HIERARCHICAL FEATURE SELECTION FOR BIOLOGICAL DATA

Witold Jacak[a], Karin Proell[b]

[a] Department of Software Engineering
Upper Austrian University of Applied Sciences Hagenberg, Austria
[b] Department of Medical Informatics and Bioinformatics
Upper Austrian University of Applied Sciences Hagenberg, Austria

(a) witold.jacak@fh-hagenberg.at (b) karin.proell@fh-hagenberg.at

ABSTRACT
In this paper we present feature selection in biological data by combining unsupervised learning with supervised cross validation. Unsupervised clustering methods are used to perform a clustering of object-data for a chosen subset of input features and given number of clusters. The resulting object clusters are compared with the predefined original object classes and a matching factor (score) is calculated. This score is used as criterion function for heuristic sequential feature selection and a cross selection algorithm.

Keywords: Index Terms— classification, clustering, feature selection, sequential feature selection

1. INTRODUCTION

Classification of biological data means to develop a model that will divide biological observations into a set of predetermined classes N. Typically a biological data set is composed of many variables (features) that represent measures of biological attributes in biological experiments. A common aspect of biological data is its high dimensionality that means data dimension is high, but the sample size is relatively small. This phenomenon is called high dimensionality-small sample problem (Kwak et al. 2002, Maiorana et al. 2008, Zhang et al. 2000). The smaller the sample, the less accurate are the results of classification and the amount of error increases. In traditional statistical classification procedures a probability model must be assumed or estimated in order to calculate the posterior probability upon which the classification decision is performed. One major limitation of the statistical models is that they work well when the underlying assumptions are satisfied. Users must have a good knowledge of both data properties and model capabilities before the models can be successfully applied.

Using all possible features for model creation does not necessarily give the best performance. A model with less features (variables) is faster to construct and easier to interpret, especially in biological data mining where a domain expert should interpret and validate such a result. Using classical supervised clustering and classification methods could lead especially in case of small sample sets, to a faster overfitting of a model during the training phase and to worse prediction performance.

Generally, the feature selection problem deals with choosing those input variables from the measurement space (all input variable) that are most predictive for a given target and reduce the feature space. The main objective of this process is to retain the optimum number of input variables necessary for the target recognition and to reduce the dimensionality of the measurement space so that an effective and easily computable model can be created for efficient data classification. Appropriate feature (input variable) selection can enhance the effectiveness of an inference model (Kohavi and John 1997).

2. FEATURE SELECTION SYSTEM

Feature Selection is a process that identifies a subset of original features. The goal is to reduce the number of features by removing redundant, irrelevant and noisy data in order to improve the predictive accuracy of the model. A typical feature selection process uses four main steps: subset generation, subset evaluation, stopping condition and result validation. Subset generation is a search procedure that produces candidate feature subsets for evaluation based on a certain search strategy. Each candidate subset is evaluated and compared with the previous best one according to a certain objective function. If the new subset turns out to be better it replaces the previous best subset. The process of subset generation and evaluation is repeated until a given stopping condition is satisfied. Then the best subset usually needs to be validated by prior knowledge or different test via real-world datasets (Huan 2005).

Subset generation: Search for subsets starts with an empty set and adds features successively or start with a full set and successively remove features or start with both ends and add and remove features simultaneously. For a dataset with N features there exist 2^N candidate subsets. Such a number of candidate subsets is unfeasible even for moderate numbers of features. Therefore a search strategy in needed in order to direct the feature selection process as it explores the space of all possible combination of features. Therefore different strategies haven applied for candidate subset identification: complete, sequential and random search. In our work we apply traditional sequential forward selection for generating subsets of features. These algorithms use greedy local heuristic search, which incrementally adds and/or deletes features to obtain a subset of relevant features with respect to the response. This search gives up completeness and there is a risk to lose optimal subsets. There are other variations to this greedy hill-climbing approach like sequential backward selection or bi-directional selection. Algorithms with sequential search are easily implemented and produce results faster because the order of the search space is usually $O(N^2)$. The problem

with these sequential approaches is that they gravitate toward local minima due to the inability to re-evaluate the usefulness of features that were previously added or discarded (Huan 2005).

Subset evaluation: As mentioned above each newly generated feature subset needs to be evaluated by a criterion. This measurement is returned by an objective function which is applied to the feature subset monitoring the "goodness" of a candidate set- a feedback signal used by the search strategy to select new candidates. Objective functions are divided in two groups. Filters evaluates feature subsets by their information content, typically interclass distance, statistical dependence or information-theoretic measures and wrappers use a pattern classifier, which evaluates feature subsets by their predictive accuracy (recognition rate on test data) by statistical resampling or cross-validation (Hua 2005). In our work we make use of the wrapper approach and present a method for objective function specification, called matching factor, combining unsupervised classification of data with prior knowledge about classes of the data.

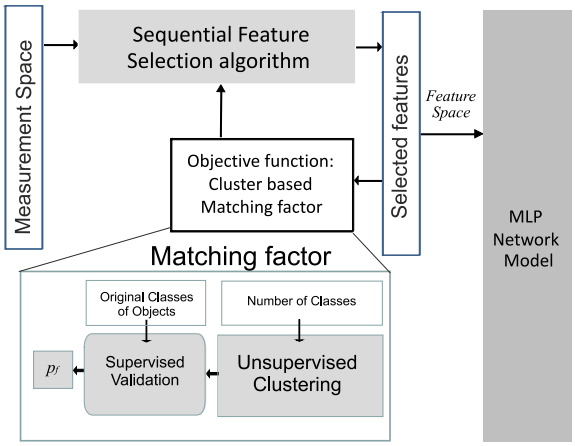

Fig. 1: Feature selection system based on unsupervised learning

Result validation: In real-world applications we usually do not have prior knowledge about relevant features to directly measure the results. We have to rely on indirect methods by monitoring the change of predictive accuracy with the change of features. In case of a classification task we can use a classification error rate as a performance indicator for a selected feature subset, simply by conducting the "before-and-after" experiment to compare the error rate of the classifier learned on the full set of features and that learned on the selected subset (Huan 2005). The system we propose is presented in Fig.1.

2.1. Objective function

Applying a wrapper approach for objective function we can give a generalized overview (Huan 2005) of the used algorithm for sequential feature selection (see Fig.2.).

input:	$L(f_0,f_1,\ldots\ldots,f_{n-1})$	// training data set with N features
	S_0	// a subset from which to start the search
	δ	// a stopping criterion
Output: S_{best}		// an optimal subset
01: begin		
02:	initialize: $S_{best} = S_0$;	
03:	pf_{best}= eval(S_0, D, mf _calculation);	//evaluate S0 by matching //factor objective function
04:	do begin:	
05:	S=generate(D);	// generate next subset for evaluation
06:	pf = eval(S, D, mf _calculation);	//evaluate S by matching // factor objective function
07:	if (*pf* is better than pf_{best})	
08:	$pf_{best} = pf$;	
09:	$S_{best} = S$;	
10:	end until (δ is reached);	
11:	return S_{best};	
12: end		

Table 1: Wrapper Algorithm with matching factor calculation as objective function

For each generated subset S, the sequential forward algorithm evaluates its goodness using the matching factor calculation as an objective function for feature subset S (see Table 1). Our objective function - called matching factor calculation - applies different clustering methods on each chosen subset of features.

Although we have knowledge of classes of the input data a priori we do not make use of this knowledge initially but have clustering methods (Ye and Liu H 2002) define a partitioning of the input data using simply the well-known number of classes as input parameter to the clustering methods (Thangavel, Shen and Pethalakshmi 2006, Törmä 1996). An average value on the results of the applied clustering methods is calculated which is finally compared to the original assignment of classes of the input data and the matching factor is calculated.

The algorithm for matching factor calculation works as follows and can be seen in full detail in Jacak, Proell and Winkler 2014.

Let $I= \{1,...,N\}$ be a set of indexes of target classes L and $J= \{1,...,N\}$ be a set of indexes of cluster set K generated from pure input data of objects with unsupervised learning using for example SOM-SL neural network. Let $L_i \in L$ be a subset of objects belonging to original class $i \in I$ and $K_j \in K$ is a subset of objects belonging to cluster $j \in J$ created by the clustering algorithms. For each $i \in I$ and $j \in J$ we define the Jaccard matching coefficient c_{ij} as follows

$$c_{ij} = \frac{|L_i \cap K_j|}{|L_i \cup K_j|} \qquad (1)$$

We use this coefficient to define the global matching factor p_f between the generated cluster and the original target classes in an iterative manner. For $n=1,..,N$ we calculate

$$p^n_{ij} = \begin{cases} c_{ij} & \text{if } c_{ij} = max\{c_{lk} \mid l \in I_{n-1}, k \in J_{n-1}\} \text{ and} \\ & s_i = \Sigma(c_{ik} \mid k \in J_{n-1}) = min \text{ and} \\ & s_j = \Sigma(c_{lj} \mid l \in I_{n-1}) = min \\ 0 & other \end{cases} \qquad (2)$$

where $I_0 = I$, $J_0 = J$, and $J_n = J_{n-1} - \{j\}$, $I_n = I_{n-1} - \{i\}$. Calculation stops when J_n and I_n are empty sets.

It is easy to observe that the $N \times N$ sized matrix $P = [p_{ij}]$ has only N elements greater zero and represent the best allotment of the generated clusters to the original classes. The global matching factor p_f can be defined as mean value of non-zero elements of P:

$$p_f = avg(p_{ij}) \qquad \text{where } p_{ij} > 0 \qquad (3)$$

The matching factor describes the score of recognizing the original target classes by unsupervised clustering which is uses subsets of input data without prior knowledge of classes. This factor uses posterior knowledge to calculate the score for validating the goodness on a chosen subset of features which should maximize the matching factor.

2.2. Hierarchical feature selection – Cross feature selection

In the above description we applied standard sequential feature selection using the matching factor calculation as objective function. In this section we present an alternate approach for feature selection which can be applied for at least four target classes. We call this approach hierarchical or cross feature selection.

When original target classes $L = \{L_1, L_2, \ldots, L_N\}$ are known then it is possible to construct different subsets of original classes. The subset of original classes

$$\{L_i, L_k, \ldots, L_m\} \subset L$$

is grouped together into one new class, called hyper-class

$$LC_i = \bigcup \{L_i, L_k, \ldots, L_m\}$$

This hyper-class contains all objects data from its components. The new non overlapping hyper-classes set LC_i $(i=1\ldots K)$ should cover a set of original classes, i.e.

$$\bigcup LC_i = \bigcup L \text{ and } \cap LC_i = \varnothing$$

Fact:

Let LC^k denote the k-th experiment (k = 1,..,M) with chosen hyper-classes, where $LC^k = \{LC_i^k \mid i=1,..,K^k\}$ and LC_i^k. is i-th hyper class in the k-th experiment. Let $F^k \subset \{1,...,m\}$ be the set of selected sensitive features for the best recognition of the hyper classes in this experiment, calculated by sequential feature selection algorithm.

When the chosen new hyper-classes sets LC^k (k = 1,..,M) meet the condition

$$\cap LC^k = L$$

then the cross selected sensitive features for recognition of the full original target L could be defined as

$$F = \bigcup F^k .$$

Such selected features based on cross comparison of subsets of target classes can be compared to feature sets found by the forward/backward sequential selection procedure for original target classes. The union of both candidates' sets of features can be used by exhaustive search to find the optimal final feature space.

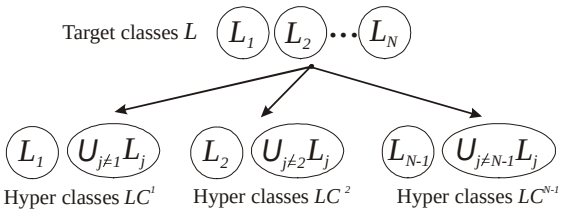

Fig. 2: Construction of hyper-classes for N-1 experiments

The easiest way to find hyper-classes is the breakdown of the whole collection of classes into two groups. The first group contains only one class and second one a union of all the others (see Fig. 2).

$$LC^i = \{L_i, \cup\{L_j \mid j \neq i\}\}$$

So we have only two hyper classes. Such a construction of hyper-classes conducted for each class separately (for $i=1\ldots N-1$), satisfies the conditions of fact and can be used in cross method of selection.

$$\cap \{LC^i \mid i=1\ldots N-1\} = L$$

If we make such groupings for particular classes, combining parameters for each group represent a new selection for recognition all L classes.

$$F_{cross\ selection} = \cup F^i$$

where F^i is the set of selected features for recognition of LC^i hyper-classes.

The hyper-class construction algorithm used in the cross selection method required $N-1$ times running of process for the sequential feature selection.

2.3. Result evaluation

After the feature selection phase a multi-layer perceptron (MLP) network is used to classify the training and validation dataset in the reduced feature space. The classification performance of each feature set is compared with the classes indicated in the original dataset. We used a two-layer perceptron network (MLP) with four neurons in the hidden layer.

3. EXPERIMENTS AND RESULTS

In this paper a cross feature selection method independent of the underlying classification model and based on clustering algorithms with posterior validation (matching factor) has been proposed. For our experiments we used a salmon growth experiment using data from fish sampled following 90 days of feeding (Epstein 2012, Gu et al. 2013). A collection of biological data for 288 fish is used. Each fish is described with 89 features max. Features are just labeled to have analyst blind to the actual variables measured. The measurement space – the full feature set - is therefore $F = (f1,\ldots, f89)$. Missing values in the input samples were replaced by the mean value of the parameters into consideration from the respective class and all data were normalized.

The fish belong to four predefined classes (target classes) L = {A,B,C,D}. Each class corresponds to a certain kind of food. We did the following class labeling depending on the

given values of GMmatrix and GMvalue in the experiment data (see Table 2).

Class	GMmatrix	GMvalue
A	1	a
B	1	b
C	2	a
D	2	b

Table 2: Class labeling depending on the given values of GMmatrix and GMvalue

Applying sequential forward selection and cross feature selection the following results were provided:
The empirical normal distribution PDF 1 (continuous line) of ratio of matching factor calculated for cross selected features to matching factor for full set of features is presented in Fig 3. This figure presents also the empirical normal distribution PDF 2 (dotted line) of ratio of matching factor calculated for forward sequential selected features to matching factor for full set of features. These ratios can be interpreted as quality of target classes' recognition.

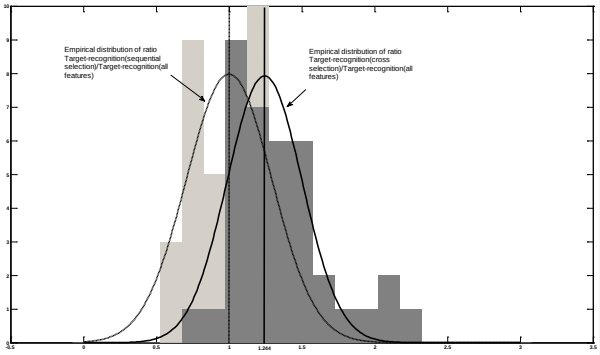

Fig 3: Empirical distribution of target recognition quality of cross selection method (pdf 2-continuous line) and forward sequential selection method (pdf 1-dotted line).

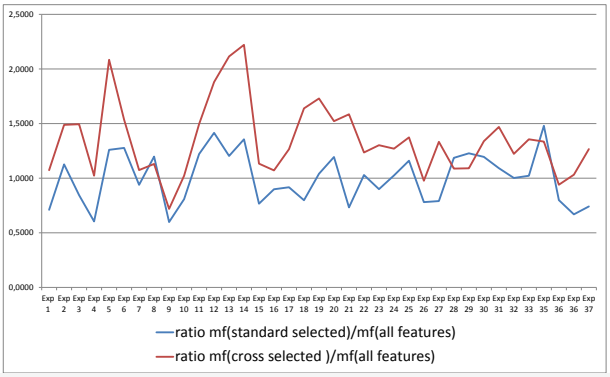

Fig 4: Comparison of matching factor ratio between standard features selection and cross features selection methods

The expectation value of empirical distribution PDF 1 is 1,244 for cross feature selection with matching factor objective function and the expectation value of empirical distribution PDF 2 is 1,01 for standard sequential feature selection. Cross selection method finds average 2,29 times more features than the standard sequential selection procedure. A direct comparison of recognition quality between cross selection found features and standard forward sequential selection found features shows that the cross selection method has about 20 % better recognition quality (see Fig.4).

4. CONCLUSION

In this paper a feature selection method independent of the underlying classification model and based on clustering algorithms with posterior validation has been proposed. The analysis was performed on biological data. The presented method can be classified as a heuristic method, which obtains an effective feature reduction with almost the same correct classification rate. Heuristics is designed on an unsupervised learning approach. The optimal number of feature space dimensions is therefore difficult to determine. Future works are to extend the analysis to datasets with different number of samples and to further investigate the distance measures to assess the impact on classification performance.

REFERENCES

Epstein M. (ed.), 2012, Proc. Of GMSAFOOD Conference, Vienna

Gu J., Krogdahl A., Sissener N.H., Kortner T.M., Gelencser E., Hemre G.I., Bakke A.M., 2013, "*Effects of oral Bt-maize (MON810) exposure on growth and health parameters in normal and sensitised Atlantic salmon, Salmo salar L"*, British Journal of Nutrition 2013 Apr 28

Jacak W., Proell, Winkler, K., S., 2014, Neural Networks based Feature Selection in Biological Data Analysis. in: Klempous, R., J. Nikodem, W. Jacak, Z. Chacko, eds. *Advanced Methods and Applications in Computational Intelligence*, Springer, pp.79-93

Kohavi R., John G. H., 1997, "*Wrappers for feature subset selection*," Artificial Intelligence, vol. 97, no. 1-2

Kwak N., 2002, Choi C., "*Input Feature Selection for Classification Problems"*, IEEE Transactions on Neural Networks, Vol. 13, No. 1

Huan L., 2005, "*Toward integrating feature selection algorithms for classification and clustering"*, Knowledge and Data Engineering, IEEE Transactions on, Vol. 17 , Issue 4

Maiorana F., 2008, "*Feature Selection with Kohonen Self Organizing Classification Algorithm"*, World Academy of Science, Engineering and Technology, Proceedings of WASET, Vol. 45

Thangavel K., Shen Qiang, Pethalakshmi A., 2006, "*Application of Clustering for Feature Selection Based on Rough Set Theory Approach*", AIML Journal, Volume (6), Issue (1)

Törmä M., 1996, *Self-organizing neural networks in feature extraction"*, Intern. Arch. of Photogrammetry and Remote Sensing, Vol. XXXI, Part 2

Ye H., Liu H., 2002, "*A SOM-based method for feature selection*", Proceedings of the 9th International Conference on Neural Information Processing (ICONIP'O2), Vol. 3

Zhang G. P., 2000, "*Neural Networks for Classification: A Survey*", IEEE Transactions on Systems, Man, And Cybernetics—Part C: Applications And Reviews, Vol. 30, No. 4