 \cup f_5$	969	193	50	40	0.1037	0.0537	0.0524	0.05
$f_2 \cup f_3$	1676	269	130	107	0.1440	0.0780	0.04	0.038
$f_2 \cup f_4$	953	291	240	198	0.0906	0.0547	0.05	0.0484
$f_2 \cup f_5$	1208	309	140	103	0.119	0.0877	0.0579	0.0509
$f_3 \cup f_4$	1101	466	200	116	0.1104	0.1149	0.097	0.091
$f_3 \cup f_5$	1252	151	30	25	0.223	0.1033	0.1033	0.0803
$f_4 \cup f_5$	1196	373	80	64	0.0811	0.0784	0.0446	0.0405

confirm the performance of the hybrid approach for feature selection relative to filter one.The relief hybrid algorithms enable us to select a reduced number of features having a strong discrimination capacity of proteins families.

TABLE III
Comparison between the classification error rate using the hybrid3 Relief algorithm and the CFS filter algorithm

Proteins	Before selection	After selection	
families	Error rate	Error rate(CFS)	Error rate(RH3)
$f_1 \cup f_2$	0.2163	0.0837	0.0512*
$f_1 \cup f_3$	0.2043	0.1340	0.1085*
$f_1 \cup f_4$	0.1917	0.0317*	0.0417
$f_1 \cup f_5$	0.2463	0.1167	0.05*
$f_2 \cup f_3$	0.2060	0.0380	0.038*
$f_2 \cup f_4$	0.1281	0.0500	0.0484*
$f_2 \cup f_5$	0.1649	0.0474*	0.0509
$f_3 \cup f_4$	0.2701	0.0746*	0.091
$f_3 \cup f_5$	0.3311	0.0803	0.0803*
$f_4 \cup f_5$	0.1689	0.0662	0.0405*

VI. CONCLUSIONS

Within the aim of improving the performance of the filter Relief algorithm, we have developed three algorithms of features selection. The aim of these algorithms is to select a reduced number of features, the most discriminated and non redundant by adopting hybrid,the idea is to produce subset including candidate features and to exploit them in order to identify the subset of features that minimize a certain error rate of classification. Experimental results show that hybrid Relief algorithms seem a promising ways in the context of proteins classification problem where we have few examples and numerous features.

References

[1] C.Gibas and P.Jambeck ,"Introduction a la bioinformatique",*Oreilly*.
[2] G. Isabelle and E. Andr , "An introduction to variable and feature selection", *Journal of Machine Learning Research* ,2003,1157-1182.
[3] F. Mhamdi, M. Elloumi and R. Rakotomalala, "Textmining,feature selection and datamining for proteins classification " *Proc of ICTTA'04 International Conference on Information & Communication Technologies: from Theory to Applications*,IEEE,2004.
[4] W. Duch, T. Wieczorek , J. Biesiada and M.Blachnik "Comparison of feature ranking methods based on information entropy " *Proc of International joint Conference on Neural Networks(IJCNN)*,Budapest,IEEE press,1415-1420,2004.
[5] K. Kira and L. Rendell , "A practical approach to feature selection" , *the Ninth International Conference on Machine Learning,San Francisco, Morgan Kaufmann*, CA ,1992.
[6] F. Mhamdi, M. Elloumi, "Approches of Feature Construction from Biological Sequences", *In the 3rd IAPR International Conference on Pattern Recognition in Bioinformatics* (PRIB 2008).
[7] H. Mhamdi and F. Mhamdi , "Etude comparative d'algorithmes filtres pour la selection de motifs biologiques" , *Conference Francophone sur l'Extraction et la gestion de connaissances*,IRIT,Toulouse,2013.
[8] A. Murzin, S. Brenner, T. Hubbard, C. Chothia, Scop, "A Structural Classification of Proteins Database for the Investigation of Sequences and Structures", *Journal of Molecular Biology*, p536540,1995
[9] L. Yu et H. Liu, "Feature selection for high-dimensional data: a fast correlation based filter solution",*In the 20th International Conference on Machine Learning (ICML03)*, p 856863, 2003.
[10] O.Chapelle, J. Weston et A. Elisseeff, "Embedded methods", *Spinger verlag*, p137-195, 2006.
[11] T.Hastie, R. Tibshirani et J. Friedman, "The Elements of Statistical Learning:Data mining,Inference,and Prediction", *Spinger verlag*, 2001.