Approche connexionniste pour l'extraction de profils cas-témoins du cancer du Nasopharynx à partir des données issues d'une étude épidémiologique

Khalid Benabdeslem*, Mustapha Lebbah**, Alexandre Aussem* et Marilys Corbex***

*PRISMa, Université Lyon1,
8, Avenue Niels Bohr, 69622 Villeurbanne Cedex, France
{kbenabde, aaussem}@bat710.univ-lyon1.fr
** Université Paris 13, UFR de Santé,
Médecine et Biologie Humaine (SMBH) - Léonard de Vinci- LIM&BIO
74, rue Marcel Cachin 93017 Bobigny Cedex France
lebbah@limbio-paris13.org,
***CIRC, Unité d'épidimiologie génétique
150, cours Albert Thomas, 69280 Lyon Cedex 08, france
corbex@iarc.fr.

Résumé. Dans cet article, nous présentons un système de découverte de connaissances à partir de données issues d'une étude épidémiologique cas-témoins du cancer du Nasopharynx (NPC). Ces données étant obtenues par une collecte de questionnaires, elles ont d'une part, la particularité d'être qualitatives et, d'autre part, de présenter des valeurs manquantes. Prenant en compte ces deux dernières contraintes, le système que nous proposons suit une démarche d'exploration de données qui consiste à (1) définir une procédure de codage des données qualitatives en présence de valeurs manquantes ; (2) étudier les propriétés de l'algorithme des cartes auto-organisatrices de Kohonen et son adaptation à ce type de données dans un cadre de découverte et de visualisation de groupes homogènes des cas cancer / non-cancer ; (3) post-traiter le resultat de cet algorithme par une classification automatique pour optimiser le nombre de groupes ainsi trouvés, et (4) donner une interprétation sémantique des profils extraits de chaque groupe. L'objectif général de cette étude est d'éclater le profil statistique global de la population étudiée en un ensemble de profils types (cancer ou non-cancer) et d'extraire pour chaque profil l'ensemble de variables explicatives du NPC à partir d'une cartographie bidimensionnelle.

1 Introduction

L'algorithme des cartes auto-organisatrices de Kohonen, Kohonen (1994) représente un véritable outil de visualisation des données multidimensionnelles. Il permet de convertir des relations statistiques complexes et non-linéaires entre les données de grande dimension en une simple relation géométrique sur une topologie réduite. Cet algorithme permet de compresser

l'information tout en préservant les relations topologiques et métriques les plus importantes à partir de l'espace des données primaires. De plus, l'algorithme de Kohonen définit un niveau d'abstraction par la possibilité d'interprétation qui devient plus facile avec une carte bi-dimensionnelle, à la fois simple et significative comparée à l'espace initial des données. Bien que l'algorithme de Kohonen décrit une méthode connexionniste qui appartient à la famille des algorithmes neuronaux, il peut être formulé par une méthode de classification statistique type : nuées dynamiques. Ce formalisme transforme le problème d'auto-organisation en un problème d'optimisation. Dans ce contexte, nous décrivons une variante de classification neuronale non-supervisée proposée par Lebbah et al. (2000). Cette variante, appelée carte topologique binaires, est dédiée aux données qualitatives et consiste en la recherche d'une classification automatique d'un nuage de points $App = \{(z_i, p_i), i = 1..N\}$ où l'individu $z_i = (z1_i, z2_i, ..., z_i^d)$ muni de la pondération p_i appartient à l'ensemble des données binaires $\beta^d = \{0, 1\}^d$. Cette variante s'inspire de la version nuées dynamiques de Kohonen, Anouar (1998), mais utilise un critère spécifique pour déterminer l'ordre topologique. Dans cet article, nous étudions le comportement des cartes topologiques binaires face à des données qualitatives présentant des valeurs manquantes. Ces données sont issues d'une étude épidémiologique cas-témoins du cancer du nasopharynx (NPC). Cette base est constituée d'une population divisée, de manière équitable, en deux cas : cancer et non-cancer. D'une part, notre étude consiste à extraire des profils de gens atteints du NPC et des gens qui ne le sont pas et d'autre part, elle consiste à déterminer pour chaque profil extrait, l'ensemble des variables explicatives de la population qui lui est associée. L'objectif de cette étude vise donc à détecter dans le profil statistique général du NPC, des profils "types" sous formes de groupes de population homogènes. Chaque groupe, étant un représentant d'un cas particulier de la population globale, est muni d'un ensemble de caractéristiques résultants de la classification non-supervisée faite par les cartes topologiques binaires et optimisée par une classification statistique automatique.

2 Données qualitatives et codage

Il existe de nombreuse variables, dites discrètes, ne pouvant prendre par nature qu'un nombre restreint de valeurs Marchetti (1989). Citons par exemple les variables associées aux caractéristiques physiques tel que la taille (grande, moyenne, petite) ou encore à la situation familiale (célibataire, veuf, divorcé, marié). Les variables ainsi définies sont appelées variables qualitatives. Elle se répartissent en deux groupes : les variables qualitatives ordinales et les variables qualitatives nominales. Si l'on utilise un codage adapté, les données qualitatives deviendront des données binaires. Les codages utilisés le plus souvent sont : (a)*Le codage binaire additif :* Ce codage permet essentiellement de rester cohérent avec la notion d'ordre entre les modalités d'une variable.(b) *Le codage disjonctif complet :* Ce tableau résulte de la transformation, par le codage disjonctif complet, d'un tableau de variables qualitatives nominales encore appelé questionnaire multiple. Une seule modalité est choisie pour chaque variable, TAB.1.

Que les données initiales soient dans l'espace des données avec modalités ou après transformation dans l'espace des données binaires, nous aboutissons à des caractéristiques identiques Leich et al. (1998). L'espace des données binaires peut être muni de la distance euclidienne, il est souvent beaucoup plus intéressant de le munir de distances adaptées permettant de mieux traduire ses particularités. Dans ce papier nous utilisons la distance de Hamming appelée $\mathcal{H}$. Le calcul de cette distance entre deux individus z_1 et z_1 se fait alors à partir de la table de

contingence établie à partir des vecteurs binaires qui leurs sont associés, TAB.2. La distance de Hamming binaires entre z_1 et z_2 est le nombre de composantes différentes entre ces deux points.$\mathcal{H}(z_1, z_2) = b + c = |z_1 - z_2|$

Modalités	Codage additif	Codage disjonctif
1	1 0 0	1 0 0
2	1 1 0	0 1 0
3	1 1 1	0 0 1

TAB. 1 – *Codage des modalités.*

z_2/z_1	1	0
1	a	b
0	c	d

TAB. 2 – *Table de contingence.*

Le nuage de points App, peut maintenant être caractérisé par sa caractéristique de valeur centrale associée à la distance de Hamming, c'est le centre médian. Le nuage étant inclus dans l'espace β^p, il admet pour centre médian un point de ce même espace. Par définition le centre médian du nuage App est tout point $\omega^j = (\omega^1, \omega^2, ..., \omega^d)$ de β^d minimisant l'inertie du nuage définit par la distance de Hamming $\sum_{i=1}^{I} p_i \mathcal{H}(z_i, \omega)$ ce qui signifie que, pour tout j, ω^j minimise :$\sum_{i=1}^{I} p_i |z_i^j - \omega^j|$

Les données étant binaires, ω^j est la médiane binaire de l'ensemble des valeurs prises par la variable j sur l'ensemble des individus. La médiane est la valeur 1 ou 0 correspondant à la plus grande sommation des pondérations de la valeur 1 et 0. Dans le cas particulier où z_i sont munis d'une même pondération $(p_i = 1, \forall i)$ la règle fournit une médiane ayant une interprétation particulièrement simple, ω^j est alors la valeur 0 ou 1 la plus souvent choisie par les individus sur la variable j.

Valeurs manquantes

Le problème des valeurs manquantes est un véritable problème de recherche. Ceci étant, il existe un certain nombre de façons pour détourner ce problème, par exemple en les remplaçant par la médiane ou par l'apprentissage d'un prédicteur automatique,...etc. Cependant, dans notre cas, les variables sont qualitatives. Nous proposons, pour une raison de simplicité, de définir une modalité supplémentaire pour les valeurs manquantes. Cette pseudo-solution, ne pose aucun problème dans le cas d'un codage disjonctif. Cependant, le problème se pose pour le codage additif où l'ordre entre les modalités est important. Nous proposons donc de définir la modalité des valeurs manquantes de telle façon à ce qu'elle soit la plus proche de la médiane entre toutes les autres valeurs de la variable.

3 La carte topologique binaire

Nous rappelons ici comment l'utilisation de la médiane peut permettre de définir un modèle de carte auto-organisatrice adapté aux données binaires. Comme pour le modèle classique des cartes topologiques, nous utilisons un réseau de neurones avec une couche d'entrée pour les entrées et une carte possédant un ordre topologique de k cellules. La prise en compte dans la carte C de la notion de proximité impose de définir une relation de voisinage topologique. Les neurones sont répartis aux noeuds d'un maillage. Comme dans le cas de l'algorithme de Kohonen nous définissons la topologie de la carte à l'aide d'un graphe non orienté et la distance $\delta(c, r)$ entre deux cellules c et r étant la longueur du chemin le plus court qui sépare la cellule c et r. Afin de modéliser la notion d'influence d'un neurone r sur un neurone c, qui dépend de leur proximité, on utilise une fonction à la fonction noyau $\mathcal{K}$ ($\mathcal{K} \geq 0$ et $\lim_{|x|\to\infty} \mathcal{K}(x) = 0$).

L'influence mutuelle entre deux cellules c et r est définie par la fonction $\mathcal{K}(\delta(c, r))$. A chaque cellule c de la grille est associé un vecteur de poids binaire w_c de dimension d. L'ensemble des poids associés constitue l'ensemble des référents noté $\mathcal{W}$. L'auto-organisation de la carte va maintenant se faire à l'aide du formalisme des nuées dynamiques et donc par l'intermédiaire de la minimisation d'une fonction de coût.

Pour utiliser l'algorithme des nuées dynamiques, Diday et C.Simon. (1976), nous avons utilisé la fonction de coût $\mathcal{E}(\phi, \mathcal{W})$ déjà définit dans Lebbah et al. (2000), adaptées aux traitements des données binaires . Donc la fonction de coût à minimiser est alors :

$$\mathcal{E}(\phi, \mathcal{W}) = \sum_{z_i \in App} \sum_{r \in C} \mathcal{K}(\delta(\phi(z_i), r))\mathcal{H}(z_i, w_r) \tag{1}$$

Où ϕ affecte chaque observation z à une cellule unique de la carte C.
La minimisation de la fonction de coût est réalisée à l'aide d'une procédure itérative en deux phases :

1. **Phase d'affectation** : mise à jour de la fonction d'affectation ϕ associée à l'ensemble $\mathcal{W}$ fixé. On affecte chaque observation $\mathbf{z}$ au référent défini à partir de l'expression suivante :

$$\forall \mathbf{z},\ \phi(z) = \arg\min_{c}(\mathcal{H}(z, w_c)) \tag{2}$$

2. **Phase d'optimisation** : La fonction d'affectation étant fixée à sa valeur courante, choisir le système de référents qui minimise la fonction $\mathcal{E}(\phi, \mathcal{W})$ dans l'espace β^m. ce point n'est autre que le centre médian de $\mathcal{A}pp$ lorsque chaque observation $\mathbf{z}_i$ est pondérée par $p_c = K^T(\delta(\phi(\mathbf{z}_i), c))$. Chaque composante $w_c = (w_c^1, ..., w_c^k, ..., w_c^d)$ est calculée comme suit :

$$w_c^k = \begin{cases} 0 & \text{si } \left[\sum_{z_i \in App} \mathcal{K}(\delta(c, \phi(z_i)))(1 - z_i^k)\right] \geq \\ & \left[\sum_{z_i \in App} \mathcal{K}(\delta(c, \phi(z_i)))z_i^k\right] \\ 1 & \text{sinon} \end{cases},$$

La minimisation de $\mathcal{E}(\phi, \mathcal{W})$ s'effectue par itérations successives jusqu'à stabilisation des deux phases.

Dans la pratique nous avons utilisé une fonction noyaux $\mathcal{K}^T(\delta(c,r)) = \exp(\frac{-0.5\delta(c,r)}{T})$ en faisant varier le paramètre T entre deux valeurs T_{max} et T_{min}. On obtient alors pour décrire la carte un ensemble de référents binaires $\mathcal{W}$. Ces référents sont du même genre que les données initiales : Le décodage (additif ou exclusif) de différents vecteurs permet l'interprétation symbolique des référents trouvés.

4 Résultats

Nous appliquons la méthode aux données d'une étude épidémiologique cas-témoins du cancer du nasopharynx (NPC). Pour clarifier le rôle de l'environnement dans l'étiologie du NPC, le CIRC a mené en 2004 une étude cas-témoins multicentrique dans la région endémique du Maghreb. Le NPC présente une incidence très variable selon les régions du monde. C'est un cancer relativement rare sauf en Chine, en Asie du sud-est et au Maghreb, où les taux d'incidence sont élevés. Dans ces régions, le NPC est un problème majeur de santé publique. Les études ont suggéré l'existence d'un grand nombre de facteurs de risques environnementaux incluant habitudes alimentaires et environnement domestique et professionnel. Afin de focaliser la présente analyse sur la forme adulte du NPC seuls les individus âgés de plus de 35 ans recrutés dans lŠ étude ont été sélectionnés, soit un total de 986 individus, dont 499 sont atteint par le cancer (les cas), et 487 ne le sont pas (les témoins). Chaque individu est décrit par 61 caractères.

Dans les 61 caractères, il y a la variable à expliquer 1-" NPC " (c'est la variable du cancer) ; les autres sont les variables explicatives. Lexique : 2-âge, 3-sexe, 4-niveau dŠinstruction, 5-catégorie professionnelle, 6-habitat dans lŠenfance, 7-habitat a lŠage adulte, 8-parents consanguins, 9-fréquentes otites, 10- fréquentes angines, 11- fréquentes rhume, 12-asthme, 13-eczéma, 14-allergie, 15-exposition aux engrais chimiques et pesticides, 16- exposition aux produits chimiques, 17- exposition aux fumées, 18- exposition aux poussières, 19- exposition aux formaldéhyde, 20-consommation dŠalcool, 21- consommation de tabac, 22- consommation de neffa, 23- consommation de cannabis, 24-25-type de logement enfant. et adulte., 26-27-lits séparés enfant. et adulte., 28-29-animaux dans la maison enfant. et adulte., 30-31-aération cuisine enfant. et adulte., 32-33-aération maison enfant. et adulte., 34-35- exposition aux fumées d'encens enfant. et adulte., 36-37- exposition aux fumées de kanoun et tabouna enfant. et adulte., 38-39- exposition aux fumées de feu de bois enfant. et adulte., 40-41-42-allaité et age au sevrage et modalité de sevrage, 43-contact avec la salive adulte par le sol ou les aliments, 44-traitements traditionnels dans lŠenfance, 45- consommation de piment, 46-47- consommation de smen et graisse enfant. et adulte., 48-49-légumes fruits agrumes enfant. et adulte., 50-51-harrissa maison enfant. adulte., 52-53-harrissa industrielle enfant. adulte., 54-55-protéines maison enfant. adulte., 56-57-protéines industrielles enfant. adulte., 58-59-conserves légumes industrielles enfant. adulte., 60-61-conserves légumes maison enfant. adulte.

Nous appliquons la variante des cartes topologiques binaires sur les données décrites ci-dessus avec une architecture de (10×10) cellules. l'apprentissage de cette carte fournit pour chaque cellule un référent w_c prenant en compte les deux codages : disjonctif et additif. Fig.1 représente la distribution de la population sur la carte. Fig.2 représente la répartition en distinguant les cas cancer et cas non-cancer. Grâce à la cartographie obtenue, nous pouvons déjà effectuer quelques analyses sur la répartition des individus. En effet, il existe 5 neurones vides, i.e. des neurones qui n'incluent aucun individu. Ces neurones représentent la propriété de lis-

44	31	21	25	21	11	14	15	17	15
18	11	5	0	14	13	3	0	2	8
33	5	5	11	9	1	6	12	5	16
37	14	13	4	4	6	7	4	6	0
11	16	5	4	3	12	8	0	10	11
14	8	3	13	8	15	4	6	4	0
22	5	5	7	5	9	5	4	6	5
13	14	6	7	2	10	6	7	4	8
8	10	5	8	3	13	4	8	1	16
37	6	12	7	6	21	9	7	9	5

FIG. 1 – *Carte topologique 10 × 10 avec les cardinalites des neurones*

sage, une des propriétés fortes qui caractérisent les cartes auto-organisatrices. Les neurones sont mélangés, ce qui veut dire que quelques individus n'ayant pas le cancer, même s'ils sont minoritaires dans un neurone, peuvent avoir les mêmes caractéristiques que les individus ayant le cancer !

Pour optimiser le nombre de neurones obtenu dans la carte, nous avons appliqué la méthode des K-means sur les référents avec différentes partitions et nous avons choisi celle qui minimise le mieux l'indice de boulding. Le nombre de classes optimal trouvé étant égal à 6 (Fig.3). A partir de l'indice optimal indiqué dans cette dernière figure, nous avons partitionné la carte en 6 grandes classes (Fig.4). Dans cette figure, nous avons constaté les statistiques suivantes : Classe 1 : la zone bleue qui regroupe 13% de la population dont 67% ayant le cancer. Classe 2 : La zone mauve, regroupant 13% de la population dont 53% ayant la cancer. Classe 3 : la zone verte représentant 25% de la population dont 65% de Non-cancer. Classe 4 : zone rouge, représentant 18% dont 60% ayant le cancer. Classe 5 et 6 (31% de la population) : zones jaune et orange, représentent des zones de conflit, que nous ne pouvons interpréter.

Au vu de ces résultats, deux classes attirent notre attention : la classe 1 et la classe 3 en raison de : 1) la proportion anormale d'individus atteints du cancer au regard des 50% dans la population d'origine, et 2) le nombre significatif d'individus dans ces classes. Pour caractériser les profils des individus dans ces classes, nous avons étudié la distribution des variables explicatives. La divergence de Kullback-Leibler Bishop (2006) notée $KL(p||q)$ (i.e., $-\sum_x p(x)\ln(q(x)/p(x))$ dans le cas discret) est classiquement utilisée pour mesurer la dissimilarité entre la distribution d'origine $p(x)$ d'une variable et celle observée dans la classe considérée, $q(x)$. Aussi, nous avons classé les 60 variables explicatives dans chaque classe dans l'ordre décroissant de la divergence KL. Le principe sous-jacent est de dire qu'une variable est d'autant plus discriminante que sa distribution dans la classe considérée est significativement modifiée par rapport à la distribution d'origine.

Dans la classe 1 pour laquelle il y a une sur-représentation des individus atteints du NPC,

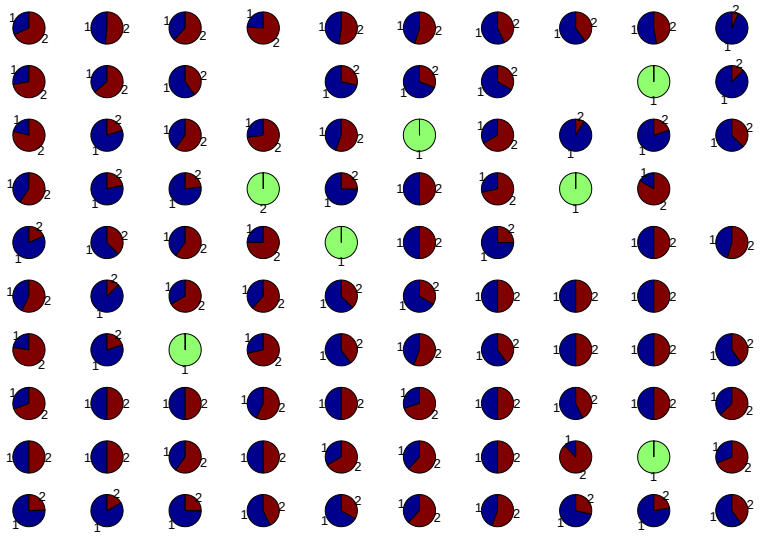

FIG. 2 – *Distribution des cas cancer :1 et non-cancer :2.*

on observe que 3 variables (50, 45 et 51) ont une valeur de KL élévée (0.75, 0.59 et 0.52 respectivement) au regard des autres (<0.25). Elles sont associées à la consommation de nourriture pimentée et d'harrissa à l'enfance et à l'âge adulte. Plus précisément, on observe dans cette classe que 60% des individus ont consommé de l'harrissa à l'enfance alors que ce taux n'est que de 20% dans la population étudiée.

Dans la classe 3 caractérisé par un sous-représentation des individus atteints du NPC, on observe que la variable 26 a une valeur de KL élévée (0.8) suivie de 32, 30 et 24 (0.38, 0.32, 0.31) au regard des autres (<0.25). La variable 26 est associée au lits séparés à l'enfance, les autres portent sur l'aération de la maison, de la cuisine et de la catégorie du logement à l'enfance. On observe dans cette classe que 90% des individus avaient des lits séparés à l'enfance alors que ce taux n'est que de 40% dans la population étudiée. De même, 22% ont vécu dans une maison bien aérée durant l'enfance alors que ce taux n'est que de 2% dans la population étudiée.

A titre d'exemple, nous illustrons pour la classe 1, la distribution de la variable 50 (Harissa maison enfant) sur la carte topologique obtenue (Fig.5). Cette figure représente l'impact des modalités de la variable 50 sur les individus atteins du NPC. La première modalité est distribuée sur tous les neurones de la carte, ce qui est normal à cause du codage additif de la variable. La quatrième modalité n'est représentée sur aucun neurone. Il s'agit d'une modalité additionnelle pour représenter les valeurs manquantes de la variable. La deuxième et la troisième modalités sont fortement présentes dans la classe étudiée (zone bleue).

Au final, ce type d'analyse nous renseigne sur l'existence de profils statistiques distincts

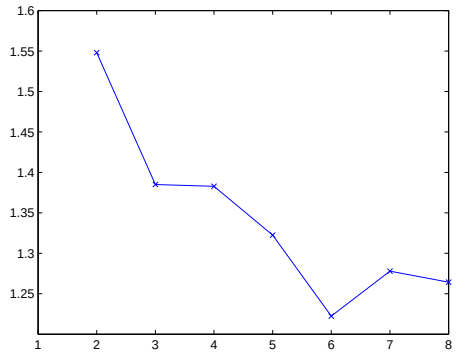

FIG. 3 – *Indice de boulding pour le choix optimal du nombre de classes.*

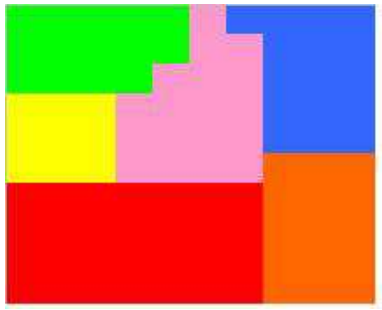

FIG. 4 – *Carte topologique partitionnée en 6 classes.*

dans la population étudiée. L'analyse des profils dans chacune des classe nous a permis d'identifier des facteurs corrélés avec le cancer (ou l'absence du cancer). Par ce type d'analyse nous avons pu dresser des profils statistiques sémantiques des différentes catégories d'individus atteints ou non du cancer et ainsi extraire l'ensemble de variables explicatives de chaque profil.

5 Conclusion

Dans ce papier, nous avons présenté un système connexionniste pour l'extraction de profils cas témoins du cancer du nasopharynx (NPC) à partir des données issues d'une étude épidémiologique. Ce système utilise un codage spécifique aux données qualitatives représentant des valeurs manquantes. Basé sur une carte topologique binaire, le système ainsi développé a permis d'une part, de trouver des groupes homogènes à partir d'une population globale regroupant des cas cancer et non-cancer et d'autre part, d'extraire les variables explicatives de chaque profil extrait. Nous avons pu grâce à ce système fournir aux épidémiologistes un outil

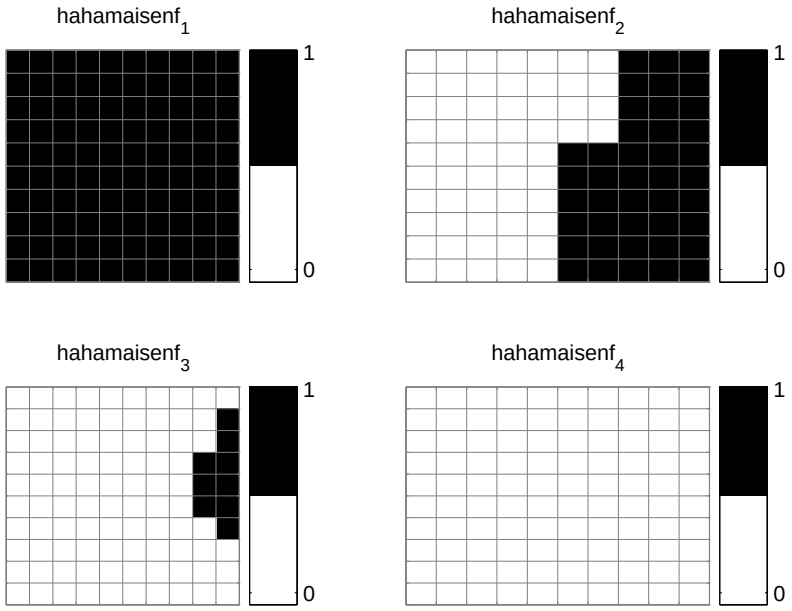

FIG. 5 – *Carte topologique décrivant l'importance des modalités de la variable Harissa maison enfant.*

d'aide à la décision qui leurs fournit un véritable outil de visualisation à partir des données multidimensionnelles. Cet outil permet également, d'éclater le profil général de la population des gens atteints du NPC en un ensemble de profils "type". Chaque profil étant caractérisé par un ensemble de variables explicatives des cas cancer ou non-cancer. Le système développé sera prochainement comparé avec d'autres méthodes d'extraction de connaissances.

Références

Anouar, F. Badran, F. S. (1998). Probabilistic self-organizing map and radial basis function. *Neurocomputing 20*, 83–96.

Bishop, C. (2006). *Pattern recognition and machine learning*. Berlin: Springer.

Diday, E. et C.Simon. (1976). *An Introduction to Symbolic Data*. New York: Springer.

Kohonen, T. (1994). *Self-Organizing Map*. Berlin: Springer.

Lebbah, M., S. Thiria, et F. Badran (2000). Topological map for binary data. In *European Symposium on Artificial Neural Networks (ESANN'00)*, pp. 267–272.

Leich, F., F. Weingessel, et A. Dimitriadou (1998). Competitive learning for binary data. In *International Conference on Articial Neural Networks mining (ICANN'98)*, pp. 779–784.

Approche connexionniste pour l'extraction de profils cas-témoins du cancer

Marchetti, F. (1989). *Contribution a la classification de données binaires et qualitatives*. Thèse de doctorat, Université de Metz.

Summary

In this paper, we present a knowledge discovery system from nasopharyngeal (NPC) data. These data have two characteristics: they are qualitative, and present missing values in the database. Considering these two constraints, the system that we propose follows a data mining methodology. It consists in: (1) determining a coding procedure for qualitative data with missing values; (2) Studying the proprieties of topological map of Kohonen and its adaptation for this kind of data; (3) optimizing the number of classes obtained by this algorithm using an automatic clustering method and (4) giving a semantic interpretation of the extracted profiles. Our aim is to find different profiles from the general one. For each profile, we give the group of explicative variables of NPC from a bi-dimensional map.