Hierarchical kernel applied to mixture model for the classification of binary predictors

Seydou Sylla, Stephane Girard, Abdou Kâ Diongue, Aldiouma Diallo, Cheikh Sokhna

Hierarchical kernel applied to mixture model for the classification of binary predictors

Seydou Nourou SYLLA*
Université Gaston Berger,Saint-Louis, Sénégal, nourou01@yahoo.fr

Stéphane Girard
INRIA & LJK, Grenoble, France, stephane.girard@inria.fr
Abdou Ka Diongue
LERSTAD-Université Gaston Berger, Saint-Louis, Sénégal,kadiongue@yahoo.fr
Aldiouma Diallo
URMITE-IRD, Dakar, Sénégal, aldiouma.diallo@ird.fr
Cheikh Sokhna
URMITE-IRD, Dakar, Sénégal, cheikh.sokhna@ird.fr

Abstract

Diagnosis systems often use structured data. These data have a hierarchical structure related with the questions asked during the interview with the doctor or the survey taker in charge of verbal autopsies. The hierarchical nature of these questions leads to consider this aspect when analyzing medical data. Thus, it is recommendable to choose a similarity measure that takes into account this issue to better represent the reality. We propose the introduction of a kernel taking into account the hierarchical structure and of the data interactions between sub-items in supervised binary classification methods This kernel can integrate the knowledge from the application domain relative to how the features of the problem are organized. In general, we focus on problems whose features can be hierarchically structured. As part of this work, these hierarchies are represented by trees on two levels. Our main contribution is the proposal of a kernel that simultaneously takes into account the hierarchical appearance and the interaction between variables. The proposed kernel has shown a good classification performance on a complex set of medical data including a high number of predictors and classes.

Keywords: hierarchical kernel; Diagnosis systems; classification; similarity..

1 Introduction

The diagnostic methods often use structured data. These data has a hierarchical structure at the questions asked during the interview with the doctor or the investigator in case of verbal autopsies. The hierarchical aspect of the questions asked during the interview needs to be considered when analyzing the medical data. Thus, it is advisable to choose a similarity measure taking into account this aspect in order to better represent reality. The hierarchical classification structure allows taking into account the a priori knowledge on the data. The a priori is represented by a structure on the variables or characteristics of the data set.
To take into account the structure of data, an overall kernel is used on heterogeneous types (video, sound, text, ...) or hierarchical struct (items, sub-items, ...). Each set of these features may require a different kernel. So, to build a global kernel, it is possible to set a kernel for each of these features and combine them linearly or multiplicatively.
In this article, we propose the use a hierarchical kernel binary data. It outlines the introduction of a kernel taking into account the hierarchical structure and interactions between sub-items in the supervised classification methods. This kernel can integrate knowledge in the domain of application. This knowledge is relative to how the characteristics of the problem are organized. In general, we focus on problems whose

characteristics can be structured hierarchically. As part of this work, these hierarchies are represented by trees on two levels.
The article is organized as follows. In paragraph 3 , we first describe the different stages of the process that leads to the choice of this kernel. We show that the resulting formulation is suitable for trees on two levels and the interaction of sub-variables. Through these two formulations, we study the characteristics of defined nucleus and characterize the relationship between the levels of this tree. The performance of the new classification method is illustrated on verbal autopsy data in section 4. Concluding remarks are provided in section 5.

2 Binary classification using a kernel function

We focus on the new classification method, referred to as pgpDA, has been proposed [1]
The principle of pgpDA is as follows. Let us consider a learning set $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ where $\{x_1, \ldots, x_n\}$ are assumed to be independent realizations of a random binary vector $X \in \{0,1\}^p$. The class labels $\{y_1, \ldots, y_n\}$ are assumed to be realizations of a discrete random variable $Y \in \{1, \ldots, K\}$. It indicates the memberships of the learning data to the L classes denoted by $C_1, \ldots, C_K$, *i.e.* $y_i = k$ means that x_i belongs to the kth cluster C_k for all $i \in \{1, \ldots, n\}$ and $k \in \{1, \ldots, K\}$.
Let κ be a symmetric non-negative bivariate function $\kappa : \{0,1\}^p \times \{0,1\}^p \to \mathbb{R}^+$. In the following, κ is referred to as a kernel function and additional conditions will be assumed on κ. For all $k = 1, \ldots, K$, the function $\rho_k : \{0,1\}^p \times \{0,1\}^p \to \mathbb{R}^+$ is obtained by centering the kernel κ with respect to the class C_k:

$$\rho_k(x, x') = \kappa(x, x') - \frac{1}{n_k} \sum_{x_\ell \in C_k} (\kappa(x_\ell, x') + \kappa(x, x_\ell)) + \frac{1}{n_k^2} \sum_{x_\ell, x_{\ell'} \in C_k} \kappa(x_\ell, x_{\ell'}),$$

where n_k is the cardinality of the class C_k, *i.e.* $n_k = \sum_{i=1}^n \mathbb{I}\{y_i = k\}$ and with $\mathbb{I}\{.\}$ the indicator function. Besides, for all $k = 1, \ldots, K$, let us introduce the $n_k \times n_k$ symmetric matrix M_k defined by $(M_k)_{\ell,\ell'} = \rho_k(x_\ell, x_{\ell'})/n_k$ for all $(\ell, \ell') \in \{1, \ldots, n_k\}^2$. The sorted eigenvalues of M_k are denoted by $\lambda_{k1} \geq \cdots \geq \lambda_{kn_k}$ while the associated (normed) eigenvectors are denoted by $\beta_{k1}, \ldots, \beta_{kn_k}$. In the following, $\beta_{kj\ell}$ represents the ℓth cordinate of β_{kj}, for $(j, \ell) \in \{1, \ldots, n_k\}^2$. The classification rule introduced in [1], Proposition 2 affects $x \in \{0,1\}^p$ to the class C_i if and only if $i = \arg\min_{k=1,\ldots,K} D_k(x)$ with

$$\begin{aligned} D_k(x) = & \frac{1}{n_k} \sum_{j=1}^{d_k} \frac{1}{\lambda_{kj}} \left(\frac{1}{\lambda_{kj}} - \frac{1}{\lambda} \right) \left(\sum_{x_\ell \in C_k} \beta_{kj\ell} \rho_k(x, x_\ell) \right)^2 + \frac{1}{\lambda} \rho_k(x, x) \\ & + \sum_{j=1}^{d_k} \log(\lambda_{kj}) + (d_{\max} - d_k) \log(\lambda) - 2 \log(n_k) \end{aligned} \quad (1)$$

where $d_{\max} = \max\{d_1, \ldots, d_K\}$ and

$$\lambda = \sum_{k=1}^K n_k (\text{trace}(M_k) - \sum_{j=1}^{d_k} \lambda_{kj}) \Big/ \sum_{k=1}^K n_k (r_k - d_k) \, .$$

Here, r_k is the dimension of class C_k once mapped in a nonlinear space with the kernel κ. In practice, one has $r_k = \min(n_k, p)$ for a linear kernel and $r_k = n_k$ for the nonlinear kernels. See [1], Table 2 for further examples. Moreover, let us highlight that only the eigenvectors associated with the d_k largest eigenvalues of M_k have to be estimated. This property is a consequence of the crucial assumption of this method: The data of each class C_k live in a specific subspace (of dimension d_k) of the space (of dimension r_k) defined by the kernel κ. This assumption allows to circumvent the unstable inversion of the matrices M_k, $k = 1, \ldots, K$ which is usually necessary in kernelized versions of Gaussian mixture models, see for instance [2]. In practice, d_k is estimated thanks to the scree-test of Cattell [3] which looks for a break in the eigenvalues scree. The selected dimension is the one for which the subsequent eigenvalues differences are smaller than a threshold t. The threshold t can be provided by the user or selected by cross-validation.The implementation of this method requires the selection of a kernel function κ which measures the similarity between two binary vectors.

3 Hierarchical kernel associated with binary observations

3.1 Structure Data and notations:

In a survey, there are often issues called main. For each main issue, there are issues called secondary. Secondary questions are asked only if the answer to the main question is positive. By formalizing this concept, the variable X_j represents the answer to the main question j. For each given X_j there were q_j responses to secondary issues noted by the sub-variables $Z_1^j, \ldots, Z_{q_j}^j$.. Thus, referring to the case of verbal autopsy data included:

- The randoms variables $X = (X_j, j = 1, \ldots, p)$ define the answers to the main questions representing the symptoms and the socio-demographic variables.
- The randoms variables $Z = (Z_\ell^j, \ell = 1, \ldots, q_j, j = 1, \ldots, p)$ define the answers to the secondary questions representing the q_j sub-variables for each variable X_j.
- The randoms variables $Y = (Y_k, k = 1, \ldots, K)$ define the explanatory variables representing the physician's answers (cause of death).

These hierarchies are represented by a two-level tree structure, as that shown in Figure 1 .
The first level represents the answers to main questions. The second level represents the sub-variables that is to say, the answers to the secondary issues of each main issue.

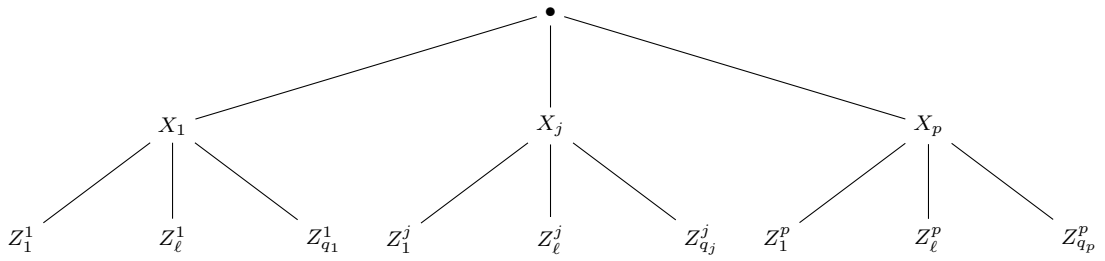

Figure 1: Example tree with two levels associated with the explanatory variables

In addition, the following lemma sets the relationship between the levels of the tree.

3.2 Hierarchical Kernel for binary data

The goal is to build a kernel takeing into account the hierarchical data structure and the interaction of sub-variables. We focus on issues where the explanatory variables of a data set can be structured in a tree. In this structure, the characteristics of sub-variables are located in the bottom level of the tree. The first level identifies the principal variables to which the sub-variables belong.
Below we propose a new kernel that takes into account interactions between variables. Interaction must have significant relevance and be capable of providing additional information on the diagnostic method to to improve the accuracy of the results.
Our principle is based on the transformation of the dissimilarity between two main variables X_j and $X_{j'}$ of a combination of dissimilarities between the main variables X_j and $X_{j'}$ and their respective sub-variables $Z_\ell^j, \ell = 1, \ldots, q_j$ and $Z_{\ell'}^{j'}, \ell' = 1, \ldots, q_j^{'}$.
For each variable X_j there were q_j sub-variables $Z_1^j, \ldots, Z_{q_j}^j$. In addition, we have:

$$X_j = \max\{Z_1^j, \ldots, Z_\ell^j\} \qquad = 1 - \prod_{\ell=1}^{q_j}(1 - Z_\ell^j) = \sum_{\ell=1}^{q_j}(-1)^{\ell-1} \sum_{k=1}^{\ell} \sum_{|i|=k} Z_{i_1} \ldots Z_{i_k}$$

where $|i| = k$ denotes the size of the multi-index $i = (i_1, \ldots, i_k)$.
Calculing $\|x - x'\|^2$ was :

$$\begin{aligned}\|x - x'\|^2 &= \sum_{j=1}^{p} \left[\prod_{\ell=1}^{q_j} (1 - z_\ell^j) - \prod_{\ell=1}^{q_j} (1 - z_\ell^{'j}) \right]^2 \\ &= \sum_{j=1}^{p} \sum_{\ell=1}^{q_j} \sum_{k=1}^{\ell} \sum_{|i|=k} s_{kji}^2 + R\end{aligned}$$

where $s_{kji} = \left(z_{i_1}^j \ldots z_{i_k}^j - z_{i_1}^{'j} \ldots z_{i_k}^{'j} \right)$ and R the sum of the double products.
By defining:

$$SC(z, z') = \sum_{j=1}^{p} \sum_{\ell=1}^{q_j} \sum_{k=1}^{\ell} \sum_{|i|=k} s_{kji}^2$$

there is therefore the decomposition

$$\|x - x'\|^2 = SC(z, z') + R.$$

A dissimilarity measure is defined for all $\gamma \in [0, 1]$ by:

$$D((x, z), (x', z')) = \gamma SC(z, z') + (1 - \gamma)R = (1 - \gamma)\|x - x'\|^2 + (2\gamma - 1)SC(z, z').$$

By asking:

- $D_x(x, x') = \|x - x'\|^2$,
- $D_z(z, z') = SC(z, z')$,

Previous dissimilarity measure can be rewritten:

$$D((x, z), (x', z')) = (1 - \gamma)D_x(x, x') + (2\gamma - 1)D_z(z, z').$$

Using the kernel construction method proposed [4], introducing the kernel:

$$\kappa_{\text{SGH}}((x, z), (x', z')) = \kappa_x(x, x')^{1-\gamma} \kappa_z(z, z')^{2\gamma - 1} \tag{2}$$

where,

- $\kappa_x(x, x') = \exp(-\|x - x'\|^2 / 2\sigma_x^2)$ is the RBF kernel,
- $\kappa_z(z, z') = \exp(-SC(z, z') / 2\sigma_r^2)$.

More generally in the kernel (2)

1) For the main variables X, we can choose a kernel of the form:

$$\kappa_x(x, x') = \exp\left(\frac{S(x, x')}{2\sigma_x^2} \right). \tag{3}$$

   where S is the similarity measure.
   This similarity measure S can be chosen by the measures defined in our formalism introduced in [4].

2) The interactions between sub-variables Z are considered to be of order r with the following kernel:

$$\kappa_z(z, z^{'}) = \exp\left(\frac{SC_{(r)}(z, z')}{2\sigma_r^2} \right) \tag{4}$$

   where

   (a) r the number of interactions,

(b) $SC_{(r)}$ in the truncated version of r on SC:

$$SC_{(r)}(z,z') = \sum_{j=1}^{p}\sum_{k=1}^{r}(q_j+1-k)\sum_{|i|=k} s_{kji}^2$$
$$= \sum_{j=1}^{p} sc_{(r,j)}$$

(c) $sc_{(r,j)}$ the interactions between r sub variables j defined by

$$sc_{(r,j)} = \sum_{k=1}^{r}(q_j+1-k)\sum_{|i|=k} s_{kji}^2$$
$$= \sum_{k=1}^{r}(q_j+1-k)\sum_{|i|=k}\left(z_{i_1}^j\ldots z_{i_k}^j - z_{i_1}^{'j}\ldots z_{i_k}^{'j}\right)^2$$

By combining 1) and 2) defines the hierarchical kernel of interactions of order r following:

$$\kappa_{\mathrm{SGH}}((x,z),(x',z')) = \kappa_x(x,x^{'})^{(1-\gamma)}\kappa_z(z,z^{'})^{(2\gamma-1)}$$

where κ_x give by (3) and κ_z par (4).
For some values of γ, it appears that the RBF kernel can be found for binary data in some cases.
If $\kappa_x = \kappa_{\mathrm{RBF}}$ then

- $\gamma = \frac{1}{2} \Rightarrow \kappa_{\mathrm{SGH}}\left((x,z),(x',z')\right) = \kappa_{\mathrm{RBF}}(x,x')$,
- $\gamma = 1$ et $r = 1 \Rightarrow \kappa_{\mathrm{SGH}}\left((x,z),(x',z')\right) = \kappa_{\mathrm{RBF}}(z,z^{'})$,
- $\gamma = \frac{2}{3}$ et $r = 1$
  $\Rightarrow \kappa_{\mathrm{SGH}}\left((x,z),(x',z')\right) = \kappa_{\mathrm{RBF}}((x\cup z),(x^{'}\cup z^{'}))$.

4 Experiments

4.1 Datasets

Verbal autopsy Data The goal of verbal autopsy is to get some information from family about the circumstances of a death when medical certification is incomplete or absent. In such a situation, verbal autopsy can be used as a routine death registration. A list of p possible symptoms is established and the collected data $X = (X_1,\ldots,X_p)$ consist of the absence or presence (encoded as 0 or 1) of each symptom on the deceased person. The probable cause of death is assigned by a physician and is encoded as a qualitative random variable Y. We refer to [5] for a review of automatic methods for assigning causes of death Y from verbal autopsy data X. In particular, classification methods based on Bayes' rule have been proposed, see [6] for instance.
Here, we focus on data measured on the deceased persons during the period from 1985 to 2010 in the three IRD (Research Institutr for Development) sites (Niakhar, Bandafassi and Mlomp) in Senegal. The dataset includes $n = 2.500$ individuals (deceased persons) distributed in $K = 18$ classes (causes of death) and characterized by $p = 100$ variables (symptoms).

4.2 Comparison with levels of interaction

We note that the classification rates associated with level of interaction $r = 3$ are higher than in the interaction $r = 1$ and $r = 2$. For $\gamma = 0.5$, the classification rate is invariant on the order of interaction. This is explained

by the fact that for $\gamma = 0.5$, the proposed kernel does not take into account the interactions and is calculated only based on the main variables. The highest classification rate is obtained for $\gamma = 0.67$ with a level of interaction equal to $r = 3$. The table 1 summarizes the classification rate depending on the level of interaction and the value of γ.

interactions	**r= 1**		**r = 2**		**r = 3**	
γ	**CCR (learning set)**	**CCR (test set)**	**CCR (learning set)**	**CCR (test set)**	**CCR (learning set)**	**CCR (test set)**
0.5	76.21	67.44	76.21	67.44	76.21	67.44
0.6	83.50	74.33	85.77	76.48	86.59	77.07
0.67	84.20	74.92	86.50	76.95	86.93	**77.19**
0.7	84.53	**75.25**	86.63	**77.00**	86.93	77.14
0.8	84.32	74.95	84.94	75.76	85.10	75.57
0.9	83.15	73.72	83.97	74.50	83.01	73.21
1	71.36	61.52	75.09	64.91	74.72	64.59

Tableau 1: Summary of correct classification rate for $\gamma \in [0.5, 1]$

5 Conclusion

This work was motivated by the consideration of the hierarchical aspect of the questions in the interview with the physician. We proposed a kernel that takes account a tree structure of the levels of response to the questions during the interview and the interactions of symptoms. This kernel implemented in the method ppgda presents consistent classification performance. A good diagnosis is obtained often accurate by the presence or absence of symptoms but particularly their interaction.
Our main contribution is the proposal of a kernel simultaneously taking into account the hierarchical appearance and interaction variables. The proposed kernel has good classification performance on a complex set of diagnostic data (high number predictors and classes).
An adaptation of this structured kernel on the graphical data might be useful on many issues.
This work could be extended to the classification of mixed quantitative and binary data by specifying the interactions of the variables.

References

[1] C. Bouveyron, M. Fauvel, and S. Girard., "Kernel discriminant analysis and clustering with parsimonious Gaussian process models.," *Statistics and Computing*, pp. 1–20, 2014.

[2] M. Dundart and D. Landgrebe, "Toward an optimal supervised classifier for the analysis of hyperspectral data," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 42, no. 1, pp. 271–277, 2004.

[3] R. Cattell, "The scree test for the number of factors," *Multivariate Behavioral Research*, vol. 1, no. 2, pp. 245–276, 1966.

[4] S. Sylla, S. Girard, A. Diongue, A. Diallo, and C. Sokhna, "A classification method for binary predictors combining similarity measures and mixture models," *Dependence Modeling*, vol. 3, pp. 1090–1096, 2015.

[5] B. Reeves and M. Quigley, "A review of data-derived methods for assigning causes of death from verbal autopsy data.," *International Journal of Epidemiology*, vol. 26, no. 5, pp. 1080–1089, 1997.

[6] P. Byass, D. Huong, and H. V. Minh, "A probabilistic approach to interpreting verbal autopsies: methodology and preliminary validation in vietnam," *Scandinavian Journal of Public Health*, vol. 31, no. 62 suppl, pp. 32–37, 2003.