Multi-Objective Semi-Supervised Feature Selection and Model Selection Based on Pearson's Correlation Coefficient

Frederico Coelho[1], Antonio Padua Braga[1], and Michel Verleysen[2]

[1] Universidade Federal de Minas Gerais,
Belo Horizonte, MG, Brazil
{fredgfc,apbraga}@ufmg.br
www.ufmg.br
[2] Universite Catholique de Louvain,
Louvain-la-Neuve, Belgium
michel.verleysen@uclouvain.be
www.uclouvain.be

Abstract. This paper presents a Semi-Supervised Feature Selection Method based on a univariate relevance measure applied to a multiobjective approach of the problem. Along the process of decision of the optimal solution within Pareto-optimal set, atempting to maximize the relevance indexes of each feature, it is possible to determine a minimum set of relevant features and, at the same time, to determine the optimal model of the neural network.

Keywords: Semi-supervised, feature selection, Pearson, Relief.

1 Introduction

In recent years, especially in the fields of bioinformatics and web-based information retrieval, the problem of Semi-Supervised Learning [6] (SSL) has gained increased interest. Broadly speaking, the problem involves the construction of classifiers with very limited labeling information and large amount of unlabeled data. Particularly in these areas, new samples are easily generated but model induction from input-output data is faced with scarce data due to the high cost for labeling. Due to the availability of a large amount of untagged input data, the question that arises is whether to use or not such a huge amount of information in model induction.

The general problem is characterized by the induction of a model from the labeled dataset $D_L = \{\mathbf{x}_i, y_i\}_{i=1}^{N_L}$ considering also the structural information contained in the unlabeled set $D_U = \{\mathbf{x}_i\}_{i=1}^{N_U}$[1]. The approaches for such a problem usually involve jointly solving the supervised problem defined by D_L and the unsupervised one defined by D_U [1,4].

[1] N_L and N_U are, respectively, the sizes of the labeled and unlabeled datasets, x_i is the i^th observation and y_i is the class label of i^th observation.

I. Bloch and R.M. Cesar, Jr. (Eds.): CIARP 2010, LNCS 6419, pp. 509–516, 2010.

Feature selection in such a framework is also faced with the same problem of dealing with small input-output samples under the availability of large amounts of untagged data, so the problem should also be handled in both fronts. Clearly, unsupervised feature selection methods [12,7] could be applied to the whole dataset $D_L \cup D_U$, but disregarding the labels $y_i \in D_L$ could represent loss of (available) information. Therefore, Semi-Supervised Feature Selection (SSFS) is also faced with the problem of selecting features by considering both datasets D_L and D_U [24,21,3]. In addition, the Supervised Learning (SL) problem characterized by D_L, involves the many issues related to supervised learning, such as minimizing both the empirical and the structural risks of the model [22]. Feature selection with embedded [3] or wrapped [14] models should also take into consideration such general issues in order to guarantee reliability in the search for representative models.

In this paper, we present a new SSFS method that allows both the selection of a classifier from a set of neural networks candidate solutions generated by a multi-objective (MOBJ) learning method [17] and the selection of relevant features for such a model. The Pareto-set solutions of the MOBJ method are obtained according to the general statistical learning principles [22], by minimizing both the empirical and the structural risks, represented by the sum of squared errors $\sum e^2$ and the norm of the neural network weight vectors $||\mathbf{w}||$ [17]. Once the Pareto-Optimal solutions are generated in a supervised manner, by considering only D_L, they also yield labels for D_U, since each Pareto-Optimal classifier is valid in the whole input domain. Therefore, for each Pareto-set solution S_k, there is a labeling $D_U^k\{\mathbf{x}_i\, \hat{y}_i^k\}_{i=1}^{N_U}$ for D_U. The aim of the feature selection method is to find the optimal solution S_* that maximizes the separability of $D_L \cup D_U^k$; the features with the highest relevance indexes (RI) are then selected.

The method can be regarded both as Semi-Supervised Learning (SSL) and as SSFS, since labeled and unlabeled data are used for both model and feature selection. The final classifier selected is the one that maximizes the RI of both the labeled and unlabeled data and the features selected are those that yield the highest RI.

In this paper, the general idea of the method is presented and the obtained results are very consistent as discussed at the end. The general organization of the paper is as follows. Section 2 deals with the Semi-supervised Learning and Feature Selection; Section 3 aboard the Multi-objective learning and section 3.1 presents the proposed method. After that the results are shown and a discussion and conclusions take place.

2 Semi-Supervised Learning

In supervised learning the methods need labeled data for training, however, labeling can be difficult, expensive and time consuming.The reason for that lies in the frequent requirement of specific human experience efforts to label patterns. In contrast, unlabeled data can be easy to obtain. In order to handle both types of information, there are many SSL methods in the literature, however, most

algorithms are based on a pre-established assumption about the unlabeled data, such as data set contiguity or low density in the margin region [6, Introduction] [2,9]. The assumption usually imposes a strong bias in the kind of solutions that may be achieved by the algorithm, although it is an important principle to compensate the missing labeling information.

Recent works in the field can be mentioned like the one in [23] where labeled and unlabeled data are integrated using the clustering structure of unlabeled data as well as the smoothness structure of the estimated class priors. In [15] the authors combined transductive inference with the *Multi-relational data mining* (MRDM) classification. Other interesting work is presented in [20], where authors applied transductive learning to K-Nearest Neighbors (KNN). A long list of references in the area can be found in [13].

The algorithm described in this paper is based on the separability assumption between classes. The decision making procedure is based on a relevance index for features that estimates separability. A restricted set of Pareto-Optimal [5] solutions is obtained from the labeled data and a decision making procedure is accomplished in order to select the one that maximizes the relevance index over labeled and unlabeled data.

2.1 Semi-Supervised Feature Selection

Semi-supervised feature selection is based on the same principles of SSL. The goal is, therefore, to select features in the framework of a very small number of labeled data and a large number of unlabeled samples. It is clear, however, that feature selection does not depend uniquely on labeled data, since redundancy elimination methods can be applied to the whole dataset regardless of any existing labels [16]. Nevertheless, in order to estimate a relevance index for features, a quantitative measure of how an individual feature or a group of features discriminates the likelihood of classes, should be considered. Fischer Linear Discriminant [8] and Relief [11] are examples of such Supervised Feature Selection approaches. In the absence of labels, one may search for some structural information in the data in order to accomplish Unsupervised Feature Selection [12,16]. The use of information coming from both sources is the goal of Semi-Supervised Feature Selection (SSFS), which has been the subject of many recent publications [24,21,3].

3 Multi-Objective Learning

It is well known that learning algorithms that are based only on error minimization do not guarantee good generalization performance models. In addition to the training set error, some other network-related parameters should be adapted in the learning phase in order to control generalization performance. The need for more than a single objective function paves the way for treating the supervised learning problem with multi-objective optimization (MOBJ) techniques [19].

Usual approaches explicitly consider the two objectives of minimizing the sum of squares error and the norm of the weight vectors. The learning task is carried on by minimizing both objectives simultaneously, using vector optimization methods. This leads to a set of solutions that is called the Pareto-optimal set [5], from which the best network for modeling the data is selected. Finding the Pareto-optimal set can be interpreted as a way for reducing the search space to a one-dimensional set of candidate solutions, from which the best one is to be chosen. This one-dimensional set exactly follows a trade-off direction between flexibility and rigidity, which means that it can be used for reaching a suitable compromise solution [19].

The decision-making strategy from the Pareto-set is clearly described by a third objective function, such as validation error or separation margin, that also needs to be optimized. The choice of the decision-making objective function defines the kind of solution that one aims to obtain. The strategy described in this paper aims at selecting the solution that maximizes the separability of classes, measured by the Pearson's Correlation Coefficient [18], as will be described in the next section.

3.1 Multi-Objective Semi-Supervised Feature Selection (MOBJ-SSFS)

In general, the MOBJ learning problem can be defined according to the Equation 2

$$w^* = arg\min \frac{1}{n}\sum_{k=1}^{n}\left(d_k - y\left(w, x_k\right)\right)^2 \qquad (1)$$
$$Subject\ to\ :\ \|w\| \leq \lambda_i$$

where w and w^* are respectively the weight and the optimal weight vectors, n is the number of observations, d_k is the expected class label of observation k, y_k is the class label found by the neural network and λ_i is the norm constraint value.

Basically, what we do is the following:

- train a Multi-Layer Perceptron with the labeled set D_L, using the ellipsoid method [25] to solve the MOBJ problem described above, for different values of λ_i. This procedure will generate a Pareto-optimal set of solutions, each one representing a different classifier;
- for each different classifier:
  - Label D_U;
  - Calculate Pearson's RI for all features of set $D_L \cup D_U$;
- Select the best solution according to one of two strategies discussed ahead

For each solution in the Pareto-set a ranking of features is obtained. The interpretation of the feature ranking information is accomplished in such a way that the solution that yields a better class separation according to Pearson's correlation coefficient is selected.

We would like to select the solution that maximizes RI, however, there is no guarantee that there is a single solution that jointly maximizes RI for all features, so the solution selected is the one that maximizes the majority of features (Strategy 1). In addition to selecting a solution from the Pareto-set, this strategy also comes-up with a ranking of features, that is obtained considering $D_L \cup D_U$ and classifier's performance resulted from MOBJ learning. SSFS can now be accomplished by taking into consideration the resulting ranking of features. An alternative strategy is to select the solution that maximizes the most relevant features among all the Pareto-set solutions (Strategy 2).

4 Results and Discussions

A Multi-Layer-Perceptron (MLP) Neural Network (NN) was trained with the MOBJ algorithm described in the previous sections for the Wisconsin Breast Cancer data from the UCI repository. The data set has 683 samples (patients) with 9 features each one. In order to observe the approach for different proportions of labeled and unlabeled data, the model was trained 30 times for different values of ($\rho = N_L/N_U$), as can be seen in Figures 1 and 2. The results are presented for the two selection strategies (Strategies 1 and 2) described in the previous paragraph. A benchmark result is also presented in the two graphs for comparison purposes. It is always the lowest curve in the graph (smaller error) and was obtained by selecting each solution from the prior knowledge of the correct labels of all patterns.

Figure 2 shows the absolute classification errors for different features sets. The set composed by features 2,3 and 6 ($S_1 = \{F_2, F_3, F_6\}$) was selected by the MOBJ-SSFS method. The set $S_2 = \{F_1, F_4, F_5, F_6, F_8\}$ was selected by FS-redundancy method [16] and set $S_3 = \{F_1, F_3, F_6, F_8\}$ was selected by RELIEF [10].

The proposed method has an interesting property: while it performs feature selection it also yields Pareto-set selection, i.e. one can use this method also as a decision-making strategy in a MOBJ learning. As expected, the larger the labeled set size the smaller the classification error. The classification errors of networks trained with features selected by recurrence in the highest Person's index positions, and trained networks chosen by Maximum Pearson criteria, are close and have similar performance for different values of ρ. In other words, training the MLP with feature set $S_1 = \{F_2, F_3, F_6\}$ leads to results very close to the benchmark, regardless of the proportion between labeled and unlabeled data. It's also interesting to notice that the average final classification errors with the reduced data set S_1 has lower variance than the obtained with Strategy 1.

Relief RI [10] was also calculated for each possible combination of features subsets in order to identify which are the features that in presence of other combinations of features are always well ranked. The results are very consistent with MOBJ-SSFS. Subset S_1 is well ranked, i.e. even in the presence of one or more features they receive the highest relief's indexes. Depending on Relief's parameter k other subset of features stands (S_3), and it's considered in our

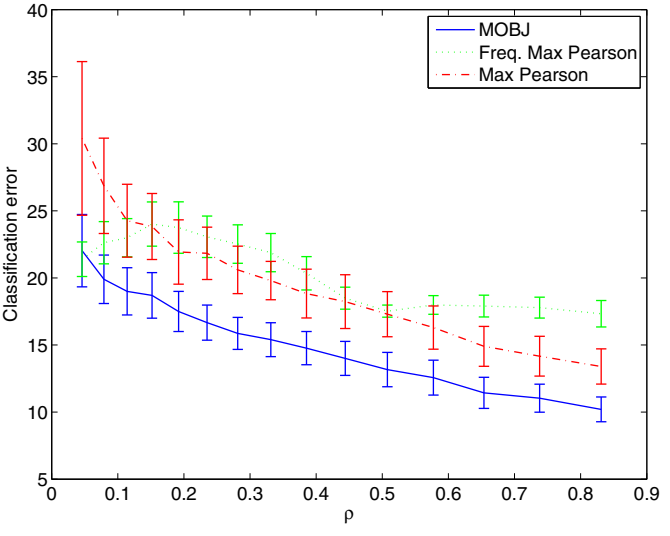

Fig. 1. The solid curve shows the real absolute classification error for the entire data set (samples and features) after training MLP with each amount of N_L defined by ρ. The dash dot one shows the error obtained by solution whose majority of features reaches max Pearson's index. The dot curve shows the error after training MLP using only features 6,2 and 3 that mostly has the three higher indexes in each solution of Pareto.

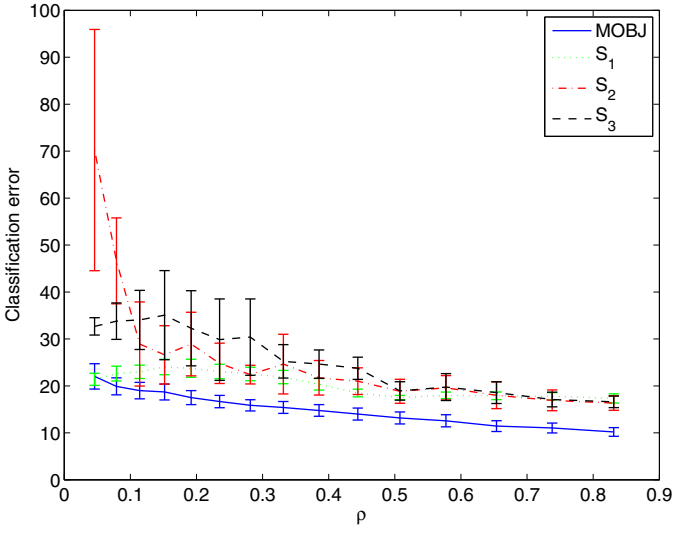

Fig. 2. Set's comparison

results. However, the subset S_1 of features perform better. Features 2,3 and 6 in S_1 are respectively *uniformity of cell size*, *uniformity of cell shape* and *Bare nuclei*. Features 1 and 8 are *Clump Thickness* and *Normal Nucleoli*.

The method proposed here should not be regarded as a wrapper nor embedded method since it does not use directly the classification results to select model parameters. A MLP was used, although other possible approaches like Support Vector Machines can be applied. The same holds for the rank *metric* used. Here we applied the *Pearson's correlation coefficient* because of its simplicity and because it is simpler to manipulate as an univariate method, but other metrics like *Relief* or *Fischer score* can also be applied.

5 Conclusions

The general concepts of a new SSFS method was presented. The results indicated that the selected features are consistent leading to coherent results for model selection. One interesting issue is that even for small values of ρ, the method was capable to select the feature subset S_1 leading to good classification results and with good stability, when compared to other subsets and even when compared with the results considering all features together. Finally, the ability of choosing one solution from MOBJ Pareto-set, when performing feature selection, is an interesting characteristic of the presented method, since it integrates model selection and feature selection under the frameworks of semi-supervised learning and statistical learning theory.

References

1. Niyogi, P., Belkin, M.: Semi-supervised learning on riemannian manifolds. Machine Learning 56, 209–239 (2004)
2. Coelho, F., de Braga, A.P., Natowicz, R., Rouzier, R.: Semi-supervised model applied to the prediction of the response to preoperative chemotherapy for breast cancer. In: Soft Computing - A Fusion of Foundations, Methodologies and Applications (July 2010)
3. Le Cun, Y., Denker, J.S., Solla, S.A.: Optimal brain damage. In: Advances in Neural Information Processing Systems, pp. 598–605. Morgan Kaufmann, San Francisco (1990)
4. Cover, T., Hart, P.: Nearest neighbor pattern classification. IEEE Transactions on Information Theory 13(1), 21–27 (1967)
5. Chankong, V., Haimes, Y.Y.: Multiobjective Decision Making Theory and Methodology. Elsevier Science, New York (1983)
6. Chapelle, O., Schölkopf, B., Zien, A. (eds.): Semi-Supervised Learning. MIT Press, Cambridge (2006)
7. Dy, J.G., Brodley, C.E.: Feature selection for unsupervised learning. J. Mach. Learn. Res. 5, 845–889 (2004)
8. Fisher, R.A.: The use of multiple measurements in taxonomic problems. Annals Eugen. 7, 179–188 (1936)

9. Kasabov, N., Pang, S.: Transductive support vector machines and applications in bioinformatics for promoter recognition. In: Proc. of International Conference on Neural Network & Signal Processing, Nangjing. IEEE Press, Los Alamitos (2004)
10. Kira, K., Rendell, L.A.: The feature selection problem: Traditional methods and a new algorithm. In: AAAI, Cambridge, MA, USA, pp. 129–134. AAAI Press and MIT Press (1992)
11. Kira, K., Rendell, L.A.: A practical approach to feature selection. In: ML 1992: Proc. of the Ninth International Workshop on Machine Learning, pp. 249–256. Morgan Kaufmann Publishers Inc., San Francisco (1992)
12. Kruskal, J., Wish, M.: Multidimensional Scaling. Sage Publications, Thousand Oaks (1978)
13. Liang, F., Mukherjee, S., West, M.: The use of unlabeled data in predictive modeling. Statistical Science 22, 189 (2007)
14. Lawler, E.L., Wood, D.E.: Branch-and-bound methods: A survey. Operations Research 14(4), 699–719 (1966)
15. Malerba, D., Ceci, M., Appice, A.: A relational approach to probabilistic classification in a transductive setting. Eng. Appl. Artif. Intell. 22(1), 109–116 (2009)
16. Mitra, P., Murthy, C.A., Pal, S.K.: Unsupervised feature selection using feature similarity. IEEE Trans. Pattern Anal. Mach. Intell. 24(3), 301–312 (2002)
17. Parma, G.G., Menezes, B.R., Braga, A.P., Costa, M.A.: Sliding mode neural network control of an induction motor drive. Int. Jour. of Adap. Cont. and Sig. Proc. 17(6), 501–508 (2003)
18. Press, W.H., Teukolsky, S.A., Vetterling, W.T.: Numerical recipes in C (2nd ed.): the art of scientific computing. Cambridge University Press, New York (1992)
19. Takahashi, R.H.C., Teixeira, R.A., Braga, A.P., Saldanha, R.R.: Improving generalization of MLPs with multi-objective optimization. Neurocomputing 35(1-4), 189–194 (2000)
20. Wu, J., Yu, L., Meng, W., Shu, L.: Kernel-based transductive learning with nearest neighbors. In: Li, Q., Feng, L., Pei, J., Wang, S.X., Zhou, X., Zhu, Q.-M. (eds.) APWeb/WAIM 2009. LNCS(LNAI), vol. 5446, pp. 345–356. Springer, Heidelberg (2009)
21. Tenenbaum, J.B., de Silva, V., Langford, J.C.: A Global Geometric Framework for Nonlinear Dimensionality Reduction. Science 290(5500), 2319–2323 (2000)
22. Vapnik, V.N.: The nature of statistical learning theory. Springer, New York (1995)
23. Wang, J., Shen, X., Pan, W.: On efficient large margin semisupervised learning: Method and theory. J. Mach. Learn. Res. 10, 719–742 (2009)
24. Zhang, D., Zhou, Z.-h., Chen, S.: Semi-Supervised Dimensionality Reduction. In: SIAM Conference on Data Mining (SDM), pp. 629–634 (2007)
25. Bland, R.G., Goldfarb, D., Todd, M.J.: The Ellipsoid Method: A Survey. Operations Research 29(6), 1039–1091 (1980)