UNSUPERVISED LEARNING APPROACH TO FEATURE SELECTION IN BIOLOGICAL DATA ANALYSIS

Witold Jacak[a], Karin Proell[b]

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

ABSTRACT
In this paper we present a novel method for scoring function specification and feature selection by combining unsupervised learning with supervised cross validation. Unsupervised clustering methods (k-means, one dimensional Kohonen SOM, fuzzy c-means) 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 novel cross selection algorithm.

Keywords: Index Terms— classification, clustering, feature selection, k-mean, fuzzy c-mean, Kohonen SOM, MLP

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 the 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.

Traditional statistical classification procedures such as discriminant analysis are built on the Bayesian decision theory (Kwak et al. 2002, Raghavendra et al. 2010, Silva et al. 2012). In these procedures, a probability model must be assumed 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 only 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 (Kohavi et al. 1997, Törmä et al. 1996). The recent research activities in classification problem have established that neural networks are a promising alternative to various conventional classification methods (Chen et al. 2006, Maiorana et al. 2008, Su et al. 2000, Raghavendra et al. 2010, Zhang et al. 2000, Ye et al. 2002). Neural networks are data driven self-adaptive systems in a way that they can adjust themselves to the data without any explicit specification of a functional or distributional form for the underlying model. Neural networks are able to estimate the posterior probabilities, which provide the basis for establishing classification rules and performing statistical analysis (Zhang et al. 2000, Ye et al. 2002). The effectiveness of neural network classification has been tested empirically, and depends on the quality of input data.

There are many reasons to perform feature selection before developing the classification model. Feature selection problems are found in all supervised and unsupervised neural network learning tasks including classification, regression, time-series prediction, and clustering (Lee et al. 2004, Maiorana et al. 2008, Raghavendra et al. 2010, Törmä et al. 1996).

Using all features to create the model does not necessarily give the best performance (peaking phenomenon) (Maiorana et al. 2008, Pal et al. 2002). 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 model. 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. Previously performed feature selection could avoid overfitting and improve the model performance in prediction.

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 constitute 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.

Feature selection methods can be grouped in four categories (Kwak et al. 2002, Maiorana et al. 2008, Zhang et al. 2000).

1) Filter techniques select the features by looking only at the intrinsic property of input data and ignore feature dependencies. These techniques are independent of the classification model (Kwak et al. 2002).

2) Wrapper methods use the hypothesis model for search of a feature subset. Hypothesis models are usually

constructed in a supervised way during the training phase (e.g. Bayes model or multi-layer perceptron neural network MLP) but they increase the risk of overfitting especially in case of small sample sizes (Kohavi et al. 1997).

3) Embedded techniques in which the search for an optimal subset of features is built into classifier construction (Kohavi et al. 1997, Su et al. 2000, Thangavel et al. 2006).

4) Feature selection after classification uses discriminant analysis to find sensitive variables (Kwak et al. 2002, Pal et al. 2002).

These techniques can be used by exhaustive search or heuristic search. The exhaustive search method tries to find the best subset among 2^m candidate subsets, where m is the total number of features. Such a search could be very time consuming and practically unacceptable; even for a medium–sized feature set. Heuristic search methods, which use a learning approach, reduce computational complexity (Kwak et al. 2002). There are two kinds of learning methods, supervised learning and unsupervised learning. In case of supervised learning the incorporation of prior knowledge about membership of classes into the training data is the key element to substantially improve classification performance. But full knowledge increases the risk of overfitting during the training phase of the model and as a consequence prediction performance for novel data decreases. On the other hand unsupervised learning ignores interaction with the classifier model (Faro et al.,2005, Maiorana et al. 2008, Törmä et al. 1996, Ye et al. 2002).

In this paper we present a novel method for scoring function specification and feature selection by combining unsupervised learning with supervised cross validation. A one dimensional Kohonen SOM (Self-Organizing Map) is used to perform a clustering of objects-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. Additionally, the significance of an individual feature for recognition of original classes or composed groups of original classes is calculated based on this matching factor. The results are compared and aggregated with the result of sequential feature selection to find the final sensitive feature space. Sequential feature selection searches for a subset of the features in the full model with comparative predictive power. The final result is a reduced model with only few of the original features. In the next step these features are used to build and train the MLP network model for object classification.

2. UNSUPERVISED NEURAL NETWORKS BASED MATCHING FACTOR

The four different clustering methods are used to perform clustering for given number of cluster based on object-data for chosen input features subset. Those obtained clusters are finally compared and validated with the predefined original classes and matching factors (scores) are calculated. The system for classifying object-data model consists of four procedures working in parallel: k-mean clustering, fuzzy c-mean clustering, one dimensional Kohonen SOM neural network and hierarchical clustering. Usually the output neurons of SOM are arranged in a bi-dimensional array, but we use an implementation of the network where the output neurons are arranged along a single layer (SL configuration). In the SL configuration the classes are given by the output neurons. In this case there is no topological similarity between output neurons since adjacent output neurons do not represent necessarily similar classes.

The resulting clusters are compared to one another and to the predefined target classes by using the matching factor p. A final score takes into account only the results of the independent clustering procedures with pairwise high matching factors and a high matching factor between its cluster and the original target classes. The final score is calculated as the mean value of the matching factors of these clusters. The system for calculating the matching factor is presented in Fig. 1.

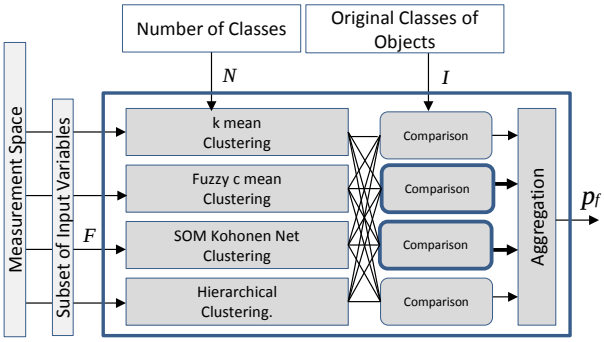


Fig. 1 Calculation of global matching score

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 SOM-SL network. 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 simply based on input data without prior knowledge of classes. This factor uses posterior knowledge to calculate the score for validating the chosen set of individual features or group feature groups. The choice of features should maximize the matching factor.

3. FEATURE SELECTION SYSTEM

The matching factor p_f is used as score for finding the sensitive group of input features. The feature selection system contains two modules: one module for sequential feature selection techniques with matching factor as quality criterion function, and one module for searching individual sensitive features, which discriminate the chosen group of composed subset of original predefine classes. The system is presented on Fig.2.

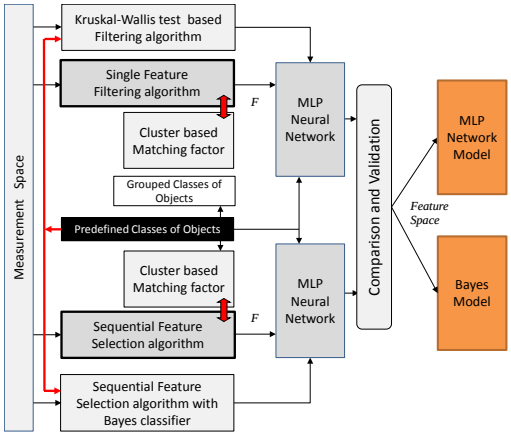


Fig. 2 Mixed feature selection system based on unsupervised learning

3.1. Features cross selection method

The subset of original classes $\{L_i, L_k, \ldots, L_m\} \subset L$ is grouped together into one new class $LC_i = \bigcup \{L_i, L_k, \ldots, L_m\}$, which contains all objects data from its components. The new classes set LC cover a set of original classes, i.e.

$\bigcup LC_i = \bigcup L$.

The new set of target classes determines the new number of classes. This number $|LC|$ is used for unsupervised clustering the previously grouped input data with clustering procedures and next to calculate the matching factor between generated clusters and the new classes set. In this way we can search for significant features, which discriminate the subsets of original classes. Usually the grouping of original classes into larger new classes is not accidental and is performed based on domain specific knowledge concerning properties of objects being classified.

For each group of target classes and also for original classes it is possible to test the significance of every single variable due to its predictivity of the target classes. After the calculation of the matching factor individually for each variable, only variables with matching factor greater than a given threshold-value are chosen from all m variables as candidates for feature space.

Let LC^k ($k = 1,..,M$) denote the k-th experiment with chosen subsets of original target classes then $F^k \subset \{1,...,m\}$ is the set of selected sensitive features for the best recognition of the new composed classes in this experiment. When the chosen new larger classes sets LC^k meet the condition

$$\cap LC^k = L \qquad (4)$$

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

$$F = \bigcup F^k \qquad (5)$$

Such selected features are compared to feature sets found by the forward/backward sequential selection procedure. Both candidates of sensitive feature sets are used to train a multi-layer perceptron neural network (MLP) to build the model and to evaluate the performance of target recognition and prediction.

4. EXPERIMENTS AND RESULTS

In this study, the aim is to test the quality of a feature selection method based on a combination of unsupervised learning and posterior validation in comparison to standard statistic algorithms. A collection of biological data for 288 objects is used in experiment. The objects belong to four predefined classes (target classes) $L = \{A,B,C,D\}$. Each object is described with 89 input measurement variables (parameters). The measurement space - full feature set - is therefore $F = (f_1, \ldots, f_{89})$.

The task is to find a subset of input variables $F_i \subset F$ with a minimal number of sensitive features, which recognize the four target classes without loss of classification quality. This subset will be used to develop a model for object classification. Additionally the feature sets, which recognize two subsets of the original target classes were identified. In the first the classes A,B and C,D are combined together i.e. $LC^1 = \{\{A \cup B\}, \{C \cup D\}\}$ and in the second the classes A,C and B,D i.e. $LC^2 = \{\{A \cup C\}, \{B \cup D\}\}$. Since $LC^1 \cap LC^2 = L$, the union of sensible features for recognition of LC^1 and LC^2 could be used as a discriminant feature set for the whole target classes set L.

For all groups of target classes we calculate the matching factor p_f individually for each input variable and compare it to statistical significance of Kruskal–Wallis one-way analysis of variance test results. We consider a threshold equal to 80% of maximal value of p_f to filter the first candidate set of significant features. The matching factor for single variable, the inverse of the p-value of Kruskal-Wallis test and filtered set of features for recognition of four original classes is presented in Fig.3.

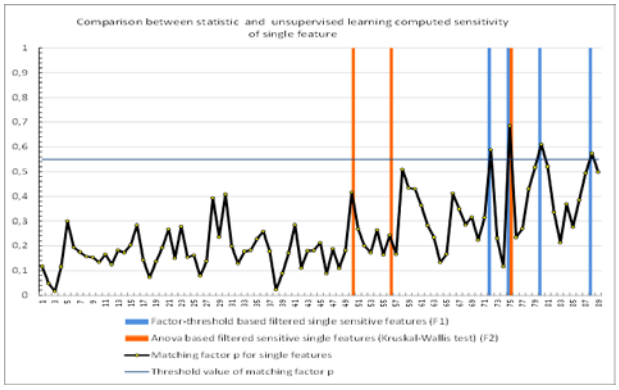
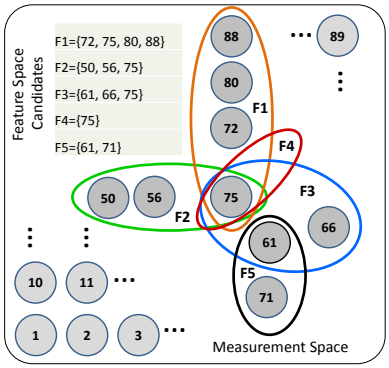


Fig. 3 The matching factor for single variable and filtered set of features for recognition of four original classes

The whole collection of data was used as input features for a forward sequential feature selection with the SOM matching factor p_f as performance criterion and additionally we performed a second sequential feature selection with a Bayes classifier as performance function. The selected features in both approaches are presented in Fig. 4.

Both methods applied for previously described classes LC^1 and LC^2 found only one single sensitive feature in each of both cases, $\{f_{61}\}$ and $\{f_{71}\}$ respectively (see Fig.4.). The union of these is used as the new features subset of candidates for feature space. The results of the applied feature selection techniques are presented in Fig.4.



Method	Selected Features	Prediction performance
Factor-threshold based filtered single sensitive features (F1)	F1={72, 75, 80, 88}	87,8%
Anova based filtered sensitive single features (Kruskal-Wallis test) (F2)	F2={50, 56, 75}	78,3%
Sequential feature selection with matching factor p as performance function (F3)	F3={61, 66, 75}	95,7%
Sequential feature selection with Bayes classifier as performance function (F4)	F4={75}	73,9%
Features selected as union of discriminants of subsets of target classes (F5)	F5={61, 71}	99,2%

Fig.4. Features selected with different algorithms for recognition of 4 original object classes

After the feature selection phase a multi-layer perceptron network is used to classify the training and validation dataset in the reduced feature space. The classification results are compared with the classes indicated in the original dataset.

The results of feature selection were validated with a two-layer perceptron network (MLP) with four neurons in the hidden layer. The data set was divided into training set - 70 % of samples and test set -30% of samples. The correct classification rate to original class in percent is presented in Fig. 5.

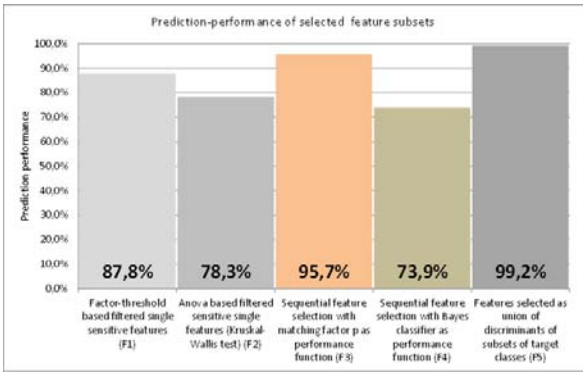


Fig.5. Comparison of classification-performance of selected features subsets

The feature spaces found using the combined unsupervised clustering and the posterior validation for matching factor calculation show a better classification rate for predefined original classes as features selected with classical statistic methods (Fig. 5). It demonstrates the feasibility and effectiveness of the proposed novel method for score calculation, which is independent of the classification model.

5. CONCLUSION

In this paper a feature selection method independent of the underlying classification model and based on clustering algorithms of the Kohonen SOM family, k-mean, fuzzy c–mean 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

Chen Y., Abraham A., Yang B., "*Feature selection and classification using flexible neural tree*", Neurocomputing, Vol. 70, 2006

Faro G., Giordano D., Maiorana F., "*Discovering complex regularities by adaptive Self Organizing classification*," Proceedings of WASET, Vol. 4, 2005:

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

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

Lee K., Booth D., State K., Alam P., "Backpropagation and Kohonen Self-Organizing Feature Map in Bankruptcy Prediction" in Neural Networks in Business Forecasting, edited by G. Peter Zhang, , Idea Group Inc., 2004

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

Su Mu-Chun, Chang Hsiao-Te, "*Fast Self-Organizing Feature Map Algorithm"*, IEEE Transactions on Neural Networks, Vol. 11, no. 3, 2000

Pal S. K., De R. K., Basak J., "*Unsupervised Feature Evaluation: A Neuro-Fuzzy Approach*", IEEE Transactions on Neural Networks, Vol. 11, No. 2, 2000

Raghavendra B. K., Simha J. B., "*Evaluation of Feature Selection Methods for Predictive Modeling Using Neural Networks in Credits Scoring*", Int. J. Advanced Networking and Applications, Vol-02, Issue: 03, 2010

Silva B., Marques N., "*Feature clustering with self-organizing maps and an application to financial time-series for portfolio selection*", Proceedings of Intern. Conf. of Neural Computation, ICNC, 2010

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

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

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

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