ESANN 2013 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2013, i6doc.com publ., ISBN 978-2-87419-081-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100131010.

GA-KDE-Bayes: An Evolutionary Wrapper Method Based on Non-Parametric Density Estimation Applied to Bioinformatics Problems

Maria Fernanda Wanderley[1] and Vincent Gardeux[2] and
René Natowicz[3] and Antônio P. Braga[1] *

1- Graduate Program in Electrical Engineering - Federal University of Minas Gerais
Av. Antônio Carlos 6627, 31270-901, Belo Horizonte, MG, Brazil

2- L@RIS - EISTI
Avenue du Parc, Cergy, France

3- ESIEE-Paris - University of Paris-Est
Noisy-le-Grand, France

Abstract. This paper presents an evolutionary wrapper method for feature selection that uses a non-parametric density estimation method and a Bayesian Classifier. Non-parametric methods are a good alternative for scarce and sparse data, as in Bioinformatics problems, since they do not make any assumptions about its structure and all the information come from data itself. Results show that local modeling provides small and relevant subsets of features when comparing to results available on literature.

1 Introduction

Density estimation with parametric models are based on the principle that data has fixed structure and that global model parameters can be induced from sampled data. However, when data is scarce and sparse, such a global assumption may lead to biased estimators that are not capable of inducing a reliable general model. An alternative to overcome this conflict between the global target and the lack of information, is to use non-parametric estimators, which do not rely on a preestablished structure and that are based on localized models in order to construct a more general representation of the underlying problem.

Generative models for classification and feature selection depend on the validity of the induction principle adopted and on the representativeness of the sampled data. This seems to be a dilemma in most function induction problems, since usually a general function is aimed, but in most real problems the dataset is not large enough to guarantee global convergence conditions. Adopting localized non-parametric models does not overcome the dilemma, but provides a more realistic assumption under the adverse conditions.

Wrapper methods for feature selection require a dataset model in order to accomplish the selection task, therefore the robustness of the selected features depend on the chosen wrapper model. In this scenario, Bioinformatics problems are particularly challenging, since the number of confirmed cases to provide a

*This work were supported by CNPq/BR.

ESANN 2013 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2013, i6doc.com publ., ISBN 978-2-87419-081-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100131010.

representative dataset is usually not large enough to guarantee global convergence. In order to compensate the lack of inductive samples, more sophisticated models to control structural and effective model capacities need to be developed.

In this paper we present an evolutionary wrapper method for feature selection that is based on Kernel Density Estimation(KDE) [1], which is non-parametric estimator, and on Bayes Classification [2]. The method is consistent with the capacity control principle since its smoothness is selected according to the dataset performance. Results indicate that the evolutionary method provide representative small subsets of features which performs better than discriminant functions.

2 Proposed Method

2.1 Kernel Density Estimation(KDE)

A kernel estimator of kernel K is defined by the following expression [3]:

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K(\frac{x - X_i}{h}) \tag{1}$$

where h is the window width, $X_1, X_2, \ldots, X_n$ are independent and identically distributed (iid) samples from a random variable and K is a kernel function that needs to satisfy the condition from Eq. 2, being the gaussian function the most frequently used.

$$\int_{-\infty}^{\infty} K(x)dx = 1 \tag{2}$$

The decision of the window width h takes an important role on the density estimation, since this parameter defines the smoothness of the estimated density function. In this work we used the value of h presented on [3], which proposes to balance the trade-off between bias and variance of the asymptotic mean integrated squared error of the estimation, given by:

$$h = 1.06 * \sigma * n^{-\frac{1}{5}}. \tag{3}$$

2.2 Multidimensional Kernel Density Estimation

Let $\mathbf{X} = (x_1, \ldots, x_d)^T$ be a vector of random variables of d dimensions. The objective is to estimate the joint PDF of random variables $x_1, \ldots, x_d$: $f(x) = f(x_1, \ldots, x_d)$. The multidimensional kernel estimator can be seen as an expansion from the univariate case (subsection 2.1):

$$\hat{f}_h(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h_1 * \ldots * h_d} K\left(\frac{x_1 - x_{i1}}{h_1}, \ldots, \frac{x_d - x_{id}}{h_d}\right). \tag{4}$$

In this case, the kernel K is a multidimensional function which also need to satisfy the property from Eq.2 and, given that the value of h is always the same, it is assumed that the data is equally distributed in all dimensions [3].

ESANN 2013 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2013, i6doc.com publ., ISBN 978-2-87419-081-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100131010.

An alternative for a multivariable kernel function is the multiplicative kernel [3]. Here, a univariable kernel function is used for each one of the dimensions with its own width h. Re-writing the equation 4 using the multiplicative kernel we have:

$$\hat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n}\left\{\prod_{j=1}^{d}\frac{1}{h_j}K\left(\frac{x_j - x_{ij}}{h_j}\right)\right\} \tag{5}$$

2.3 KDE-Bayes

The idea behind this method is to perform a local modeling of data by using a non-parametric estimator and then separate the classes with a bayesian classifier. Kernel estimators provide local inferences by using each input data to accomplish the estimation and, from this information, obtain some knowledge about the global relationships.

KDE-Bayes [4] is accomplished by non-parametric density estimation of likelihoods followed by data classification with a bayesian rule. The decision rule of a binary bayesian classifier [2] indicates that the feature vector x belongs to class C_i according to the posterior probability, $P(C_i|x)$. For a two class problem, such rule can be described by:

$$Class(x) = \begin{cases} C_1 & \text{if } P(x|C_1)P(C_1) > P(x|C_2)P(C_2), \\ C_2 & \text{otherwise,} \end{cases} \tag{6}$$

where $P(x|C_1)$ and $P(x|C_2)$ are the likelihood for the classes with respect to the vector x and $P(C_1)$ e $P(C_2)$ are the a priori probabilities of each class.

2.4 GA-KDE-Bayes: Selecting subsets of features

Given the high dimensionality of the problems presented in this work, methods that search all combinations within the search space can not be applied. To address this issue we have chosen an evolutionary method, in order to search the space more efficiently and to select proper subsets of features.

For this work, initially a population of subsets of features, encoded as decimal numbers, is randomly generated. Within generations the following steps occurs: at first, individuals are selected according to a fitness function, by using a roulette-wheel, then, given a probability, the individuals suffer a mutation, that can increase or decrease, by one, the number of features on it. Either if the individual is going to have its size increased or decreased, which is defined by a probability, the feature that is going to be removed is chosen randomly. In the end of an iteration a new population is created and the algorithm continues until the stop condition [5].

To measure the fitness of a subset of features we used two metrics, sensitivity and specificity, given by the performance of the KDE-Bayes for the individual being evaluated. In order to have a balance of both of them we used as fitness

ESANN 2013 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2013, i6doc.com publ., ISBN 978-2-87419-081-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100131010.

function f the geometric mean of the metrics of the classifier, sensitivity and specificity,

$$f = \sqrt{se * sp}.$$

The value of f becomes large when both metrics are large and when the difference between them is small [6]. As so, the best individuals are those which have balanced metrics, while individuals that are biased for one class or another receive a small fitness function value.

Differently of a classical genetic algorithm, here the crossover was not used, since the order of the features within an individual doesn't change the fitness value. Another important operator used was the elitism, which keeps the best (of some of them) individual for the next generation, preventing that its loss during the process of mutation.

3 Results

3.1 Oncology Datasets

In order to assess the performance of the method proposed in this work, six publicly available datasets in oncology were chosen: Colon (`http://genomics-pubs.princeton.edu/oncology/affydata/index.html`), Lymphoma (`http://www.gems-system.org/`), Leukemia, Brain (the previous two from `http://www.broadinstitute.org/cgi-bin/cancer/datasets.cgi`), Prostate (`http://www.gems-system.org/`), and Ovarian (`http://data.cgt.duke.edu/clinicalcancerresearch.php`). Those datasets have between 2000 and 22283 features and less than 103 samples. As so, they are good candidates for feature selection.

We compared the results from GA-KDE-Bayes with two other selection methods: δ-test, a filter method based on the optimization of a bi-objective function that aims to maximize the interclass distance and minimize the number of features, and ABEUS, a wrapper based on the optimization of the performance of a classifier [7]. The GA-KDE-Bayes parameters were: 150 individuals, 100 generations, 0.7 as probability of mutation and 1 individual is kept by the elitism. The mutation probability used is higher than usual because we wanted to have a high number of size changes at the individual.

On Table 1, we show the results for each dataset. In general, GA-KDE-Bayes selects a smaller number of features and have similar or better performance when comparing with δ-test and ABEUS methods.

3.2 Neoadjuvant Chemotherapy on Breast Cancer Patients

After assessing the performance of GA-KDE-Bayes on oncology datasets, we used gene expression data from breast cancer patients that had been treated with neoadjuvant chemotherapy. This database is composed by 133 patients from a clinical trial made at Nellie B. Connaly Breast Center from M.D. Anderson Cancer Center, from Texas University [8]. The patients are divided on those with complete response to the chemotherapy (PCR) and those who had not

ESANN 2013 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2013, i6doc.com publ., ISBN 978-2-87419-081-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100131010.

Table 1: Three-fold cross-validation of the GA-KDE-Bayes, δ-KDE-Bayes e ABEUS-KDE-Bayes. Ac = accuracy, Se = sensitivity, Sp = Specificity, PPV, NPV: positive and negative predictive values.

data	colon	lymphoma	leukemia	prostate	brain	ovaries
GA-KDE-Bayes						
#features	6.9 ±0.11	6.5 ±0.04	8.8 ±0.17	6.5 ± 0.05	5.9 ±0.05	8.8 ±0.14
Ac	0.80 ±0.0	0.83 ±0.0	0.86 ±0.0	0.80 ±0.0	0.65 ±0.01	0.74 ±0.01
Se	0.83 ±0.0	0.72 ±0.01	0.80 ±0.01	0.80 ±0.0	0.63 ±0.02	0.73 ±0.01
Sp	0.76 ±0.01	0.87 ±0.0	0.90 ±0.0	0.81 ±0.01	0.65 ±0.02	0.75 ±0.01
PPV	0.86 ±0.0	0.64 ±0.0	0.80 ±0.01	0.84 ±0.0	0.48 ±0.01	0.65 ±0.01
NPV	0.705 ±0.01	0.90 ±0.0	0.90 ±0.00	0.76 ±0.01	0.78 ±0.01	0.79 ±0.01
δ-KDE-Bayes						
#features	13.1 ± 6.72	4.2 ± 1.71	3.8 ± 1.77	7.0 ± 4.47	17.8 ± 5.55	10.0 ± 3.6
Ac	0.81 ± 0.08	0.86 ± 0.1	0.95 ± 0.0	0.90 ± 0.0	0.66 ± 0.1	0.67 ± 0.1
Se	0.88 ± 0.1	0.80 ± 0.2	0.89 ± 0.1	0.90 ± 0.1	0.47 ± 0.24	0.6 ± 0.2
Sp	0.68 ± 0.2	0.88 ± 0.1	0.98 ± 0.0	0.91 ± 0.1	0.76 ± 0.2	0.71 ± 0.2
PPV	0.84 ± 0.1	0.70 ± 0.1	0.97 ± 0.1	0.91 ± 0.1	0.62 ± 0.2	0.68 ± 0.2
NPV	0.77 ± 0.2	0.94 ± 0.1	0.95 ± 0.1	0.91 ± 0.1	0.73 ± 0.1	0.72 ± 0.1
ABEUS-KDE-Bayes						
#features	6.8 ± 4.2	4.1 ± 1.2	3.2 ± 0.8	12.5 ± 9.44	14.8 ± 6.3	4.6 ± 1
Ac	0.78 ± 0.1	0.86 ± 0.1	0.89 ± 0.1	0.85 ± 0.1	0.61 ± 0.1	0.62 ± 0.1
Se	0.84 ± 0.1	0.81 ± 0.2	0.86 ± 0.1	0.83 ± 0.1	0.40 ± 0.1	0.59 ± 0.2
Sp	0.65 ± 0.2	0.88 ± 0.1	0.91 ± 0.1	0.87 ± 0.1	0.73 ± 0.1	0.64 ± 0.2
PPV	0.83 ± 0.1	0.71 ± 0.2	0.84 ± 0.1	0.86 ± 0.1	0.47 ± 0.2	0.59 ± 0.1
NPV	0.71 ± 0.13	0.94 ± 0.0	0.93 ± 0.05	0.85 ± 0.1	0.69 ± 0.1	0.66 ± 0.1

(NoPCR). We splited the data in training set, composed by 82 cases (61 PCR cases and 21 NoPCR cases) and test set, composed by 51 patients (38 PCR cases and 13 NoPCR cases).

On Table 2 we compare the results from GA-KDE-Bayes, DLDA-30 [8] and clinical predictors. Even with less than a third of the features used by DLDA-30, GA-KDE-Bayes performs around 10% better than the former one. When comparing to the Clinical Predictors GA-KDE-Bayes also performs better but using three times more features.

Table 2: Performances of the predictors on the independent test dataset: Villejuif (51 cases). Clinical Predictors are based on age, estrogen receptor status and nuclear grade.PPV, NPV: positive and negative predictive values.

	GA-KDE-Bayes	DLDA-30	Clinical Predictors
#features	9	30	3
Accuracy	**0.86**	0.76	0.78
Sensitivity	**0.92**	**0.92**	0.61
Specificity	**0.84**	0.71	**0.84**
PPV	0.67	0.52	0.57
NPV	0.97	0.96	0.86

ESANN 2013 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2013, i6doc.com publ., ISBN 978-2-87419-081-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100131010.

4 Conclusion

This paper presented a wrapper feature selection method based on KDE and Bayes classification. Features were selected according to an evolutionary method with fitness function based on geometric mean which reward individuals that have both sensitivity and specificity metrics balanced. As can be seen on the results on Table 2 the method presented better performance than others presented on the literature.

The results suggest that bottom-up non-parametric methods can be particularly important in application domains such as Bioinformatics, that usually have high dimensional scarce datasets.

References

[1] E. Parzen. On estimation of a probability density function and mode. *The annals of mathematical statistics*, 33(3):1065–1076, 1962.

[2] Richard O. Duda, Peter E. Hart, and David G. Stork. *Pattern Classification*. Wiley-Interscience, 2 edition, 2000.

[3] B.W. Silverman. Density estimation for statistics and data analysis. *Monographs on Statistics and Applied Probability*, 1986.

[4] MFB Wanderley, AP Braga, EMAM Mendes, R Natowicz, and R Rouzier. Non-parametric kernel density estimation for the prediction of neoadjuvant chemotherapy outcomes. In *Proceedings of 32nd Annual International Conference of the IEEE EMBS (EMBC'10)*, 2010.

[5] David E. Goldberg. *Genetic Algorithms in Search, Optimization and learning*. Addison-Wesley, Massachusetts, 1989.

[6] M. Kubat, R. Holte, and S. Matwin. Learning when negative examples abound. *Machine Learning: ECML-97*, pages 146–153, 1997.

[7] V. Gardeux, R. Natowicz, M.F.B. Wanderley, and R. Chelouah. Optimization for feature selection in dna microarrays. In Patrick Siarry, editor, *Heuristics: Theory and Applications*. Nova Publishers, 2013. In publication.

[8] KR Hess, K Anderson, WF Symmans, et al. Pharmacogenomic predictor of sensitivity to preoperative chemotherapy with paclitaxel and fluorouracil, doxorubicin, and cyclophosphamide in breast cancer. *Journal of Clinical Oncology*, 24(26):4236–4244, 2006.