Samir Chergui
Courriel : chergsam@yahoo.fr
Mobile : 0658122841

Synthèse des méthodes statistiques vues dans le cadre du cours de STA211 (Entreposage et fouille de données).

Dans le cadre du cours de STA211, nous avons pu étudier des méthodes de fouilles de données, méthodes qui selon Tufféry sont l'application de techniques de statistique, d'analyse des données et d'intelligence artificielle à l'exploration et l'analyse dans a priori de grandes bases de données informatiques en vue d'en extraire des informations nouvelles et utiles pour le détenteur de ces données.

Pour essayer de condenser ce qui a été vu nous commencerons par rappeler très brièvement le contexte de l'émergence de ces nouvelles méthodes puis nous énumérerons chacune des techniques vues dans le cadre du cours en mettant en évidence pour chacune leurs apports et leurs éventuelles limites.

1/Contexte et Application du Datamining

Les dernières décennies ont été le théâtre de chamboulements majeurs liés au développement des nouvelles technologies et à leur démocratisation.
En effet, nombre de ces techniques (liées au datamining) existait mais il était difficile voire impossible de les utiliser sans coûts exorbitants. Les nouvelles technologies ont non seulement permis d'améliorer exponentiellement les temps de calculs mais ont été l'origine d'une explosion de la quantité de données disponibles et donc à traiter.
La quantité de données étant en constante et forte augmentation, il est devenu nécessaire pour les entreprises et administrations d'exploiter ces données de nature très diverses (on parle de Big-Data) pour en tirer des informations permettant une hausse de la valeur ajoutée et éventuellement une compétitivité accrue.
L'exploitation de ces données nécessite l'utilisation de techniques, des compétences particulières, compétences liées à ce qui est appelé le Datamining.
Le datamining fait étymologiquement référence à la recherche d'objets de valeurs dans le cadre de fouilles minières, fouilles étant fastidieuses et parfois dénuées de résultats exploitables.
Il devient donc important de prendre en compte les caractéristiques des techniques utilisées afin de maximiser les chances d'aboutir à un résultat pertinent et exploitable.
Cela passe par une étude de Datamining, étude nécessitant de profondes connaissances non seulement techniques mais aussi contextuelles (liées à la problématique de l'entreprise).

Transformer des données en informations exploitables nécessite un processus complexe comprenant la collecte des données jusqu'à la présentation des résultats liés à ces données.

La découverte de connaissances dans les bases de données (Knowledge Discovery in Data bases) se fait de la manière suivante :

->La sélection de variables pertinentes (souvent associées à des individus)
->Le prétraitement des données afin de prendre en compte les manquements, les redondances et autres anomalies
->Transformer les données (si besoin)
-> Choisir une technique de Datamining pertinente liée à la problématique rencontrée. Nous verrons par exemple qu'il est peu pertinent d'utiliser des réseaux de neurones si l'on cherche à faire de l'interprétation.
->Enfin, il est indispensable d'interpréter les résultats obtenus et les évaluer en tenant compte des résultats empiriques. Il est souvent nécessaire de réitérer le processus en amont pour améliorer la pertinence du processus.

Pour conclure sur cette (indispensable) introduction, il faut souligner qu'il existe deux principales méthodes de Datamining liées à la structure des données :

→Les méthodes non-supervisées

Elles prennent en compte les patterns et sont caractérisés par l'étude d'analyse factorielle, classification voire méthode de recherche d'association

→ Les méthodes supervisées

Elles prennent en compte les modèles et permettent de faire de la prévision et de l'interprétation (généralement) plus facilement que les méthodes non-supervisées.
Elles sont caractérisées par l'analyse de modèle de régression ou de classement selon la nature de la variable à expliquer. Il existe de nombreux modèles différents qui ont chacune leurs spécificités.

Il est possible de mixer ces deux méthodes, ceci afin d'améliorer la robustesse du modèle.

2/Analyse de données : méthodes descriptives (ACP, ACM, Classification hiérarchique, K-means)

Dans le cadre du processus Knowledge Discovery in DataBases, il est indispensable (dans un premier temps) d'analyser les données séparément une par une pour identifier les patterns et permettre une interprétation pertinente qui sera source de profit pour le Dataminer.

On commence donc par l'analyse uni variée pour identifier la distribution de chaque individu, l'existence d'éventuelles erreurs aberrantes, on peut à ce titre regarder un histogramme ou constater si les distributions suivent une loi normale via l'analyse de la moyenne, la médiane, l'écart interquartile ou la variance.

Ensuite, nous pouvons regarder s'il existe des corrélations entre individus via l'analyse bi-variée. Un coefficient de corrélation qui serait par exemple supérieur à 0,8 en valeur absolue montrerait un fort lien linéaire positif ou négatif. Un coefficient égal à 0 montrerait

l'inexistence de lien linéaire. Il important de dire ici que l'inexistence de lien linéaire n'implique pas d'indépendance entre les variables. Pour l'analyse de variables qualitatives, on peut utiliser le khi deux pour mesurer la liaison entre deux variables.

Enfin, l'analyse multivariée permet de prendre en compte un nombre de variables plus élevé ce qui permet de gagner un gain de temps significatif lorsque le nombre de variables est très élevé.

Il existe de nombreuses méthodes d'analyse multivariée qui peuvent nous permettre de mettre en évidence la structure des données pour une éventuelle analyse ultérieure. Nous avons en effet des méthodes d'analyse factorielle mais aussi de classification qui reposant sur des approches métriques permettent de condenser un jeu de données à l'aide de projections de moindre dimension.

Commençons par parler de l'analyse en composantes principales (ACP), nous verrons juste après l'analyse factorielle des correspondances (AFC) et l'analyse des correspondances multiples (ACM) puis l'analyse factorielle multiple (AFM).

2.1/L'Analyse en composantes principales

Nous avons une matrice centrée composée de n individus et p variables (Dans le cadre de l'ACP, nous travaillons uniquement sur des variables quantitatives). Cette matrice représentable comme un nuage de points a un nombre de points égal au nombre d'individus. Ces individus ayant des profils différents ont des coordonnées différentes dans le nuage de points.
L'ACP permet d'analyser la diversité des profils des individus en visualisant la projection de chaque individu dans un sous-espace, ce sous-espace est un condensé optimal du nuage initial (représentant la matrice originelle).
Pour mesurer la qualité de projection, nous prendrons en compte le critère de l'inertie.
Plus la distance entre deux individus est importante et plus l'inertie le sera.
L'objectif de l'ACP sera donc de fournir une inertie la plus importante possible reflétant les différences entre les individus.

Un préalable indispensable est de définir l'inertie du nuage de points. Dans la mesure où l'on est le plus souvent confronté à des variables très différentes (en termes d'unités par exemple), on utilisera des variables centrées et réduites (ACP normée) pour annihiler les effets indésirables de la variance sur le nuage de points.

Enfin, les axes du nuage de points devront être orthogonaux les uns des autres et on devra prendre en compte les axes dans un ordre précis :
Le premier axe devra être l'axe maximisant l'inertie des points projetés, le deuxième axe (orthogonal au premier) devra avoir une inertie la plus grande possible mais inférieure à l'inertie du premier axe et ainsi de suite.
L'inertie de chaque axe étant appelé valeurs propres, l'objectif de l'ACP est la recherche d'axes principaux maximisant la valeur propre.

Dans le cadre de l'interprétation de l'ACP, il est indispensable d'avoir le nuage de points et le cercle des corrélations pour pouvoir mettre en évidence quelles individus se ressemblent ou non ; ces derniers sont caractérisés par un taux d'inertie en pourcentages, plus celui-ci est élevé et mieux les axes associés expliquent les données initiales.

Pour choisir un nombre d'axes dans le cadre de l'ACP, il existe plusieurs techniques telles que la règle de Kaiser (prenant en compte le nombre de valeurs propres supérieurs à 1) ou encore la règle du coude (où l'on cherche un axe dont l'inertie est largement inférieure à l'inertie de l'axe précédant, cet axe sera le premier à ne pas être pris en compte).

2.2/L'analyse factorielle des correspondances

L'analyse factorielle des correspondances noté AFC permet de résumer et de visualiser un tableau de contingence, c'est-à-dire un tableau croisant deux variables de type qualitative. Ce tableau donne au croisement de la ligne « i » et la colonne « j », le nombre d'individus prenant la modalité i de la première variable et la modalité j de la seconde variable.
L'AFC permet donc de comparer les lignes et les colonnes entre elles pour interpréter les associations lignes/colonnes pour ensuite visualiser les associations des modalités des deux variables qualitatives.

2.3/L'analyse des correspondances multiples

L'ACM est une extension de l'ACP sauf que cette fois ci, les variables prises en compte sont des variables de type qualitatives. Ici, les variables qualitatives sont recodées sous la forme d'un tableau disjonctif complet, chaque variable qualitative sera remplacée par autant d'indicatrices que le nombre de modalités qu'elle possède.
L'interprétation de l'ACM sera assez différente de l'ACP dans la mesure où même si l'on obtient les axes d'une façon similaire (conservation des axes ayant les valeurs propres les plus élevées), on devra prendre en compte les modalités des variables qualitatives en les comparant avec les coordonnées des individus prenant ces modalités.

2.4/L'analyse factorielle multiple

L'analyse factorielle multiple noté AFM permet d'étudier des jeux de données où un même ensemble d'individus est décrit par des variables structurées en groupes. L'AFM équilibre l'influence de chaque groupe de variables quelques soient leur type (quantitatif ou qualitatif).
L'AFM peut également être assimilée une ACP particulière (lorsque les variables sont uniquement quantitatives) prenant en compte des tableaux disposés les uns à côté des autres.

Après avoir vu les méthodes d'analyse factorielle, regardons maintenant les méthodes de clustering des K-means puis la classification ascendante hiérarchique.

2.5/Méthodes de partitionnement

Malheureusement, dans le cadre des analyses factorielles, il était impossible de partitionner les individus en classes à peu près homogènes.
En effet, ces méthodes ne permettent pas d'affecter un individu ou une variable à un objet ou une classe de façon automatique. Les méthodes de partitionnement permettent donc de partitionner l'ensemble des individus en K groupes.
La méthode de partitionnement dite « classique » qui sera prise en compte sera la méthode des K-means.

Cette méthode se caractérise par le rassemblement des individus autour des centres mobiles. La notion de ressemblance entre les individus se fait à l'aide de la distance euclidienne. La caractérisation de la distance est fondamentale pour comprendre ce qui différencie intrinsèquement les individus.
A partir de là, on peut définir l'inertie de l'ensemble des individus, inertie qui permettra de décrire la qualité de la partition.
Selon la relation de Huygens, nous pouvons décomposer l'inertie totale en deux types d'inerties complémentaires : l'inertie interclasses et l'inertie intraclasses.
L'inertie totale étant selon la relation de Huygens la somme de l'inertie interclasse et intra classe.
Maximiser la qualité de la partition revient à minimiser l'inertie intra-classe et/ou à maximiser l'inertie interclasse dans la mesure où l'inertie totale est constante. De cette manière, nous aurons des groupes bien séparés les uns des autres tout en ayant une importante homogénéité des individus qui sont dans le même groupe.
La recherche de la meilleure partition ne peut aboutir qu'à un minimum local car une recherche exhaustive (amenant à un minimum global) nous amènerait à un nombre de calculs beaucoup trop long.

L'algorithme de réaffectation des centres mobiles (utilisé dans la méthode des K-means) proposé par Forchy est le suivant :
a/ Initialisation : sélectionner K points dans l'espace des individus
b/Répéter le processus suivant: allouer chaque individu au centre (la classe) le plus proche au sens de la distance choisie, calculer le centre de gravité de chaque classe qui devient le nouveau noyau et s'arrêter si le critère d'inertie intraclasse ne diminue plus (=l'inertie interclasse n'augmente plus).

2.6/Classification ascendante hiérarchique

La classification ascendante hiérarchique permet de regrouper les individus ou variables selon leurs similarités à l'aide d'une suite de partitions imbriquées les unes dans les autres.
Comme avec les K-means, il est nécessaire de définir au préalable une métrique commune entre les individus, métrique qui nous permettra d'identifier les individus les plus proches pour ensuite définir une partition à K classes (ceci en coupant l'arbre).
Dans la CAH, on peut choisir directement le nombre de groupe que l'on souhaite en coupant l'arbre. Il est aussi important de réfléchir à comment effectuer les regroupements entre individus ou groupes d'individus, c'est le critère d'agrégation.
Plusieurs critères d'agrégation peuvent être utilisés mais la plus connue est la distance de Ward qui permet de conserver une inertie interclasse maximale.

3/ Autres méthodes non-supervisées

Parmi d'autres méthodes non-supervisées, nous allons parler des règles d'association, des cartes de Kohonen et l'analyse des données multivues.

3.1/ Règles d'association

L'objectif des règles d'association est de trouver des règles liées à la structure de sous-ensembles. L'origine de ces règles étant marketing.
On recherchait en effet quels produits étaient achetés conjointement et à quel moment. Par exemple, lorsqu'un client achète du pain et du beurre, il achète aussi 90% du temps du lait. On peut soit rechercher des règles dont le support (proportion de transactions contenant l'ensemble des éléments de (au moins) deux sous-ensembles) est supérieur ou égal à un certain seuil ou rechercher d'autres règles dont la confiance (proportion de transactions contenant les éléments de A et B par rapport aux transactions contenant les éléments de A) est supérieure à un seuil minimal.
La pertinence des règles et les seuils étant donnés par le dataminer.

L'approche métier est indispensable pour fixer le support et la confiance car l'on peut vite obtenir un nombre de règles très conséquent. Il existe de nombreux algorithmes différents permettant de limiter le temps de calcul.
On peut énumérer l'algorithme Apriori, Partition ou Eclat (chacun ayant ses avantages et inconvénients), algorithmes qui permettent d'obtenir les mêmes ensembles fréquents car la recherche de ces ensembles est déterministe. La mesure de l'intérêt des règles peut aussi être obtenue à l'aide de tests statistiques (de significativité par exemple).

3.2/Cartes de Kohonen

Les cartes de Kohonen permettent de former des groupes d'individus homogènes dans des sous-espaces de l'espace des individus, c'est une méthode de classification qui permet donc de comprendre et analyser la structure des données multidimensionnelles.
La méthode est similaire à celle des K-means sauf que les centres des classes des individus sont contraints de rester dans leurs sous-espaces dédiés.
Les cartes de Kohonen permettent donc de faire de la classification en grande dimension.
Basées sur les réseaux de neurones, il s'agit d'un processus incrémental d'auto-organisation qui cherche à projeter des données dans un espace de faible dimension tel un réseau à deux couches entièrement connectées.
La carte est choisie indépendamment de la structure des données et donc ne permet pas de capter des informations liées à la structure de ces données, les voisinages entre classes peuvent être choisis de manière variée. Les classes peuvent être disposées sur des grilles rectangulaires voire hexagonales.

L'algorithme de Kohonen est le suivant :
1/Initialisation : initialiser les vecteurs référents initiaux de manière aléatoire, fixer la structure et la taille de la carte puis fixer le nombre d'itérations.

2/Choix au hasard d'une observation à chaque étape, on compare l'ensemble des référents pour déterminer la classe gagnante (celle dont le référent est le plus proche au sens d'une distance donnée)
3/Phase d'adaptation : Détermination de la taille du voisinage, du pas d'apprentissage et modification des poids des neurones et adaptation des poids des neurones.

Les cartes de Kohonen permettent aussi d'effectuer un prétraitement avant d'effectuer une classification ascendante hiérarchique en réduisant au préalable la complexité des données considérées.

3.3/Analyse de données multi vues

L'analyse de données multi vues représente des méthodes d'analyse de plusieurs tableaux de données de façon simultané. Il peut être en effet utile d'analyser plusieurs tableaux conjointement lorsque ceux-ci mettent en avant pléthore de variables caractérisant les individus. Finalement, l'analyse multi vues permet de comparer des groupes de variables et d'analyser la typologie des individus simultanément.
Il existe 4 étapes liées à l'analyse multi vues :

- ➔ Étude de l'interstructure : On assimile des tableaux à des objets, objets dont le choix dépend de la méthode d'analyse employée puis on compare les objets entre eux pour y déceler d'éventuels groupes homogènes.
- ➔ Compromis : Détermination d'un espace commun de représentation pour résumer l'ensemble des données des objets
- ➔ Étude de l'intrastructure : on analyse le compromis pour voir les éventuelles ressemblances entre les objets
- ➔ Trajectoires : on compare les profils des individus ou des variables selon les différents groupes.

Les méthodes liées à l'analyse de données multi vues que l'on a pu voir sont :

- ➔ La double ACP : L'ACP n°1 prend en compte le nuage des centres de gravité comme étude de l'interstructure et l'ACP n°2 permet la recherche d'un compromis résumant au mieux les différents nuages (maximisation de l'inertie en sélectionnant les meilleurs axes (les valeurs propres seront sélectionnées par ordre décroissant)
- ➔ STATIS/ STATIS DUALE : uniquement sur variables quantitatives, ces méthodes s'appliquent sur des tableaux ayant soit des individus identiques (STATIS) soit des variables identiques (STATIS DUALE)
- ➔ L'AFM : voir page 4

4/Méthodes supervisées

L'objectif des méthodes supervisées est de prédire une variable Y à l'aide de variables explicatives notées X.
Parmi les méthodes supervisées, nous allons parler des arbres de décision, des support vector machine (SVM) et des réseaux de neurones.

4.1/Arbres de décision

Les arbres de décision sont des méthodes non paramétriques permettant de partitionner ou segmenter l'espace des variables explicatives en un certain nombre de régions. Chaque région est ensuite associée à un modèle simple qui permet de prendre en compte une prise de décision. Les arbres de décisions permettent d'étudier des données quelles que soient leurs nature, l'interprétation des arbres de décision étant directe et intuitive.
La construction d'un tel arbre se déroule en 4 étapes :

- ➔ Établir pour chaque nœud l'ensemble des divisions admissibles et déterminer un critère pour sélectionner la « meilleure »
- ➔ Définir une règle pour déclarer un nœud comme terminal ou intermédiaire
- ➔ Affecter chaque nœud terminal à l'un des groupes
- ➔ Estimation du coût associé à l'arbre

L'étape 1 est relatif au critère de division, celui-ci dépend de la nature de la variable (binaire, quantitative, nominale ou ordinale par exemple) et est basé sur la notion d'impureté à l'aide de l'indice de Gini par exemple. Les divisions ne se font tant qu'aucun des nœuds descendants n'est vide.

L'étape 2 est ici la plus délicate car il est nécessaire d'élaborer un seuil qui ne soit ni trop faible (donnera un arbre trop grand et faible taux de mal classés) ni trop élevé (taux de mal classés trop élevé), il existe plusieurs méthodes permettant d'estimer un seuil pertinent dont la méthode CART (élaboration d'un grand arbre puis élagage des branches les moins informatives) ou Breiman (on rajoute une procédure d'estimation des taux d'erreur des arbres de la séquence pour choisir un arbre satisfaisant, cela peut se faire par validation croisée par exemple à l'aide d'un échantillon test).

L'étape 3 consiste à affecter un nœud noté « a » qui a un coût moyen d'erreur de classement minimal. Ici, le risque de la règle d'affectation présente un biais car il est estimé à l'aide des données ayant servis pour la construction de l'arbre. Pour pallier ce biais, on peut faire un échantillon test puis une validation croisée.

4.2/ Support Vector Machine (SVM)

Les support vector machines SVM sont une famille d'algorithmes dédiés à la classification de type supervisée et aux problèmes de régression.
L'objectif des SVM est de trouver un hyperplan de l'espace des variables explicatives qui sépare de manière optimale les observations parmi une infinité d'hyperplans existants.
L'hyperplan optimal est la séparation linéaire qui donnera la plus grande marge entre les vecteurs supports.
Les vecteurs supports sont les points représentants les bords de la marge, ces points étant les plus difficiles à classifier, ils représentent un « worst-case scenario » dans le sens où si l'on sait classifier ces points alors à fortiori, il sera plus facile de classifier les points moins ambigus. En effet, on peut intuitivement se rendre compte que si l'on prend en compte un point autre que le vecteur support dans une classe donnée, alors la marge s'agrandit et plus le risque d'erreur de classification se réduit.
L'un des objectifs est donc de maximiser la marge afin d'obtenir un taux d'erreur de classification le plus faible possible.

Dans la plupart des cas, il n'existera pas de séparation linéaire pour la simple raison que les données ne sont jamais aussi simples. Dans le cas de vecteurs non linéairement séparables, il va falloir prendre en compte un paramètre de régularisation qui décide à quel point on peut faire des erreurs. Plus ce paramètre est « petit », plus les erreurs sont autorisées et plus la marge est large. L'enjeu sera donc de trouver le meilleur compromis entre peu d'erreurs et une marge importante. Il est possible d'optimiser ce paramètre de régularisation via la validation croisée.

En guise de conclusion, nous pouvons dire que les SVM permettent d'avoir d'excellent résultats en termes de classification, sont utilisables pour une multitude d'objets (images, sons, séquences ADN…) et permettent de sélectionner fiablement des variables.
L'inconvénient étant qu'il est difficile à interpréter les résultats que l'on obtient à l'instar des réseaux de neurones que l'on verra juste après. Il est aussi difficile et laborieux d'obtenir les paramètres optimaux pour le problème auquel on est confronté.

4.3/Réseaux de neurones

Les réseaux de neurones sont souvent considérés comme des boites noires aux résultats impressionnants que ce soit pour le traitement des sons, de l'image, des vidéos, la conduite autonome, l'imagerie médicale…
Ces réseaux peuvent être assimilés à des machines à prédire des valeurs ou des probabilités dont les paramètres sont appris grâce à des algorithmes efficaces d'optimisation de la vraisemblance. Les réseaux de neurones peuvent s'appliquer pour faire de la classification supervisée, de la classification (avec notamment les cartes de Kohonen) voire de la régression.

De façon plus concrète, les réseaux de neurones sont caractérisés par une multitude de neurones connectés les uns avec les autres. Commençons par appréhender ce qu'est un neurone.
Un neurone prend en entrée une observation « x » appartenant à l'ensemble des réels R de dimension « m », calcule le produit scalaire entre l'observation « x » et un vecteur de dimension « m » de poids noté « w » et ajoute un biais noté « b ».
Nous avons donc l'équation suivante : Y = f(w^t*x+b).
L'enjeu sera ici de sélectionner un poids et un biais permettant de minimiser le taux d'erreur, cela se fait classiquement en utilisant un algorithme d'optimisation de type descente de gradient.
Le principe des réseaux de neurones est d'empiler des couches de réseaux de neurones afin d'obtenir des fonctions de plus en plus riches et de plus en plus complexes.
Les couches cachés (c'est-à-dire les couches intermédiaires du réseau de neurones) peuvent approximer n'importe quelle fonction s'il y a un nombre suffisant d'unités cachés en output.

Pour avoir plus de choix sur la relation entre la variable d'entrée « x » et les probabilités en sortie, on peut utiliser la technique de « rétropropagation ». Cette technique consiste à interpréter les valeurs de la couche de sortie comme une nouvelle représentation de la variable d'entrée pour ensuite utiliser ce vecteur en entrée d'un autre algorithme

d'apprentissage cela nous amène à avoir une seconde couche en sortie la première couche devenant une couche cachée).
Il existe pléthore de méthodes permettant de rendre une couche cachée plus complexe et minimisant l'erreur. Le risque cependant est de tomber dans le sur-apprentissage si les modèles sont trop complexes par rapport au volume de données, le dataminer devra donc chercher à trouver la bonne architecture (profondeur du noyau) permettant la meilleure performance.

4.4/Agrégation de modèles

L'agrégation de modèles est le dernier ensemble de méthode que nous allons voir au cours de cette synthèse, il s'agit de méthodes supervisées qui ont pour but d'améliorer l'ajustement par combinaison ou l'agrégation d'un grand nombre de modèles en évitant le sur-ajustement. Le meilleur modèle sera celui qui maximise un critère de performance sur un ensemble test. La performance pouvant être la minimisation de l'erreur ou l'amélioration de la robustesse du modèle.
Le principe de l'agrégation de modèles est de moyenner les prédictions de plusieurs modèles pour réduire la variance et le taux d'erreur de prédiction. On peut soit utiliser plusieurs échantillons du jeu de données en y appliquant la même méthode (rééchantillonnage) soit utiliser différentes méthodes pour le même échantillon (hybridation).
Dans la mesure où il est difficile d'avoir un grand nombre d'échantillons indépendants, on peut utiliser des échantillons bootstrap (en tirant au hasard des observations dans l'échantillon, le tirage s'effectuant avec remise) qui permettent de diversifier l'échantillon pour couvrir différentes régions de l'espace des données d'apprentissage (l'objectif étant de minimiser l'erreur).
La diversification par échantillonnage comprend de nombreuses méthodes que nous allons voir telles que :

→ Le bagging (méthode aléatoire) qui permet la classification des arbres de décision en réduisant la variance de l'estimateur. Cela se fait à l'aide d'échantillons aléatoires qui moyennés (données quantitatives) ou par majorité absolue (données qualitatives) permettent de réduire la variance. C'est un algorithme simple à comprendre et à mettre en pratique cependant, le logiciel doit évaluer un nombre suffisant d'arbres ce qui peut amener à un temps de calcul important.
→ Les forêts aléatoires (méthode aléatoire) qui sont une amélioration du bagging, ajoutent une sélection aléatoire de variables pour rendre les arbres plus indépendants par le choix par le choix au hasard des variables qui entrent dans les modèles ; cela permet une baisse de la corrélation entre ceux-ci. Les forêts aléatoires sont utiles lorsque le nombre de variables est important. L'interprétation des forêts aléatoires se fait à l'aide de critères tels que la fréquence d'apparition de chaque variable ou encore les critères Mean Decrease Accuracy et Mean Decrease Gini.

→Le boosting (méthode adaptative) améliore les performances de n'importe quel algorithme d'apprentissage utilisé pour les classifieurs faibles, on part du bagging puis on accorde un poids plus important aux observations mal ajustées ou mal prédites. De cette manière, on concentre les efforts sur les observations les plus difficiles à ajuster.