E.M.C.A , (Environnement pour les Méthodes de Classification Automatique de données multidimensionnelles) basé sur une modélisation Objet

Abdessamed Réda GHOMARI
Institut National de formation en Informatique (I.N.I)
B.P 68M, Oued Smar 16270 Alger -ALGERIE-
Tel: (213) (2) 51.60.77 - 51.63.91 Fax: (213) (2) 51.61.56
Email: sisia@ist.cerist.dz

RESUME : Ce papier a pour objet de présenter un environnement pour les méthodes de classification automatique (E.M.C.A) destiné à fournir à des utilisateurs statisticiens et gestionnaires novices à tous les niveaux de l'entreprise, une boîte à outils pour la typologie de données et l'interprétation de résultats.
E.M.C.A s'appuie sur une démarche de conception orientée objet. L'intérêt se trouve illustré par les apports en matière de modularité et de continuité temporelle (greffe facile de nouveaux programmes de classification). L'implémentation de la première version du logiciel en Borland C++ sous Windows garanti la simplicité, la banalité d'emploi, l'assistance et la flexibilité recherchés par la majorité de concepteurs de produits statistiques.

MOTS CLES: Analyse de Donnés, Classification, groupe, partition, hiérarchie, objet, Base de données, qualité logiciel.

ABSTRACT: The object of this paper is to present E.M.C.A. This environment provide statisticians and inexperimented managers in all firm levels, with data typology tool box.
E.M.C.A lean on object-oriented conception step. Its interest to provide modularity and temporal continuity (easy transplant of new clustering programs). The first package implementation version with Borland C++ under Windows garantees simplicity, banality use, help and flexibility aimes at by all statistic product developers.

KEYWORDS: Data analysis, classification, group, partition, hierarchy, object, data basis, package quality,

INTRODUCTION : Nous constatons aujourd'hui, que la compréhension à travers les techniques d'apprentissage numérique, notamment par l'analyse de données multidimensionnelles ne cesse de s'amplifier. En effet les logiciels sont devenus une nécessité inéluctable, offrant un atout majeur pour les analystes et statisticiens grâce notamment à la rapidité des traitements et la capacité de mémorisation des données..
L'analyse des données cherche à extraire d'une grande masse de données multidimensionnelles les "informations utiles"[MAR91]. Cette synthèse peut être effectuée à l'aide de méthodes de visualisation (Analyse factorielle) , de méthodes de structuration (Classification) ou de méthodes d'explication (Discrimination,...) [AUR90.4]. C'est au deuxième groupe de méthodes que nous nous intéressons.
En plus, il convient de noter que malgré la pluralité des produits offerts sur le marché, ces derniers restent en dessous des attentes des utilisateurs éprouvant des difficultés dans la gestion

du potentiel informationnel en amont et surtout en aval du processus de classification (respectivement les tableaux de données et les résultats à interpréter) [JAM89].
Dans l'optique d'une meilleure gestion de cette complexité, EMCA, projet en cours à l'INI, est le fruit d'un travail de conceptualisation de l'activité classificatoire grâce à une analyse orientée objet [GHO94]. Il garantit une facilité d'emploi et des possibilités d'exploitation considérables.
Le présent papier est organisé comme suit: une brève présentation de synthèse du domaine est l'objet de la section I. Une approche de modélisation de l'activité de classification à travers le formalisme objet [COA92] est présentée à la section II.
La section III se focalise elle sur la présentation de la première version de l'environnement pour les méthodes de classification automatique (EMCA) à travers son architecture détaillée suivi d' explications des fonctions assurés par le logiciel. Le produit a été développé en Borland C++ sous MS- Windows.

I LA CLASSIFICATION : PRESENTATION DE SYNTHESE

La nature offre un grand nombre de populations qu'il est souhaitable de répartir en catégories. Chaque discipline scientifique sollicite des classifications. Le terme le plus couramment utilisé, est la *classification* ou *classification automatique*. On trouve aussi *taxinomie* (ou taxonomie) en biologie et zoologie et *nosologie* en médecine.[MAR91]
Le terme classification recouvre plusieurs significations (trois processus et un résultat) selon le contexte dans lequel il est utilisé. Le sens qui lui est donné en analyse de données est celui de la distribution en classes ou catégories (décision par la distance). Le sens qui lui est communément donné en intelligence artificielle est de celui de classement (procédé d'apprentissage). Etant donné un objet, trouver sa classe d'appartenance.[HAT91]. La classification est aussi vue comme processus de discrimination fonctionnelle.[BEL92]
Enfin, comme résultat: Faire émerger, d'un ensemble de données, une structure particulière qui restitue l'essentiel de l'information tout en réduisant la masse de données [JAM89]. Les étapes par lesquelles passe l'utilisateur d'une méthode de classification sont récapitulés dans le schéma ci-dessous [CHA81].

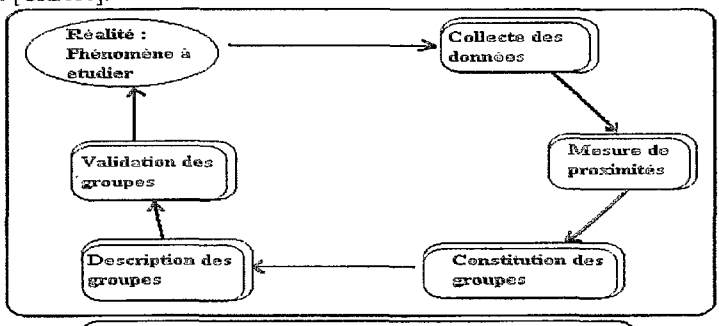

FIG 1: LES ETAPES DE LA CLASSIFICATION [CHA 81]

Typologie des méthodes de Classification: La combinaison des critères et des hypothèses simplificatrices permet de classer les algorithmes en grandes familles comme le font Joyce et Channon (1966), Benzecri (1973), Andberg (1973), Everitt (1974) et Berti-Bourouche (1975). [CHA81]

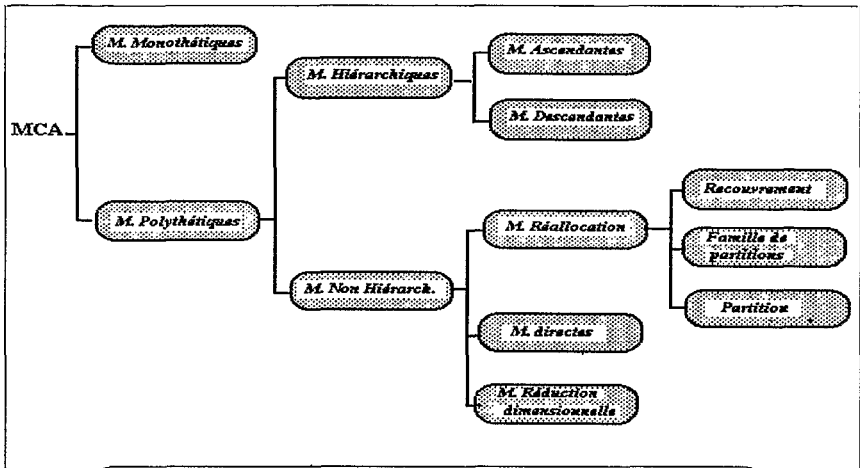

FIG 2 : TYPOLOGIE DES METHODES DE CLASSIFICATION

II. MODELISATION DE L'ACTIVITE CLASSIFICATION DES DONNEES

II.1 Justifications et opportunités:

Mettre l'accent sur la conceptualisation de l'activité de classification de grands ensembles de données par des formalismes est d'un intérêt majeur. D'abord la préoccupation de classification est commune à plusieurs disciplines, ensuite le domaine est riche par: [CEL89]

① le nombre de problèmes soulevés ② le choix des objets et des variables (caractères)
③ le codage des données ④ la variété de mesures de ressemblances ou dissemblances
⑤ la structure des classes ⑥ les critères à optimiser ⑦ le choix de l'algorithme.

En plus pour un maximum de flexibilité des outils à offrir à l'utilisateur, une réflexion primaire indépendante des particularités des programmes de classification automatique existants sur le marché est plus que nécessaire. Elle aura comme avantages de:

❑ Garantir un pouvoir de synthèse et de communication (à travers les modélisation des besoins et des points communs des statisticiens).

❑ Faire bénéficier le domaine d'analyse de données multidimensionnelles de formalismes comme support efficace dans l'étude statistico-informatique du problème.

❑ Créer une base informationnelle (base de données) suffisamment riche et extensible pour l'interprétation et la gestion des données.

II.2 Formalisation par le modèle objet:

Le grand pouvoir d'expression d'un modèle objet repose avant tout sur cette possibilité d'une part de manipuler directement des ensembles d'objets et d'autre part, la relation généralisation /spécialisation qui structure l'ensemble des classes d'objets.

Le modèle objet s'accorde aussi avec la démarche progressive dans laquelle les objets sont introduits et enrichis au fur et à mesure des besoin.[AUB91]

La démarche de modélisation suivie est :[COA92]

- Identifier les classes et objets constituant l'essentiel du domaine classificatoire.
- Identifier des structures de généralisation-spécialisation et de composé-composant.

- Identifier des sujets destinés guider les lecteurs à l'intérieur de modèles complexes. Notre découpage du modèle classificatoire à été opéré en trois sujets: * Gestion des données * Constitution des groupes * Interprétation des résultats
- Définir les attributs comme information d'état des objets.
- Identifier des services modélisant les comportements spécifiques.

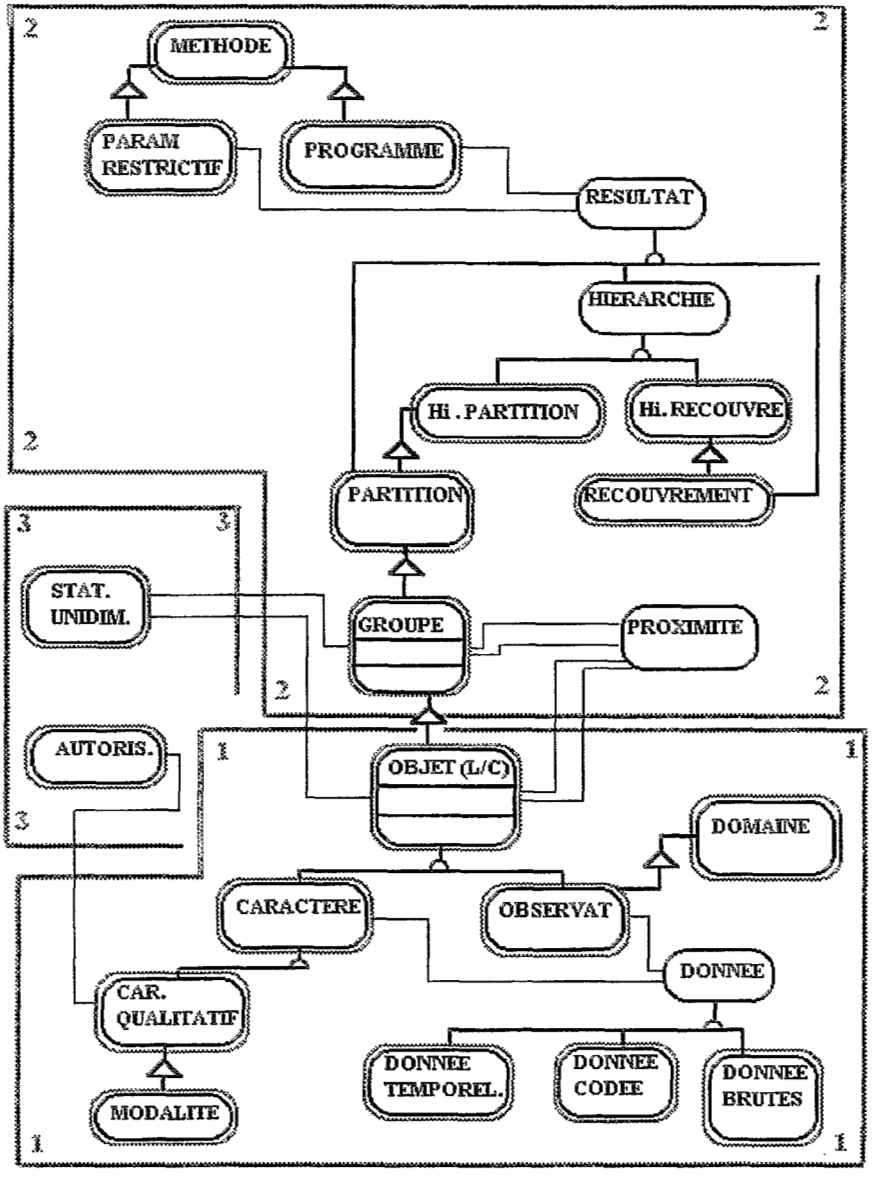

FIG 3: Modèle objet

III PRESENTATION D' EMCA

III.1 Les objectifs assignés : Cet environnement sera une réelle boîte à outils informatique pour les problèmes de typologie des données s'appuyant sur la méthodologie d'approche en analyse de données multidimensionnelles [AUR90.4] qui prend en charge la diversité des problèmes soulevés et des structures recherchées par les utilisateurs des logiciels de classification automatique.
L'implémentation objet permettra sans nul doute une greffe facile de nouveaux programmes de classification et une aisance dans la phase d'interprétation des résultats. Les résultats attendus peuvent se résumer en:
La mise en place de l'environnement informatique doit garantir des aptitudes en matière de qualité logiciel [THE88]. Il faudra assurer par ordre de priorité:

- *Facilité d'emploi:* Minimiser l'effort d'apprentissage du logiciel. L'utilisateur ne doit pouvoir s'intéresser qu'à la logique de son fonctionnement et être déchargé de tous les problèmes purement informatiques. (programmation en C++ sous MS-WINDOWS)
- *Flexibilité:* apte à supporter les évolutions sans remise en cause fondamentale de la structure existante (implémentation orientée objet).
- *Maniabilité:* minimiser l 'effort nécessaire pour l'apprentissage, mise en oeuvre des entrées et exploitation des sorties (mise en forme graphique, base de données comme réservoir structuré d'information et un rapport de cession pour une une interprétation aisée des résultats obtenus.
- *Efficacité :* La notion d'efficacité recouvre la minimisation des coûts d'exploitation, la rapidité d'exécution, la minimisation de la taille mémoire.

III.2 Architecture d'EMCA: [GHO94]

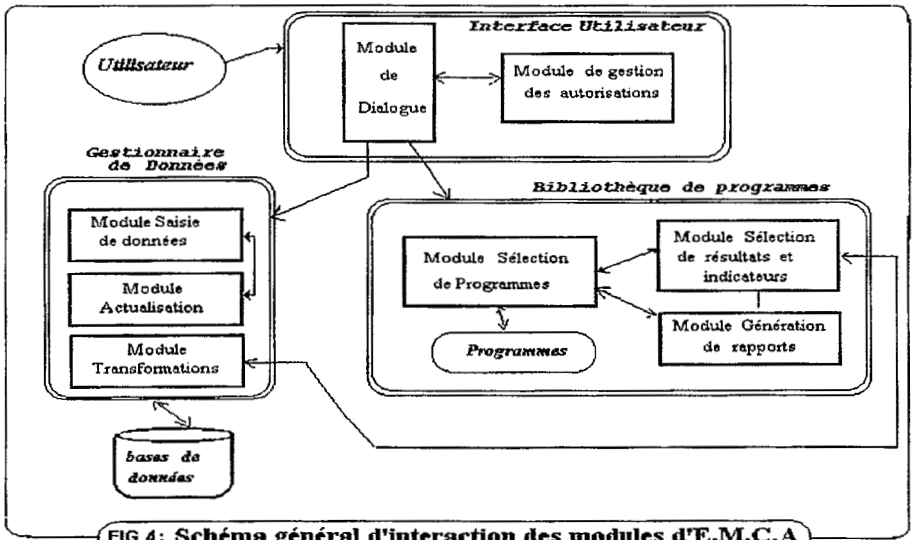

FIG 4: **Schéma général d'interaction des modules d'E.M.C.A**

III.3 Description détaillée d'EMCA:

❒ **L'interface utilisateur:** Constitue le point d'entrée du système et permet l'interaction et la coordination entre l'utilisateur et les différents modules composant le système.[BOU93].

a. Le module de dialogue: Le logiciel est organisé autour d'un écran principal et de deux autres menus déroulants regroupant toutes les opérations à déclencher par l'utilisateur.

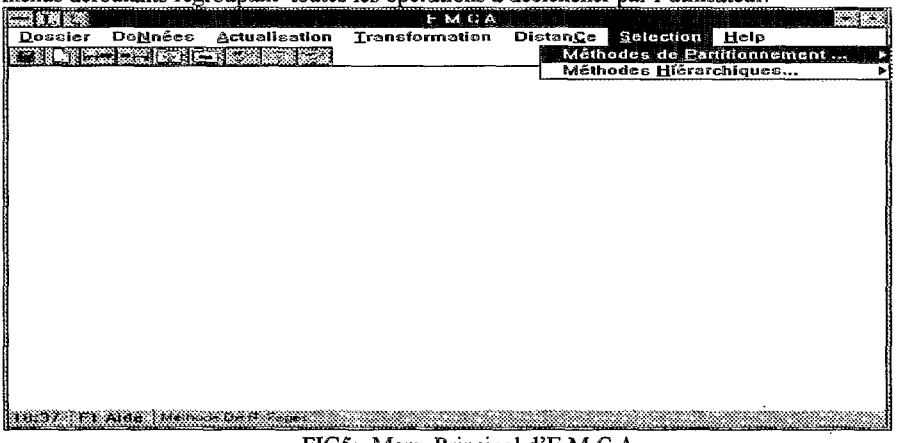

FIG5: Menu Principal d'E.M.C.A

Le menu correspondant au méthodes non hiérarchiques

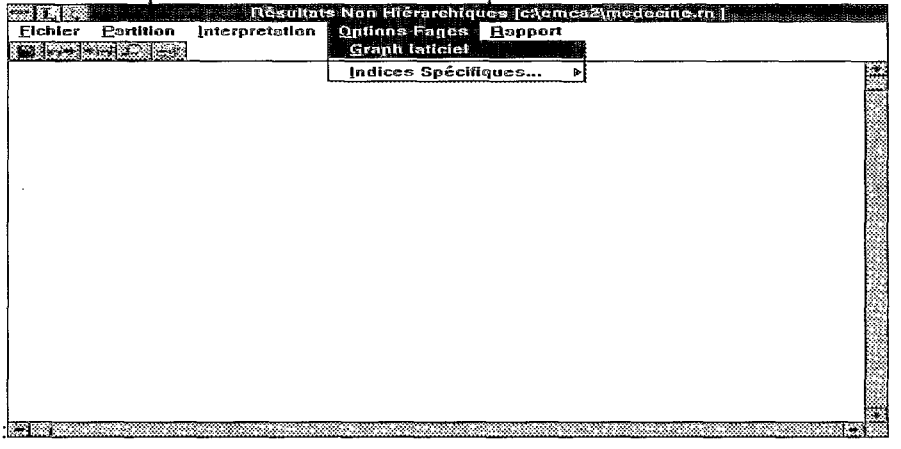

FIG6: Menu Secondaire d'E.M.C.A: Méthodes de partitionnement

Il est composé de :

Fichier : - permet l'ouverture, la sauvegarde du résultat (l'ensemble des partitions).
- l'impression de l'affichage de l'option en cours .

Partition : permet de sélectionner la partition à interpréter par l'introduction du numéro de la partition, le système affiche les éléments de chaque classe de la partition.

Interprétation : Cette option permet l'interprétation de la partition par un ensemble d'indices de description. L'interprétation des partitions repose sur la décomposition de l'inertie de la population en inertie interclasse et inertie intra-classe associées à la partition [AUR90.2].

Option-Fages : cette option permet l'interprétation à travers les indices spécifiques à la méthode de fages.

Le menu correspondant au méthodes hiérarchiques :

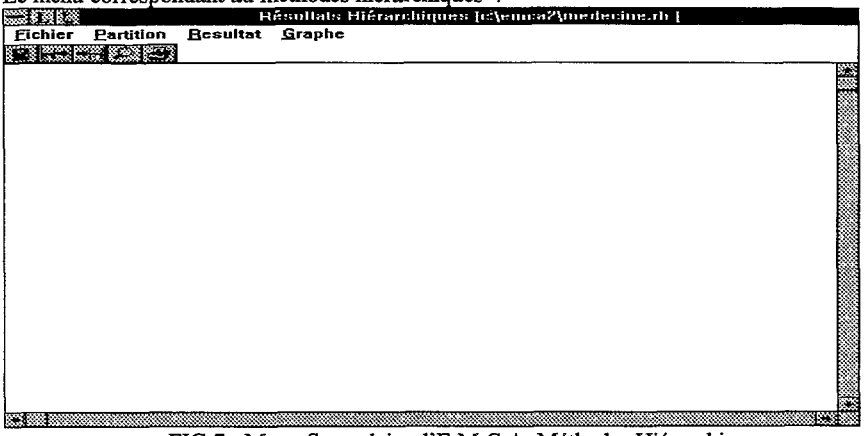

FIG 7: Menu Secondaire d'E.M.C.A: Méthodes Hiérarchiques

b.Le module « intelligent » de gestion des autorisations: aura comme rôle dans un premier instant de verrouiller les options non autorisées en cours de cession tenant compte de la méthodologie en analyse de données [AUR90.4] et des restrictions propres au domaine classificatoire. Son rôle est d'autant plus attrayant pour les utilisateurs novices.

❒ **Le gestionnaire de données**: Ce module permet la saisie des données brutes selon le type choisi : - Données individus-variables - Données de proximités.
Le gestionnaire de données est composé de :
a- Un module de saisie : permet la saisie de données à classifier (quantitatives, qualitatives, mixtes) à travers un mini-tableur intégré (grille de dimension modifiable).

- Création du dossier d'étude : Au début, Il s'agit de fournir les informations nécessaire sur les données de l'étude à travers la boite de dialogue ci-dessous.

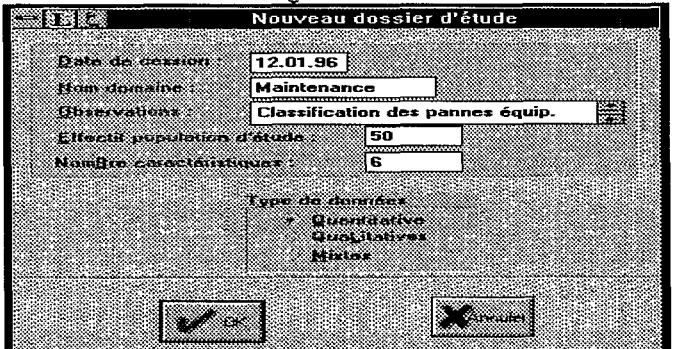

FIG 8: Boîte de dialogue: Description d'un dossier d'étude

- Saisie de l'échantillon: saisie des données par remplissage rapide de la grille offerte à l'utilisateur.

b- Un module d'actualisation: offrant la possiblité d'opérer la modification de la pondération des variables et des individus (opération très pratiquée en analyse des données).

c- Un Module de transformation: offre la possibilité d'effectuer un changement de variables. Les transformations permises sont :
- Transformation d'une variable quantitative en une variable qualitative en utilisant l'un des deux types de découpages :
- Le découpage par bornes choisies par l'utilisateur.
- Le découpage par intervalles égaux.
- Centrage des données
- Réduction des données

❒ **La bibliothèque de programmes:** Elle contient les différentes méthodes de classification implémentées et demeure facilement extensible grâce à l'approche de conception adoptée.
L'instanciation, s'est opérée sur deux méthodes de classification:
- La méthode de partitionnement NHD (*Non Hiérarchique Descendante)* [FAG80]
- La méthode CAH (*Classification Ascendante Hiérarchique selon le critère du saut minimum*) [JAM89].
L'extension de la bibliothèque en cours garantira aussi des méthodes de classifcation de données probabilistes.

a - Le Module sélection de programmes: Ce module constitue le noyau du logiciel, permet le traitement des données à classifier par une des méthodes de la bibliothèque.
- La sélection de la distance : est effectuée avant la sélection de la méthode oncernée par la cession. Une boîte de dialogue offre une liste de distances selon le tableau d'étude.

- La sélection de la méthode: Dans le cas où la méthode NHD est choisi, le système offre la possibilité de deux modes de fonctionnement de la méthode :
+ Mode manuel
+ Mode Automatique
En mode manuel (usuel pour ce type de méthodes), l'utilisateur indique le nombre de partitions à obtenir, par contre en mode automatique proposé, l'utilisateur est déchargé de cette tâche souvent difficile, le système se charge de déterminer le nombre de partitions à obtenir suivant un critère bien déterminé (arrêt si la variation en cours de la variance active est cinq fois inférieure à celle de la première) [GHO94].

b- Module Résultats et Interprétation : Le logiciel effectue l'interprétation de la classification en se basant sur des calculs numériques qui mettent en lumière les éléments saillants de la classification (calcul des indicateurs d'interprétation).[CHE94]
Pour une bonne présentation et une meilleure compréhension des résultats, le logiciel s'appuie sur la représentation graphique et met à la disposition de l'utilisateur plusieurs outils notamment:

- L'édition des valeurs des indices d'interprétation sous forme tabulaire concernant :
  + une partition : Pouvoir discriminant moyen des variables vis à vis de la partition R.
  + les variables: paramètres statistiques de chaque variable (moyenne, variance, min, max).
  + valeurs des indicateurs suivants: COR(j) (indicateur mesurant le pouvoir discriminant de la variable par rapport à la partition.) et CTR(j) (indicateur mesurant la contribution relative de la variable)
  + Paramètres statistiques par classe : (variance, moyenne , min, max, CORj, CTRj)
  + les classes : centres de gravité des classes de la partition et autres indicateurs
  + les Variables/Classes : indicateurs suivants :COR(j,l), CTR(j,l), DIS(j,l), CE(j,l)

- Graphe laticiel de filiations: permet d'afficher la suite de partitions fournis par la méthode liées par des filiations. Dans ce graphe l'utilisateur peut voir la formation des classes, le déplacement des individus d'une classe à une autre).

- Arbres hiérarchiques (ou dendogrammes): offrir à l'utilisateur la facilité de sélectionner l'arbre hiérarchique à divers niveaux jugés pertinents au cours de l'interprétation, ce qui veut dire obtenir une partition.

c. Module de génération de Rapport : A la demande de l'utilisateur, un rapport général d'interprétation de la classification est édité pour la partition souhaitée. il comprend :

- Les données du problème
- Les résultats obtenus entre autre :
  + Le degré de représentativité de la partition par rapport aux données de départ.
  + La classe la plus concentrée.
  + La classe la plus excentrée.
  + La variable de pouvoir discriminant plus fort.
  + L'importance des variables dans la formation des classes.

CONCLUSION: Le but de cet article était de montrer les apports de la modélisation objet sur deux plans: garantir des produits flexibles (réutilisation du code - extensibilité par spécialisation - Extensibilité par enrichissement - réactivité) et offrir des logiciels d'une réelle convivialité importante pour des utilisateurs novices .

Cette approche de modélisation et d'implémentation est d'autant plus attrayante car elle fournit un cadre de développement intégré et des ouvertures certaines, quant on sait l'intérêt commun de la plupart des méthodes d'analyse de données multidimensionnelles en amont et en aval du processus.

Dans la suite de nos travaux nous développons actuellement une composante didactique autour de l'environnement et à l'avenir, intégrer des techniques de validation des résultats et bénéficier des fonctionnalités d'un SGBD objet.

Une réflexion est aussi menée en collaboration avec l'université lumière LYON II, pour aboutir à un environnement intelligent pour l'apprentissage numérique.

D'autres directions de recherche pourraient être poursuivies, par exemple celle qui consisterait à doter cet environnement d'une intelligence dans le choix des méthodes et des outils adéquats au contexte statistico-informatique du domaine à étudier [CEL89].

REFERENCES BIBLIOGRAPHIQUES :

[AUB91] **AUBERT J.P/DIXNEUF P.** « *Conception et programmation par objets. Techniques, outils et applications* » Ed MASSON, 1991

[AUR90.2] **AURAY J.P/DURU G./ZIGHED A.** *"Analyse de données multidimensionnelles 2. Les méthodes de structuration"* Ed. Alexandre-Lacassagne-Lyon,1990

[AUR90.4] **AURAY J.P/DURU G./ZIGHED A.** *"Analyse de données multidimensionnelles 4. Aspects méthodologiques"* Ed. Alexandre-Lacassagne-Lyon,1990

[BEL92] **BELAID A. BELAID Y.** « *Reconnaissance des formes: Méthodes et Applications*". Ed. InterEditions, 1992, pages 127-257

[BOU93] **BOUZEGHOUB M.** "Les méthodes de conception orienté-objet: Les interfaces Homme-Machine". Labo. MASI, Univ. Pierre et Marie Curie, Actes de la 2ème Ecole Maghrébine Annaba, 6-12 Novembre 1993

[CEL89] **CELEUX G.- DIDAY E.- GOVAERT G.- LECHEVALLIER Y. RALAM BONDRAINY H.** *"Classification automatique des données:Environnement statistique et informatique"*

[CHA81] **CHANDON J.L - PINSON S.** *"Analyse typologique: Théories et applications"* Edition MASSON, 1981

[CHE94] **CHEBINE M.S / HAMZAOUI S.** « *Mise en ouvre d'outils d'aide à l'interprétation de résultats issus d'une classification automatique* » Mémoire d'Ingénieur, I.N.I, Octobre 1994

[COA92] **COAD P - YOURDON Y.** *"Analyse orientée objet"* Edition Masson, 1992

[FAG80] **FAGES R.** *"Cours de statistiques: classification automatique"* tome X , 1980

[GHO94] **GHOMARI A.R** « *E.M.C.A : Environnement pour les Méthodes de Classification Automatique de données multidimensionnelles.* » Thèse de magister, I.N.I (ALGER), Janvier 1994

[JAM89] **JAMBU M.** *"Exploration Informatique et statistique des données"* Ed. Dunod Informatique ,1989

[HAT91] **HATON J.P - BOUZID N. - CHARPILLET F. -HATON M.C - LAASRI B. - LAASRI H. - MARQUIS P. - MONDOT T. - NAPOLI A.** "L*e raisonnement en I.A, modèles,techniques et architectures pour les systèmes à bases de connaissances*" Ed. Intereditions ,1991, pages 313-361

[LER81] **LERMAN I.C** *"Classification et analyse ordinale des données"* Dunod, 1981

[MAR87] **MARTIN J.P** *"La qualité des logiciels"* Ed. AFNOR, 1987

[MAR91] **MARCOTORCHINO F.** *"La classification automatique aujourd'hui: bref aperçu historique, applicatif et calculatoire"* Ed. publications scientifiques et techniques d'IBM FRANCE / novembre 1991

[NIC88] **NICOLOYANNIS N.** *"Structures prétologiques et classification automatique: Le logiciel DEMON"* Thèse de Doctorat (Lyon I) , Decembre 1988

[THE88] **THERON P.** « *Guide pratique du Génie Logiciel* » Ed. Eyrolles, 1988