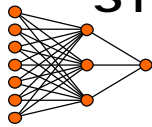


Stéphane Tufféry
Statisticien - Data Miner - Enseignant

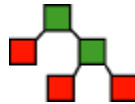


DATA MINING - SCORING

STATISTIQUE DÉCISIONNELLE



APPLICATION AU CRM



21/09/2004

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

1

Plan du cours

- Qu'est-ce que le data mining ?
- A quoi sert le data mining ?
- Les 2 grandes familles de techniques
- Le déroulement d'un projet de data mining
- Coûts et gains du data mining
- Facteurs de succès - Erreurs à éviter
- Informatique décisionnelle et de gestion
- La préparation des données
- Techniques descriptives de data mining
- Techniques prédictives de data mining
- *Logiciels et consultants*
- CNIL et limites légales du data mining
- Le text mining
- Le web mining

21/09/2004

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

2

Logiciels et consultants

21/09/2004

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



3

Les logiciels de data mining et statistique

- Il existe de nombreux logiciels de statistique et data mining sur PC :
 - faciles à installer et pas très chers
 - conviviaux avec des algorithmes de bonne qualité
 - bons pour des PME car pouvant gérer plusieurs dizaines de milliers voire quelques centaines de milliers d'individus
 - **SPAD™** de Decisia, **Alice™** de Isoft, **AnswerTree™** de SPSS, **Predict™** de Neuralware, **R** (version gratuite de S-PLUS) et **Weka** (freeware)...
- Cependant :
 - ils ne permettent pas de traiter exhaustivement de très grandes bases de données
 - ils ne mettent souvent en œuvre qu'une ou deux techniques (sauf quelques produits tels SPAD, R et Weka)

21/09/2004

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



4

Zoom sur les gros logiciels

- Certains logiciels sont conçus :
 - pour exploiter de grands volumes de données
 - pour couvrir une large palette de techniques
- Ils existent parfois en version « **statistique** » ou « **data mining** » (le 2nd étant parfois une sur-couche du 1^{er})
- Ils peuvent fonctionner en mode client-serveur
- Il s'agit de **SPSS™** et **Clementine™** de SPSS
- Et de **SAS/STAT™** et **Enterprise Miner™** de SAS
- Et de **Statistica Data Miner™** de StatSoft
- Et de **S-PLUS™** et **Insightful Miner™** de Insightful
- On peut ajouter **KXEN™**



21/09/2004

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

5

Statistique vs Data mining

Appellation commerciale	Logiciel de statistique	Logiciel de data mining
Plate-forme	PC ou client-serveur	PC ou client-serveur
Interface	Fenêtre de programmation ou menus déroulants	Icônes à déplacer et relier par des flèches
Algorithmes	Ce qui sert couramment - les arbres de décision (existent en logiciels spécifiques)	Comme « logiciel statistique » – beaucoup d'algos de statistique (ex : tests non paramétriques) et d'analyse des données (ex : analyse discriminante linéaire) + arbres de décision, réseaux de neurones, détection d'associations
Prix	Prix de revient de l'éditeur + une marge raisonnable	Plusieurs fois celui du logiciel de statistique !



21/09/2004

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

6

Logiciels de data mining 1/3 (poids légers : qq dizaines de milliers de lignes)

Produit	Spécialité (le cas échéant)	Editeur
Stat Lab		SLP InfoWare (Gemplus)
StartMiner	Réseaux de neurones – Arbres de décision	Grimmersoft
Alice	Arbres de décision	Isoft
Predict	Réseaux de neurones	Neuralware
Previa	Réseaux de neurones	Elseware
NeuroOne	Réseaux de neurones	Netral
Scenario	Réseaux de neurones	Cognos
Wizwhy	Détection d'associations	Wizsoft
WEKA		« open source » (logiciel gratuit)

21/09/2004

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



7

Logiciels de data mining 2/3 (poids moyens : qq centaines de milliers de lignes)

Produit	Spécialité (le cas échéant)	Editeur
SPAD		Decisia
AnswerTree	Arbres de décision	SPSS
4Thought	Réseaux de neurones	Cognos
KnowledgeSEEKER	Arbres de décision	Angoss
KnowledgeSTUDIO		Angoss
C5.0 (Unix) See5 (Windows)	Arbres de décision	RuleQuest Research
Data Mining Suite		Salford Systems
Darwin		Oracle
Polyanalyst		Megaputer
R		« open source » (logiciel gratuit)
S-PLUS		Insightful
SQL Server 2000	Arbres de décision – clustering	Microsoft

21/09/2004

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



8

Logiciels de data mining 3/3 (poids lourds : plusieurs millions de lignes)

Produit	Spécialité (le cas échéant)	Editeur
KXEN	algorithmes de Vapnik	KXEN
Intelligent Miner	Réseaux de neurones – classification relationnelle – détection d'associations	IBM
Statistica Data Miner		Statsoft
Insightful Miner		Insightful
SPSS (version C/S)		SPSS
Clementine		SPSS
SAS/STAT (version C/S)		SAS
Entreprise Miner		SAS

21/09/2004

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



9

Cartographie des logiciels

Logiciels multi-techniques	Decisia – SPAD Insightful – S-PLUS R (version gratuite de S-PLUS)	SAS – Entreprise Miner SPSS – Clementine Statsoft – Statistica Data Miner Insightful – Insightful Miner KXEN IBM – Intelligent Miner
Logiciels mono-techniques	Salford Systems – CART SPSS – Answer Tree Isoft – Alice Neuralware – Predict	
	Logiciels micros	Logiciels gros systèmes

21/09/2004

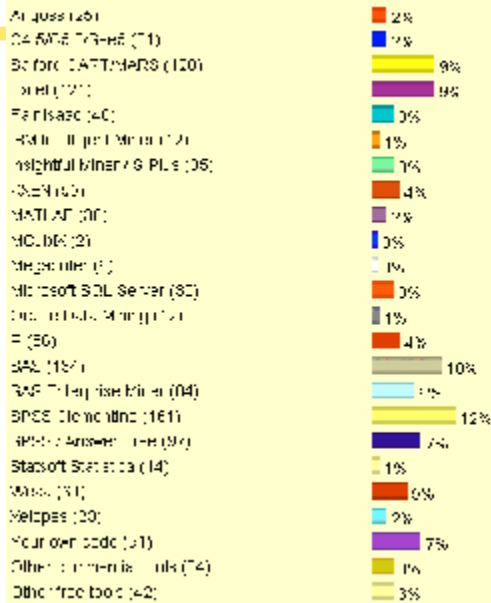
© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



10

Sondage sur www.kdnuggets.com

Data mining tools you regularly use: (500 respondents, 1324 votes total)



Sondage
effectué
en mai
2004

21/09/2004

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

13

Critères de choix d'un logiciel

- Variété des algorithmes de data mining, de statistique et de préparation des données
 - + simple d'avoir tout dans un seul outil
- Qualité des algorithmes implémentés
 - documentation éditeur pas toujours accessible
- Capacité à traiter de grands volumes de données
 - peut être cruciale à partir de plusieurs centaines de milliers d'individus à traiter
- Types de données gérées
 - exemple : choix influencé si l'entreprise possède déjà un infocentre SAS...
- Convivialité du logiciel et facilités à produire des rapports
- Prix !

21/09/2004

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

14

Ce qu'on peut attendre d'un logiciel 1/5

- Algorithmes de statistique et de data mining :
 - classement (analyse discriminante linéaire ou quadratique, régression logistique binaire ou polytomique, modèle linéaire généralisé, régression logistique PLS, arbres de décision, réseaux de neurones, k-plus proches voisins...)
 - prédiction (régression linéaire, modèle linéaire général, régression non-paramétrique, régression non-linéaire, régression PLS, arbres de décision, réseaux de neurones, k-plus proches voisins...)
 - classification (« clustering ») (centres mobiles, nuées dynamiques, k-means, classification hiérarchique, méthode mixte, réseau de Kohonen, analyse relationnelle...)
 - analyse des séries temporelles
 - analyse de survie
 - **détection des associations**

21/09/2004

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



15

Ce qu'on peut attendre d'un logiciel 2/5

- Fonctions de préparation des données
 - manipulation de fichiers (fusion, agrégation, transposition...)
 - visualisation des individus, coloriage selon critère
 - détection, filtrage et winsorisation des extrêmes
 - analyse et imputation des valeurs manquantes
 - transformation de variables (recodage, standardisation, normalisation automatique, discrétisation...)
 - création de nouvelles variables (fonctions logiques, chaînes, statistiques, maths...)
 - sélection des discrétisations, des interactions, des variables les plus explicatives et des variables les moins corrélées entre elles

21/09/2004

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



16

Ce qu'on peut attendre d'un logiciel

3/5

- Fonctions d'échantillonnage et de partition des données
 - pour créer des échantillons d'apprentissage, de test et de validation (l'échantillonnage stratifié doit être possible)
 - bootstrap, jacknife
- Fonctions statistiques
 - détermination des caractéristiques de tendance centrale, de dispersion, de forme...
 - tests statistiques de moyenne, de variance, de distribution, d'indépendance, d'hétéroscédasticité, de multicollinéarité...
- Fonctions d'analyse exploratoire des données et d'analyse factorielle
 - ACP, ACP avec rotation
 - AFC
 - ACM

21/09/2004

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



17

Ce qu'on peut attendre d'un logiciel

4/5

- Présentation des résultats
 - visualisation des résultats
 - manipulation des tableaux
 - bibliothèque de graphiques (2D, 3D, interactifs...)
 - incorporation dans un rapport
 - navigation dans les arbres de décision
 - affichage de la matrice de confusion
 - affichage des courbes de performances (ROC, lift, gain...)
 - indice de Gini, aire sous la courbe ROC
- Gestion des métadonnées
 - données définies identiquement pour tous les fichiers du projet (identifiant, cible, exclusions...)

21/09/2004

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



18

Ce qu'on peut attendre d'un logiciel 5/5

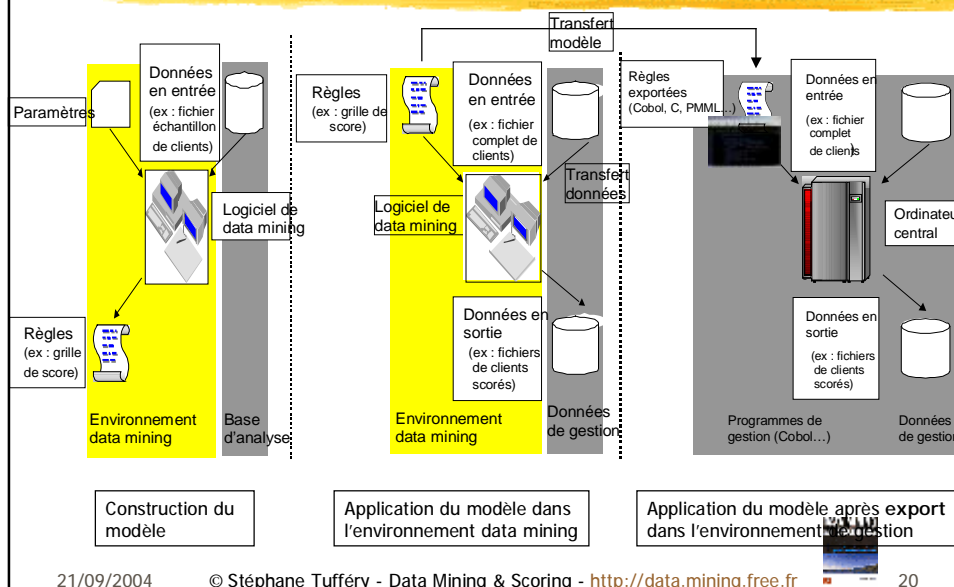
- Plates-formes supportées (Windows, Unix, MVS...)
- Formats d'entrée/sortie des données gérés :
 - tables Oracle, SQL Server, DB2, SAS, fichiers Excel, à plat...
- Langage avancé de programmation
- Enchaînements programmés de plusieurs algorithmes
- Portabilité des modèles construits (C, XML, Java, SQL...)
- Volume de données pouvant être raisonnablement traité
- Pour plus de puissance
 - architecture client-serveur : calculs sur le serveur et visualisation des résultats sur le client
 - algorithmes parallélisés
- Exécution en mode différé ou interactif

21/09/2004

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

19

Utilisation d'un logiciel



21/09/2004

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

20

Critères de performance des algorithmes 1/5

- Nous récapitulons dans les tableaux suivants les algorithmes avec leurs conditions d'utilisation sur 3 points essentiels :
 - l'absence d'hypothèses restrictives fortes préalables à la recherche
 - la capacité de traiter exhaustivement les données, en un temps raisonnable
 - la possibilité de manier des données lacunaires et de types hétérogènes (numériques ou non).

21/09/2004

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



21

Critères de performance des algorithmes 2/5

techniques	absence d' <i>a priori</i> dans la recherche	traitement toujours exhaustif des bases de données	traitement des données hétérogènes ou lacunaires
CLASSIFICATION			
classification relationnelle	oui	oui	discrétisation des var. continues
réseau de neurones de Kohonen	nb max segments spécifié	oui	les variables $\notin [0,1]$ sont transformées
méthode des centres mobiles et ses variantes	nb de segments fixé	oui	variables forcément numériques et sans valeurs manquantes
classification hiérarchique	oui, mais les segments au niveau n sont déterminés par ceux au niveau $n-1$	non (algorithme non linéaire)	oui (possibilité de traiter des variables non numériques avec une distance ad hoc)

21/09/2004

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



22

Critères de performance des algorithmes 3/5

techniques	absence d' <i>a priori</i> dans la recherche	traitement toujours exhaustif des bases de données	traitement des données hétérogènes ou lacunaires
CLASSEMENT ET PRÉDICTION			
arbres CHAID, CART, C5.0	comme la classification hiérarchique (qui est un arbre à l'envers)	non (comme la classification hiérarchique)	oui (sauf CHAID qui ne traite que les variables qualitatives)
réseau à fonction radiale de base	oui	oui	les variables $\notin [0,1]$ sont transformées
réseau de neurones à apprentissage supervisé	oui	non (pas d'apprentissage sur plusieurs centaines de variables)	les variables $\notin [0,1]$ sont transformées

21/09/2004

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

23

Critères de performance des algorithmes 4/5

techniques	absence d' <i>a priori</i> dans la recherche	traitement toujours exhaustif des bases de données	traitement des données hétérogènes ou lacunaires
analyse discriminante	relations linéaires entre les variables et hypothèses sur loi X_i/Y	oui	var. forcément numériques et sans valeurs manquantes
analyse discriminante DISQUAL (Saporta)	oui (permet de s'affranchir largement des hypothèses sur loi X_i/Y)	oui	var. sans valeurs manquantes (travaille sur les facteurs d'une ACM)
régression linéaire	relations linéaires entre les variables + autres hypothèses	oui	var. forcément numériques et sans valeurs manquantes
régression logistique	oui (simple hypothèse sur loi Y/X_i)	oui	var. sans valeurs manquantes

21/09/2004

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>

24

Critères de performance des algorithmes 5/5

techniques	absence d' <i>a priori</i> dans la recherche	traitement toujours exhaustif des bases de données	traitement des données hétérogènes ou lacunaires
ASSOCIATIONS			
analyse du panier de la ménagère	oui	oui	oui
séries chronologiques similaires	oui	oui	oui

21/09/2004

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



25

Internalisation ou externalisation 1/2

- Soit l'entreprise internalise l'activité de data mining, éventuellement avec l'aide de consultants spécialisés
- Soit elle externalise totalement cette activité, en fournissant ses fichiers de données à des prestataires spécialisés (les « credit-bureaux » pour la banque), ceux-ci lui restituant ses fichiers enrichis avec les informations de data mining qu'ils auront calculées (score, segment, etc.)
 - ne pas oublier de faire signer une clause de confidentialité
- Soit elle sous-traite la fabrication des modèles de DM, mais se les fait livrer, afin de les appliquer elle-même à ses fichiers.

21/09/2004

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



26

Internalisation ou externalisation 2/2

- L'intérêt du recours à des prestataires est de disposer immédiatement de leur savoir et de leur expérience.
- L'intérêt d'avoir des compétences en interne dans l'entreprise est de pouvoir :
 - acquérir une parfaite connaissance de ses données
 - avoir une plus grande réactivité lorsqu'une nouvelle étude est demandée
 - actualiser en permanence ses résultats
 - développer pour un coût bien plus faible quantité d'outils de score, de classification, de recherche d'association de produits... pour des besoins et des destinataires variés.

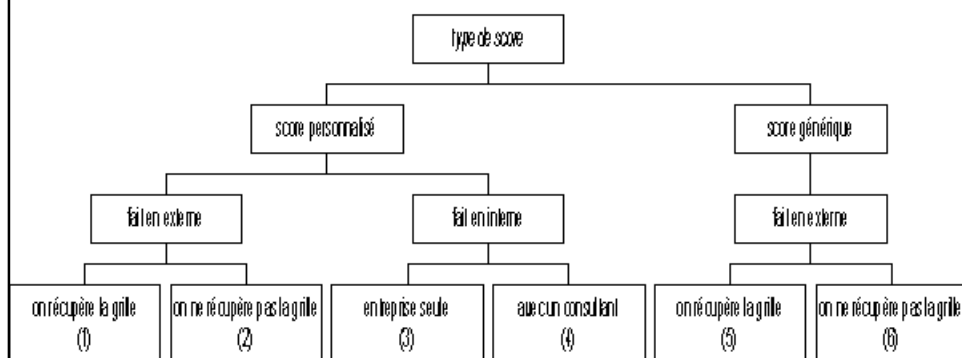
21/09/2004

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



27

Scores personnalisés et génériques



21/09/2004

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



28

Comparatif des diverses solutions

	Performance du score	Transfert de compétence	Pérennité à court terme du score	Pérennité à long terme du score	Rapidité d'obtention
(1)	+	-	+	-	-
(2)	+	-	– (sert 1 seule fois)	-	-
(3)	+	-	+	+	-
(4)	+	+	+	+	-
(5)	-	-	+	-	++ (30 jours)
(6)	-	-	– (sert 1 seule fois)	-	+

21/09/2004

© Stéphane Tufféry - Data Mining & Scoring - <http://data.mining.free.fr>



29