Université Paris 13 – Institut Galilée
Département d'Informatique
CumUE
Université Sorbonne Paris Cité

LICENCE INFORMATIQUE

3ème année

VISUALISATION DES DONNEES

Data Visualization

Projet

Visualisation des données L3-INFO 2016/2017

Projet 1 : Données de cœur SPECTF

Sources :

Les propriétaires d'origine : Krzysztof J. DSI , Lukasz Kurgan A.
Université du Colorado à Denver , Denver , CO 80217 , USA, Krys.Cios @ cudenver.edu
Lucy S. Goodenday
Medical College of Ohio , OH , U.S.A.
- Les bailleurs de fonds : Lukasz A.Kurgan , Krzysztof J. DSI
- Date: 10/01/01

Description :

SPECTF est une bonne base de données pour tester les algorithmes d'apprentissage. Elle est composée de 267 exemples décrits par 45 attributs. Attribut prédit : OVERALL_DIAGNOSIS (binaire). L'ensemble des données décrit le diagnostic d'images cardiaques du « Single Proton Emission Computed Tomography » (SPECT). Chacun des patients est classé en deux catégories : normal et anormal. La base des données de 267 séries d'images SPECT (patients) a été traitée pour extraire des caractéristiques qui résument les images SPECT originales.
En conséquence, 44 motifs caractéristiques en continu ont été créés pour chaque patient. L'algorithme de CLIP3 a été utilisé pour générer des règles de classification à partir de ces schémas. L'algorithme de CLIP3 généré règles qui étaient de 77% de précision (par rapport à des diagnostics de cardiologues).

Nombre d'instances : 267
Nombre d'attributs : 45 (44 en continu + 1 classe binaire)

L'ensemble de données est divisé en :
- Base d'apprentissage (« SPECTF.train » 80 instances)
- Base de test (« SPECTF.test « 187 cas)

La distribution des classes :
- Ensemble des données

Classe	# exemples
0	55
1	212

- Ensemble des données d'apprentissage

Classe	# exemples
0	40
1	40

- Ensemble des données de test

Classe	# exemples
0	15
1	172

Travail demandé :

Les bases de données sont disponibles à : http://archive.ics.uci.edu/ml/datasets.html

Ce qui est demandé dans ce travail de projet :

1- Effectuer une exploration des données par l'outil KNIME
2- Construire des chaines de traitement (workflow)
3- Interpréter les résultats graphiques
4- Faire une présentation PowerPoint de 5 minutes
5- Fournir une copie papier des transparents

Université Paris 13 – Institut Galilée
Département d'Informatique
CumUE
Université Sorbonne Paris Cité

LICENCE INFORMATIQUE

3ème année

VISUALISATION DES DONNEES

Data Visualization

Projet

Visualisation des données L3-INFO 2016/2017

Projet 2 : Données Courriers indésirables (E-mail SPAMs)

Sources :
Créateurs : Mark Hopkins, Erik Reeber , George Forman , Jaap Suermondt
Hewlett-Packard Labs, 1501 page Mill Rd. , Palo Alto, CA 94304
Donateur : George Forman (gforman à nospam hpl.hp.com) 650-857-7835
Générées : Juin-Juillet 1999

Description :

Le concept de « spam» est diverse : publicités pour des produits / les sites web, font chaînes de lettres, la pornographie ...
Notre collection de spams venait de notre gestionnaire de courriers et les personnes qui avaient déposé le spam. Notre collection de non–spam provenaient de travail déposée et des e-mails personnels, et donc le mot «George» et le code de zone '650 ' sont des indicateurs de non-spam. Ils sont utiles lors de la construction d'un filtre personnalisé anti-spam.

Pour des informations sur les spams :
Cranor , Lorrie F. , LaMacchia , Brian A. Spam !
Communications de l'ACM, 41 (8) :74 -83 , 1998 .

Nombre d'instances : 4601 (1813 Spam = 39,4%)
Nombre d'attributs : 58 (57 étiquettes en continu, une classe nominale)

La dernière colonne du « spambase.data » indique si l'e-mail était considéré comme du spam (1) ou pas (0). La plupart des attributs indiquent si un mot particulier ou caractère est souvent produit dans l'e-mail. Le terme de longueur attributs (55-57) mesurent la longueur des séquences de majuscules consécutives. Pour les mesures statistiques de chaque attribut, voir la fin de ce fichier.

Voici les définitions des attributs :

48 réels continus [0,100] les attributs de type word_freq_WORD = Pourcentage de mots dans l'e-mail qui correspond WORD, soit 100 * (nombre de fois où le mot apparaît dans l'e-mail) / nombre total de mots dans l'e-mail . Un " mot " dans ce cas est toute chaîne de caractères alphanumériques limitées par des caractères non-alphanumériques ou en fin de chaîne.

6 réels en continu [0,100] les attributs de type char_freq_CHAR = Pourcentage de caractères dans l'e-mail qui correspond à CHAR, soit 100 * (nombre d'occurences CHAR) / nombre total de caractères dans l'e-mail.

1 attribut continu réel [1, ...] de type capital_run_length_average = Longueur moyenne des séquences ininterrompues de lettres majuscules.

1 entier continu [1, ...] attribut de type capital_run_length_longest = Longueur de la plus longue séquence ininterrompue de lettres majuscules.

1 entier continue [1, ...] attribut de type capital_run_length_total = Somme de la longueur de séquences ininterrompues de lettres majuscules = Nombre total de lettres majuscules dans l'e-mail.

1 attribut nominal {0,1} nom de la classe = indique si l'e-mail a été considéré comme du spam (1) ou pas (0).

La distribution des classes :

Spam	1813 (39,4 %)
Non -Spam	2788 (60,6 %)

Travail demandé :

Les bases de données sont disponibles à : http://archive.ics.uci.edu/ml/datasets.html

Ce qui est demandé dans ce travail de projet :

1- Effectuer une exploration des données par l'outil KNIME
2- Construire des chaines de traitement (workflow)
3- Interpréter les résultats graphiques
4- Faire une présentation PowerPoint de 5 minutes
5- Fournir une copie papier des transparents

Université Paris 13 – Institut Galilée
Département d'Informatique
CumUE
Université Sorbonne Paris Cité

LICENCE INFORMATIQUE

3ème année

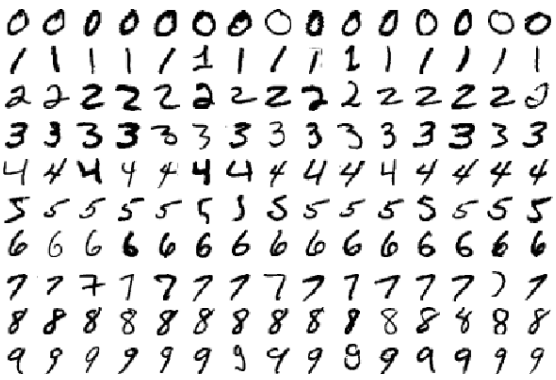
VISUALISATION DES DONNEES

Data Visualization

Projet

Visualisation des données L3-INFO 2016/2017

Projet 3 : Données chiffres manuscrits « Semeion Handwritten Digit »



Sources :

Le jeu de données a été créé par Tactile Srl , Brescia , Italie (http://www.tattile.it/) en 1994 au centre de recherche Semeion des Sciences de la Communication , Rome , Italie (http://www.semeion.it/), pour recherche sur l'apprentissage artificiel.

Description :

1593 chiffres manuscrits de près de 80 personnes ont été scannés, étirés dans une boîte 16x16 rectangulaire sous une échelle de niveaux de gris de 256 valeurs .

Ensemble de données caractéristiques : multivariées
Nombre d'instances : 1593
Caractéristiques : Entiers
Nombre d'attributs : 256
Date de Don : 2008-11-11
Tâches associées : Classification

1593 caractères manuscrits de près de 80 personnes ont été scannés, tendu dans une boîte 16x16 rectangulaire dans une échelle de gris de 256 valeurs. Chaque pixel de chaque image a été mis à l'échelle en binaire (1/0) en utilisant une valeur seuil fixe.
Chaque personne a écrit sur un papier tous les chiffres de 0 à 9, deux fois. La consigne était d'écrire le chiffre la première fois de façon normale (essayer d'écrire chaque chiffre avec précision) et la deuxième fois de manière rapide (sans précision).
Le meilleur protocole de validation pour cet ensemble de données semble être une 5x2 Cross Validation, 50 % (Apprentissage+ test) et 50 % de validation.

Cette base de données se compose de 1593 enregistrements (lignes) et 256 attributs (colonnes). Chaque enregistrement représente un chiffre manuscrit, avec une résolution de 256 gris échelle. Chaque pixel de l'image numérisée est d'abord étiré, et après mise à l'échelle entre 0 et 1 (mise à 0 à chaque pixel dont la valeur était sous la valeur 127 de l'échelle de gris (127 inclus), et mise à 1 de chaque pixel dont la valeur était plus que 127). Enfin, chaque image binaire a été réduite à nouveau dans une boîte carrée 16x16 (les 256 attributs binaires finaux).

Documents pertinents:
M Buscema, MetaNet: The Theory of Independent Judges, in Substance Use & Misuse 33(2)1998, pp 439-461.

Travail demandé :

Les bases de données sont disponibles à : http://archive.ics.uci.edu/ml/datasets.html

Ce qui est demandé dans ce travail de projet :

1- Effectuer une exploration des données par l'outil KNIME
2- Construire des chaines de traitement (workflow)
3- Interpréter les résultats graphiques
4- Faire une présentation PowerPoint de 5 minutes
5- Fournir une copie papier des transparents

Université Paris 13 – Institut Galilée
Département d'Informatique
CumUE
Université Sorbonne Paris Cité

LICENCE INFORMATIQUE

3ème année

VISUALISATION DES DONNEES

Data Visualization

Projet

Visualisation des données L3-INFO 2016/2017

Projet 4 : Données Communautés et crimes aux Etats-Unis

Sources :

Michael Redmond (Redmond ' à ' lasalle.edu) ; La Salle Université ; Philadelphie , PA , 19141 , USA-Recensement américain de 1990 , 1995 US FBI Uniform Crime Report , 1990, la loi des Etats-Unis Gestion exécution et enquête de Statistique administrative, disponible à partir de ICPSR du Michigan .

Description :

Les données combinent les données socio-économiques du recensement américain en 1990, les données de l'application de la loi de l'enquête LEMAS de 1990 aux États-Unis , et les données des crimes de 1995 du FBI DUC.
Ensemble de données : multivariées
Caractéristiques des attributs : réels
Tâches associées : Régression
Nombre d'instances : 1994
Nombre d'attributs : 128
Date : 2009-07-13

De nombreuses variables sont incluses afin que les algorithmes d'apprentissage qui sélectionnent ou apprennent des poids pour ces attributs peuvent être testés. Toutefois, les attributs n'ayant manifestement aucun rapport avec le problème cible n'ont pas été inclus ; les attributs ont été choisis s'il y avait un lien plausible avec le crime (N = 122), plus l'attribut à prédire (par habitant infractions violentes).
Les variables incluses dans l'ensemble des données impliquant la communauté, tels que le pourcentage de la population urbaine considérée, et le revenu médian de la famille, et impliquant l'application de la loi , comme par habitant le nombre d'officiers de police , et pourcentage des agents affectés à des unités de drogue.
La variable de crimes violents par habitant a été calculée en utilisant la population et la somme de variables de la criminalité considérés comme des crimes violents aux États-Unis : l'assassinat, le viol, le vol qualifié, et l'agression. Il y avait apparemment une certaine controverse dans certains Etats concernant le nombre de viols. Elles ont abouti à des valeurs manquantes pour viol, qui a abouti à des valeurs incorrectes pour les crimes violents par habitant. Ces villes ne sont pas incluses dans l'ensemble des données. Beaucoup de ces communautés ont été omises du Midwest américain.

Les données sont décrites ci-dessous sur la base de valeurs d'origine. Toutes les données numériques ont été normalisées entre 0.00 et 1.00 en utilisant une méthode de discrétisation. Une limitation est que l'enquête LEMAS était des services de police avec au moins 100 agents. Pour nos besoins, les communautés pas trouvées dans les deux ensembles de données de recensement et de la criminalité ont été omises. De nombreuses communautés sont donc absentes.

Travail demandé :

Les bases de données sont disponibles à : http://archive.ics.uci.edu/ml/datasets.html

Ce qui est demandé dans ce travail de projet :

1- Effectuer une exploration des données par l'outil KNIME
2- Construire des chaines de traitement (workflow)
3- Interpréter les résultats graphiques
4- Faire une présentation PowerPoint de 5 minutes
5- Fournir une copie papier des transparents

Université Paris 13 – Institut Galilée
Département d'Informatique
CumUE
Université Sorbonne Paris Cité

LICENCE INFORMATIQUE

3ème année

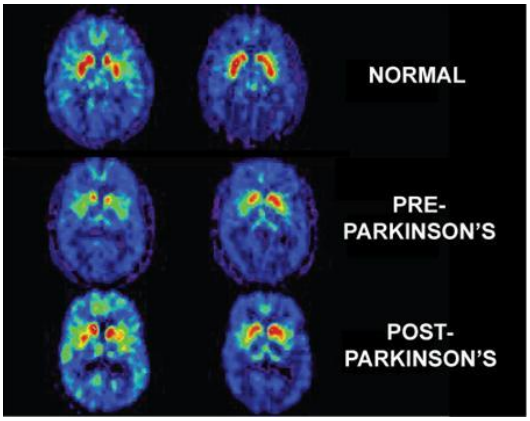
VISUALISATION DES DONNEES

Data Visualization

Projet

Visualisation des données L3-INFO 2016/2017

Projet 5 : Données maladie de Parkinson



Sources :

Le jeu de données a été créé par Max Little, de l'Université d'Oxford, dans le cadre d'une collaboration avec le Centre national pour la voix et la parole, Denver, Colorado, qui a enregistré les signaux de parole. L'étude originale publie les méthodes d'extraction de caractéristiques pour les troubles généraux de la voix.

Description :

Ensemble de données : multivariées
Nombre d'instances : 195
Caractéristiques des attributs : réels
Nombre d'attributs : 23
Date de Don : 2008-06-26
Tâches associées : Classification

Cette base de données est composée d'une série de mesures biomédicales de la voix à partir de 31 personnes, 23 ayant la maladie de Parkinson (MP). Chaque colonne de la table est une mesure particulière de la voix, et chaque ligne correspond à une des 195 enregistrements de ces personnes (colonne "nom"). L'objectif principal des données est de discriminer les personnes en bonne santé de ceux avec MP, selon "l'état" colonne qui est fixé à 0 pour la santé et une pour MP.
Les données sont au format CSV ASCII. Les lignes du fichier CSV contiennent une instance correspondant à un mode d'enregistrement de la voix. Il y a environ 6 enregistrements par patient, le nom du patient est identifié dans la première colonne.

Les attributs :
nom - ASCII nom de l'objet et le numéro d'enregistrement
MDVP : Fo (Hz) - moyenne fréquence fondamentale vocal
MDVP FHI (Hz) - fréquence fondamentale vocal maximum
MDVP : Flo (Hz) - fréquence fondamentale vocal minimum
MDVP : Jitter (%) , entamés : Jitter (Abs) , entamés : RAP , entamés : PPQ , Jitter: DDP-Plusieurs mesures de variation de la fréquence fondamentale MDVP : Shimmer , entamés : Shimmer (dB) , Shimmer : APQ3 , Shimmer : APQ5 , entamés : APQ , Shimmer : DDA - Plusieurs mesures de variation d'amplitude NHR , HNR - Deux mesures de rapport de bruit à composantes tonales dans la voix état - L'état de santé du sujet (un) - la maladie de Parkinson , (zéro) – santé RPDE , D2 - Deux mesures de complexité dynamiques non linéaires DFAE - Signal fractale échelle exposant Spread1 , spread2 , PPE - Trois mesures non linéaires de variation de fréquence fondamentale.

Référence :
'Exploiting Nonlinear Recurrence and Fractal Scaling Properties for Voice Disorder Detection', Little MA, McSharry PE, Roberts SJ, Costello DAE, Moroz IM. BioMedical Engineering OnLine 2007, 6:23 (26 June 2007)

Travail demandé :

Les bases de données sont disponibles à : http://archive.ics.uci.edu/ml/datasets.html

Ce qui est demandé dans ce travail de projet :

1- Effectuer une exploration des données par l'outil KNIME
2- Construire des chaines de traitement (workflow)
3- Interpréter les résultats graphiques
4- Faire une présentation PowerPoint de 5 minutes
5- Fournir une copie papier des transparents

Université Paris 13 – Institut Galilée
Département d'Informatique
CumUE
Université Sorbonne Paris Cité

LICENCE INFORMATIQUE

3ème année

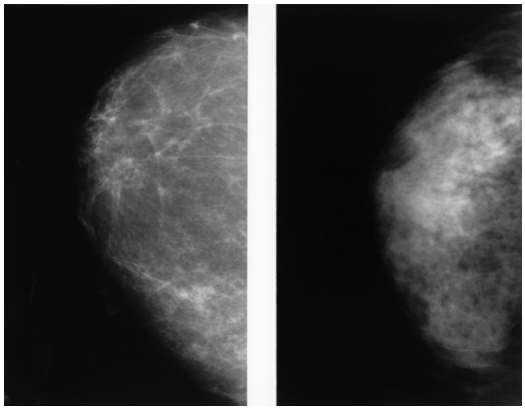
VISUALISATION DES DONNEES

Data Visualization

Projet

Visualisation des données L3-INFO 2016/2017

Projet 6 : Données Diagnostic de cancer Wisconsin (WDBC)



Sources :

Dr. William H. Wolberg, General Surgery Dept., University of Wisconsin, Clinical Sciences Center, Madison, WI 53792, wolberg@eagle.surgery.wisc.eduW. Nick Street, Computer Sciences Dept., University of Wisconsin, 1210 West Dayton St., Madison, WI 53706, street@cs.wisc.edu 608-262-6619Olvi L. Mangasarian, Computer Sciences Dept., University of Wisconsin, 1210 West Dayton St., Madison, WI 53706, olvi@cs.wisc.edu
Date: Novembre 1995

Description :

Caractéristiques sont calculées à partir d'une image numérisée d'une aiguille fine d'aspiration (FNA) d'une masse au sein. Elles décrivent les caractéristiques des noyaux de cellules présentes dans l'image .

- Champ de prédiction 2 , diagnostic: B = bénigne , M = maligne
- Ensembles sont linéairement séparables en utilisant les 30 caractéristiques d'entrée
- Une meilleure précision prédictive obtenue en utilisant un plan de séparation dans l'espace 3D. Précision estimée à 97,5 % en utilisant la Crossvalidation 10 fois.

Les résultats décrits ci-dessus ont été obtenus en utilisant la Multisurface Method-Tree (MSM-T) [K. P. Bennett, "Decision Tree Construction Via Linear Programming." Proceedings of the 4th Midwest Artificial Intelligence and Cognitive Science Society, pp. 97-101, 1992], une méthode de classification qui utilise la programmation linéaire pour la construction d'un arbre de décision. Les caractéristiques pertinentes ont été sélectionnées en utilisant une recherche exhaustive dans l'espace de 1-4 caractéristiques et 1-3 plans de séparation.

Le programme linéaire utilisé pour obtenir le plan de séparation dans l'espace à 3 dimensions est celui décrit dans :
[K. P. Bennett and O. L. Mangasarian: "Robust Linear Programming Discrimination of Two Linearly Inseparable Sets", Optimization Methods and Software 1, 1992, 23-34].

Nombre de cas : 569
Nombre d'attributs : 32 (ID, le diagnostic , 30 fonctions à valeurs réelles)

Information sur les attributs :
1) le numéro d'identification
2) Diagnostic (M = maligne , B = bénigne)
3-32) Dix fonctions à valeurs réelles sont calculées pour chaque noyau de la cellule :
    a) rayon (moyenne des distances de centre aux points sur le périmètre)

b) la texture (écart-type des valeurs de gris)
c) périmètre
d) zone
e) la douceur (variation locale de longueurs de rayon)
f) la compacité (périmètre ^ 2 / zone - 1.0)
g) concavité (gravité des portions concaves du contour) des points (nombre de parties concaves du contour)
h) concaves
i) la symétrie
j) la dimension fractale (« approximation du littoral " - 1)

Plusieurs des documents mentionnés ci-dessus contiennent des descriptions détaillées sur comment ces caractéristiques sont calculées.

Distribution de classe : 357 bénigne, maligne 212

Travail demandé :

Les bases de données sont disponibles à : http://archive.ics.uci.edu/ml/datasets.html

Ce qui est demandé dans ce travail de projet :

1- Effectuer une exploration des données par l'outil KNIME
2- Construire des chaines de traitement (workflow)
3- Interpréter les résultats graphiques
4- Faire une présentation PowerPoint de 5 minutes
5- Fournir une copie papier des transparents

Université Paris 13 – Institut Galilée
Département d'Informatique
CumUE
Université Sorbonne Paris Cité

LICENCE INFORMATIQUE

3ème année

VISUALISATION DES DONNEES

Data Visualization

Projet

Visualisation des données L3-INFO 2016/2017

Projet 7 : Données qualité du vin

Sources :

Paulo Cortez, University of Minho, Guimarães, Portugal, http://www3.dsi.uminho.pt/pcortez
A. Cerdeira, F. Almeida, T. Matos and J. Reis, Viticulture Commission of the Vinho Verde Region(CVRVV), Porto, Portugal @2009

Description:

Deux ensembles de données sont inclus, liés à des échantillons de vin « Vinho Verde » rouges et blancs, en provenance du nord du Portugal. L'objectif est de modéliser la qualité du vin basée sur des tests physico-chimiques.
Dans la référence ci-dessus, deux ensembles de données ont été créés, en utilisant des échantillons de vin rouge et blanc.
Les entrées comprennent des tests objectifs (par exemple, les valeurs de pH) et la sortie est basée sur des données sensorielles (médiane d'au moins 3 évaluations faites par des experts en vin). Chaque expert classé la qualité du vin entre 0 (très mauvais) et 10 (excellent). Plusieurs méthodes d'extraction des données ont été appliquées à un modèle ces ensembles de données selon une approche de régression. Le vecteur modèle de machine de soutien atteint les meilleurs résultats. Plusieurs mesures ont été calculées : MAD, matrice de confusion pour une tolérance d'erreur fixe (T), etc En outre, nous traçons les importances relatives des variables d'entrée (telle que mesurée par une sensibilité de la Procédure d'analyse).

Les deux ensembles de données sont liés à des variantes rouges et blancs des vins Portugais " Vinho Verde ". Pour plus de détails, consulter : http://www.vinhoverde.pt/en/ ou la référence [Cortez et al, 2009].
En raison de problèmes de confidentialité et de logistique, que des variables physico-chimique (entrées) et sensorielle (la sortie) sont disponibles (par exemple, il n'existe pas de données sur les types de raisin, vin de marque , le prix de vente de vin , etc.).
Ces données peuvent être considérées comme des tâches de classification ou de régression. Les classes sont ordonnées et non équilibrés (par exemple, il ya des vins plus normaux que des excellents ou pauvres). Des algorithmes de détection des valeurs aberrantes peuvent être utilisés pour détecter les vins rares d'excellente ou pauvre qualité. En outre, nous ne sommes pas sûr si toutes les variables d'entrée sont pertinentes.

Nombre d'instances :
- vin rouge - 1599 ;
- vin blanc - 4898.

Nombre d'attributs : 11 + attribut de sortie

Note: plusieurs des attributs peuvent être corrélés, donc il est logique d'appliquer une sorte de sélection de caractéristiques.

Les variables d'entrée (sur la base des tests physico-chimiques) :
1 - acidité fixe
2 - acidité volatile
3 - l'acide citrique
4 - sucre résiduel
5 - les chlorures
6 - dioxyde de soufre libre
7 - dioxyde de soufre total
8 – densité
9 – pH
10 - les sulfates
11 – alcool

Grandeur de sortie (sur la base de données sensorielles) :
12 - qualité (note comprise entre 0 et 10)

Travail demandé :

Les bases de données sont disponibles à : http://archive.ics.uci.edu/ml/datasets.html

Ce qui est demandé dans ce travail de projet :

1- Effectuer une exploration des données par l'outil KNIME
2- Construire des chaines de traitement (workflow)
3- Interpréter les résultats graphiques
4- Faire une présentation PowerPoint de 5 minutes
5- Fournir une copie papier des transparents

Université Paris 13 – Institut Galilée
Département d'Informatique
CumUE
Université Sorbonne Paris Cité

LICENCE INFORMATIQUE

3ème année

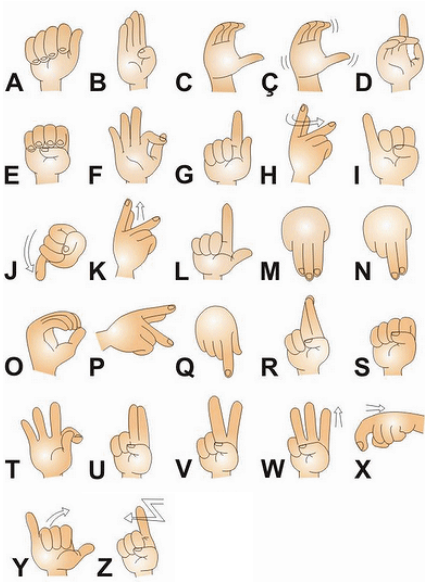
VISUALISATION DES DONNEES

Data Visualization

Projet

Visualisation des données L3-INFO 2016/2017

Projet 8 : Données Mouvement LIBRAS



Sources :

Daniel Baptista Dias (Dias , D.B.)
Sarajane Marques Peres (Peres , S. M.)
Helton Hideraldo Bscaro (Bscaro. H. H.)
{ danielbdias , heltonhb , Sarajane } @ usp.br

Université de São Paulo
School of Art , sciences et sciences humaines
São Paulo , SP , Brésil
http://each.uspnet.usp.br/each/
Date: Novembre 2008

Description :

LIBRAS, acronyme du nom portugais "Lngua Brasileira de Sinais", est l'officielle langue brésilienne des signes.

L'ensemble de données contient 15 classes de 24 cas chacun, où chacun des références de classe à une main type de mouvement dans LIBRAS. Le mouvement de la main est représenté par une courbe bidimensionnelle effectuée par la main dans une période de temps. Les courbes ont été obtenues à partir des vidéos de mouvements de la main, avec 4 différentes personnes, au cours de deux sessions. Chaque vidéo correspond à un seul mouvement de la main et a environ 7 secondes.
Chaque vidéo correspond à une fonction F dans un espace de fonctions qui est la version continue de l'ensemble de données d'entrée.
Dans le pré-traitement de la vidéo, une normalisation de temps est effectuée en sélectionnant à partir de chaque 45 trames vidéo, en fonction à une distribution uniforme. Dans chaque image, les pixels barycentre des objets segmentés (la main) sont trouvés pour composer la version discrète de la courbe F avec 45 points. Toutes les courbes sont normalisées dans l'espace unitaire.
Afin de préparer ces mouvements à analyser par des algorithmes, nous avons effectué une opération de mise en correspondance, chaque courbe F est mappée dans une représentation à 90 éléments, avec une représentation des coordonnées du mouvement.

Chaque instance représente 45 points dans un espace à deux dimensions, qui peut être tracé d'une manière ordonnée (entre 1 et 45) afin d'en tirer le chemin du mouvement.

Nombre d'instances : 360 (24 dans chacun des quinze catégories)
Nombre d'attributs : 90 numérique (double) et 1 pour la classe (entier)
Distribution des classes : 6,66% pour chacune des 15 classes.

Travail demandé :

Les bases de données sont disponibles à : http://archive.ics.uci.edu/ml/datasets.html

Ce qui est demandé dans ce travail de projet :

1- Effectuer une exploration des données par l'outil KNIME
2- Construire des chaines de traitement (workflow)
3- Interpréter les résultats graphiques
4- Faire une présentation PowerPoint de 5 minutes
5- Fournir une copie papier des transparents

Université Paris 13 – Institut Galilée
Département d'Informatique
CumUE
Université Sorbonne Paris Cité

LICENCE INFORMATIQUE

3ème année

VISUALISATION DES DONNEES

Data Visualization

Projet

Visualisation des données L3-INFO 2016/2017

Projet 9 : Données d'identification de verre

Sources :

B. allemand
Centre de recherches pour Home Office Forensic Science Service
Aldermaston , Reading, Berkshire RG7 4PN
Vina Spiehler , Ph.D. , DABFT
Diagnostic Products Corporation
(213) 776-0180 (poste 3014)
Date: Septembre 1987

Description :

L'étude de la classification des types de verre a été motivée par les enquêtes criminologiques. Sur la scène du crime, le verre peut être utilisé comme preuve ... si il est correctement identifié !

Nombre d'instances : 214
Nombre d'attributs : 10 (y compris un Id #) plus l'attribut de classe

Les attributs :
1. Numéro d'identification: 1-214
2. RI : indice de réfraction
3. Na : Sodium (unité de mesure : en pourcentage d'oxyde correspondant, attributs 4-10)
4. Mg : Magnésium
5. Al : Aluminium
6. Si: Silicon
7. K : Potassium
8. Ca : Calcium
9. Ba : baryum
10. Fe : Fer
11. Type de verre : (attribut de classe)
    - 1 building_windows_float_processed
    - 2 building_windows_non_float_processed
    - 3 vehicle_windows_float_processed
    - 4 vehicle_windows_non_float_processed (aucun dans cette base de données)
    - 5 conteneurs
    - 6 vaisselle
    - 7 projecteurs

Distribution des classes : (sur 214 cas au total)
- 163 verre de la fenêtre (la construction de fenêtres et vitres de véhicules)
- 87 flotteur traités
- 70 fenêtres du bâtiment

- 17 fenêtres de véhicules
- 76 non traitées flotteur
- 76 fenêtres du bâtiment
- 0 vitres de véhicules
- 51 non – fenêtre
- 13 conteneurs
- 9 arts de la table
- 29 projecteurs

Travail demandé :

Les bases de données sont disponibles à : http://archive.ics.uci.edu/ml/datasets.html

Ce qui est demandé dans ce travail de projet :

1- Effectuer une exploration des données par l'outil KNIME
2- Construire des chaines de traitement (workflow)
3- Interpréter les résultats graphiques
4- Faire une présentation PowerPoint de 5 minutes
5- Fournir une copie papier des transparents

Université Paris 13 – Institut Galilée
Département d'Informatique
CumUE
Université Sorbonne Paris Cité

LICENCE INFORMATIQUE

3ème année

VISUALISATION DES DONNEES

Data Visualization

Projet

Visualisation des données L3-INFO 2016/2017

Projet 10 : Données Vertebral Column Data Set

Sources :

Guilherme de Alencar Barreto (guilherme '@' deti.ufc.br) et Ajalmar Rêve vont da Rocha Neto (ajalmar '@' ifce.edu.br) , ministère de la téléinformatique génie , Université fédérale de Ceará , Fortaleza , Ceará , Brésil.
Henrique Fonseca Antonio da Mota Filho (hdamota '@' gmail.com) , Hôpital Monte Klinikum , Fortaleza , Ceará , Brésil .

Description :

Ensemble de données contenant des valeurs pour les six caractéristiques biomécaniques utilisées pour classer les patients en orthopédie en 3 classes (normale, hernie du disque ou spondilolysthesis) ou 2 classes (normaux ou anormaux).
Cet ensemble de données biomédicales a été construit par le Dr Henrique da Mota au cours d'une période de résidence médicale au sein du Groupe de recherche appliquée en orthopédie (GARO du Centre Médico - Chirurgical de Réadaptation, Lyon, France. Les données ont été organisées dans deux tâches différentes mais connexes. La première tâche consiste à classer les patients comme appartenant à une des trois catégories : Normal (100 patients) , hernie discale (60 patients) ou du spondylolisthésis (150 patients).

Pour la deuxième tâche, les catégories hernie discale et spondylolisthésis ont été fusionnées en une seule catégorie étiquetée comme «anormal». Ainsi, la deuxième tâche consiste à classer les patients comme appartenant à une des deux catégories : normale (100 patients) ou anormaux (210 patients).

Chaque patient est représenté dans l'ensemble par six attributs biomécaniques dérivés de la forme et l'orientation du bassin et de la colonne lombaire (dans cet ordre) des données : incidence pelvienne, inclinaison du bassin, l'angle de lordose lombaire, pente sacrée, rayon pelvien et qualité de spondylolisthésis. La convention suivante est utilisée pour les étiquettes de classe : DH (hernie discale), spondylolisthésis (SL), Normal (NO) et anormal (AB).

Documents pertinents :
(1) Berthonnaud, E., Dimnet, J., Roussouly, P. & Labelle, H. (2005). 'Analysis of the sagittal balance of the spine and pelvis using shape and orientation parameters', Journal of Spinal Disorders & Techniques, 18(1):40-47.

(2) Rocha Neto, A. R. & Barreto, G. A. (2009). 'On the Application of Ensembles of Classifiers to the Diagnosis of Pathologies of the Vertebral Column: A Comparative Analysis', IEEE Latin America Transactions, 7(4):487-496.

(3) Rocha Neto, A. R., Sousa, R., Barreto, G. A. & Cardoso, J. S. (2011). 'Diagnostic of Pathology on the Vertebral Column with Embedded Reject Optionâ€, Proceedings of the 5th Iberian Conference on Pattern

Recognition and Image Analysis (IbPRIA'2011), Gran Canaria, Spain, Lecture Notes on Computer Science, vol. 6669, p. 588-595.

Travail demandé :

Les bases de données sont disponibles à : http://archive.ics.uci.edu/ml/datasets.html

Ce qui est demandé dans ce travail de projet :

1- Effectuer une exploration des données par l'outil KNIME
2- Construire des chaines de traitement (workflow)
3- Interpréter les résultats graphiques
4- Faire une présentation PowerPoint de 5 minutes
5- Fournir une copie papier des transparents

Université Paris 13 – Institut Galilée
Département d'Informatique
CumUE
Université Sorbonne Paris Cité

LICENCE INFORMATIQUE

3ème année

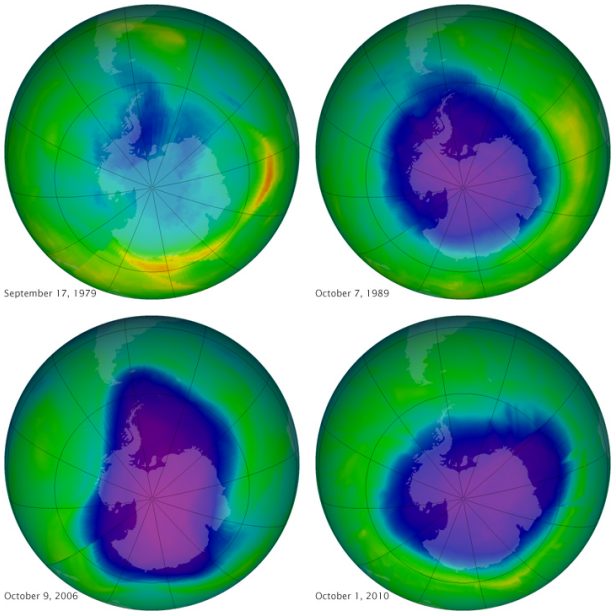
VISUALISATION DES DONNEES

Data Visualization

Projet

Visualisation des données L3-INFO 2016/2017

Projet 11 : Données Ozone Level Detection Data Set



Sources :

Kun Zhang, zhang.kun05 **'@'** gmail.com, Department of Computer Science, Xavier University of Lousiana
Wei Fan, wei.fan **'@'** gmail.com, IBM T.J.Watson Research
XiaoJing Yuan, xyuan **'@'** uh.edu, Engineering Technology Department, College of Technology, University of Houston

Description :

Deux ensembles de données d'ozone sont inclus dans cette collection. L'un est l'ensemble des données de huit heures (eighthr.data), l'autre est un ensemble de données « pic » d'une heure (onehr.data). Ces données ont été recueillies de 1998 à 2004 à Houston, Galveston et zone Brazoria.

Tout attribut commençant par T, représente la température mesurée à différents moments tout au long de la journée, et les mises en chantier avec WS indiquent la vitesse du vent à différents temps.

WSR_PK : continu . coup d'oeil la vitesse du vent - résultante (ce qui signifie en moyenne de vecteur vent)
WSR_AV : continu . vitesse moyenne du vent
T_PK : continu . pic T
T_AV : continu . T moyenne
T85 : continu . T au niveau 850 hpa (ou environ 1500 m de hauteur)
RH85 : continu . Humidité relative à 850 hPa
U85 : continu . (U vent - est-ouest direction du vent à 850 hPa)
V85 : continu . V vent - N- S la direction du vent à 850
HT85 : continu . Hauteur du géopotentiel à 850 hpa, il s'agit de la même que la hauteur à basse altitude
T70 : continu . T à 700 hpa niveau (environ 3100 m de hauteur)
RH70 : continu .
U70 : continu .
V70 : continu .
HT70 : continu .
T50 : continu . T au niveau 500 hpa (à peu près à 5500 m de hauteur)
RH50 : continu .

U50 : continu .
V50 : continu .
HT50 : continu .
KI : continu . K – Index
TT : continue . T- totaux
SLP : continu . Pression au niveau de la mer
SLP_ : continu . SLP changement de veille
Precp : continu . – précipitations

Voici les spécifications de plusieurs attributs les plus importants qui sont très appréciés par Texas Commission on Environmental Quality (TCEQ). Plus de détails peuvent être trouvés dans les documents de référence.

- O 3 - Local prévision pics d'ozone
- Niveau Upwind ozone de fond
- Facteur émissions de précurseurs liés - - EmFactor
- Tmax - Température maximale en degrés F
- Tb - température de base où commence la production d'ozone net (50 F)
- SRD - total de rayonnement solaire pour la journée
- WSA - vitesse du vent près de lever du soleil (en utilisant 09-12 mode prévisions UTC)
- WSp - Vitesse du vent en milieu de journée (15-21 en utilisant le mode de prévisions UTC)

Nombre d'instances : 2536
Nombre d'attributs : 73
1,0 | deux classes : une Journée de l'ozone , 0 : jour normale

Références :
Forecasting skewed biased stochastic ozone days: analyses, solutions and beyond, Knowledge and Information Systems, Vol. 14, No. 3, 2008.
Discusses details about the dataset, its use as well as various experiments (both cross-validation and streaming) using many state-of-the-art methods.
A shorter version of the paper (does not contain some detailed experiments as the journal paper above) is in: Forecasting Skewed Biased Stochastic Ozone Days: Analyses and Solutions. ICDM 2006: 753-764

Travail demandé :

Les bases de données sont disponibles à : http://archive.ics.uci.edu/ml/datasets.html

Ce qui est demandé dans ce travail de projet :

1- Effectuer une exploration des données par l'outil KNIME
2- Construire des chaines de traitement (workflow)
3- Interpréter les résultats graphiques
4- Faire une présentation PowerPoint de 5 minutes
5- Fournir une copie papier des transparents

Université Paris 13 – Institut Galilée
Département d'Informatique
CumUE
Université Sorbonne Paris Cité

UNIVERSITÉ PARIS 13 NORD

LICENCE INFORMATIQUE

3ème année

VISUALISATION DES DONNEES

Data Visualization

Projet

Visualisation des données L3-INFO 2016/2017

Projet 12 : Données Water Treatment Plant Data Set

Sources :

Manel Poch (igte2 **'@'** cc.uab.es)
Unitat d'Enginyeria Quimica
Universitat Autonoma de Barcelona. Bellaterra. Barcelona; Spain

Javier Bejar and Ulises Cortes (bejar **'@'** lsi.upc.es)
Dept. Llenguatges i Sistemes Informatics;
Universitat Politecnica de Catalunya. Barcelona; Spain

Description :

Cet ensemble de données provient des mesures quotidiennes de capteurs dans une usine de traitement des eaux urbaines résiduaires. L'objectif est de classer l'état de fonctionnement de l'usine afin de prédire les défauts à travers les variables d'état de la plante à chacune des étapes du processus de traitement. Ce domaine a été déclaré comme un domaine mal structuré.

Les attributs sont numériques et se présentent dans la liste ci-dessous :

1 Q-E (input flow to plant)
2 ZN-E (input Zinc to plant)
3 PH-E (input pH to plant)
4 DBO-E (input Biological demand of oxygen to plant)
5 DQO-E (input chemical demand of oxygen to plant)
6 SS-E (input suspended solids to plant)
7 SSV-E (input volatile supended solids to plant)
8 SED-E (input sediments to plant)
9 COND-E (input conductivity to plant)
10 PH-P (input pH to primary settler)
11 DBO-P (input Biological demand of oxygen to primary settler)
12 SS-P (input suspended solids to primary settler)
13 SSV-P (input volatile supended solids to primary settler)
14 SED-P (input sediments to primary settler)

15 COND-P (input conductivity to primary settler)
16 PH-D (input pH to secondary settler)
17 DBO-D (input Biological demand of oxygen to secondary settler)
18 DQO-D (input chemical demand of oxygen to secondary settler)
19 SS-D (input suspended solids to secondary settler)
20 SSV-D (input volatile supended solids to secondary settler)
21 SED-D (input sediments to secondary settler)
22 COND-D (input conductivity to secondary settler)
23 PH-S (output pH)
24 DBO-S (output Biological demand of oxygen)
25 DQO-S (output chemical demand of oxygen)
26 SS-S (output suspended solids)
27 SSV-S (output volatile supended solids)
28 SED-S (output sediments)
29 COND-S (output conductivity)
30 RD-DBO-P (performance input Biological demand of oxygen in primary settler)
31 RD-SS-P (performance input suspended solids to primary settler)
32 RD-SED-P (performance input sediments to primary settler)
33 RD-DBO-S (performance input Biological demand of oxygen to secondary settler)
34 RD-DQO-S (performance input chemical demand of oxygen to secondary settler)
35 RD-DBO-G (global performance input Biological demand of oxygen)
36 RD-DQO-G (global performance input chemical demand of oxygen)
37 RD-SS-G (global performance input suspended solids)
38 RD-SED-G (global performance input sediments)

Université Paris 13 – Institut Galilée
Département d'Informatique
CumUE
Université Sorbonne Paris Cité

UNIVERSITÉ PARIS 13 NORD

LICENCE INFORMATIQUE

3ème année

VISUALISATION DES DONNEES

Data Visualization

Projet

Visualisation des données L3-INFO 2016/2017

Projet 13 : Données Wholesale customers Data Set

Sources :

Margarida G. M. S. Cardoso, margarida.cardoso **'@'** iscte.pt, ISCTE-IUL, Lisbon, Portugal

Description :

L'ensemble de données se réfère aux clients d'un distributeur en gros. Il comprend les dépenses annuelles en unités monétaires (mu) sur les catégories de produits divers.

Les attributs sont numériques et se présentent dans la liste suivante :

1) FRESH: annual spending (m.u.) on fresh products (Continuous);
2) MILK: annual spending (m.u.) on milk products (Continuous);
3) GROCERY: annual spending (m.u.)on grocery products (Continuous);
4) FROZEN: annual spending (m.u.)on frozen products (Continuous)
5) DETERGENTS_PAPER: annual spending (m.u.) on detergents and paper products (Continuous)
6) DELICATESSEN: annual spending (m.u.)on and delicatessen products (Continuous);
7) CHANNEL: customersâ€™ Channel - Horeca (Hotel/Restaurant/CafÃ©) or Retail channel (Nominal)
8) REGION: customersâ€™ Region â€“ Lisnon, Oporto or Other (Nominal)
Descriptive Statistics:

(Minimum, Maximum, Mean, Std. Deviation)
FRESH (3, 112151, 12000.30, 12647.329)
MILK (55, 73498, 5796.27, 7380.377)
GROCERY (3, 92780, 7951.28, 9503.163)
FROZEN (25, 60869, 3071.93, 4854.673)
DETERGENTS_PAPER (3, 40827, 2881.49, 4767.854)
DELICATESSEN (3, 47943, 1524.87, 2820.106)

REGION Frequency
Lisbon 77
Oporto 47
Other Region 316
Total 440

CHANNEL Frequency
Horeca 298
Retail 142
Total 440

Références :

Cardoso, Margarida G.M.S. (2013). Logical discriminant models â€“ Chapter 8 in Quantitative Modeling in Marketing and Management Edited by Luiz Moutinho and Kun-Huang Huarng. World Scientific. p. 223-253. ISBN 978-9814407717

Jean-Patrick Baudry, Margarida Cardoso, Gilles Celeux, Maria JosÃ© Amorim, Ana Sousa Ferreira (2012). Enhancing the selection of a model-based clustering with external qualitative variables. RESEARCH REPORT NÂ° 8124, October 2012, Project-Team SELECT. INRIA Saclay - ÃŽle-de-France, Projet select, UniversitÃ© Paris-Sud 11

Abreu, N. (2011). Analise do perfil do cliente Recheio e desenvolvimento de um sistema promocional. Mestrado em Marketing, ISCTE-IUL, Lisbon

Université Paris 13 – Institut Galilée
Département d'Informatique
CumUE
Université Sorbonne Paris Cité

UNIVERSITÉ PARIS 13 NORD

LICENCE INFORMATIQUE

3ème année

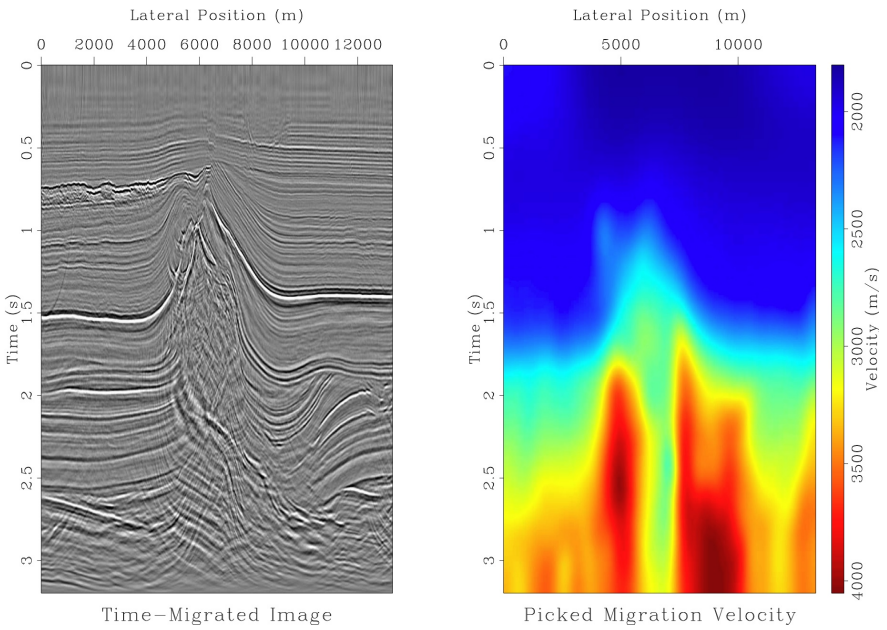
VISUALISATION DES DONNEES

Data Visualization

Projet

Visualisation des données L3-INFO 2016/2017

Projet 14 : Données Seismic-Bumps Data Set



Sources :

Marek Sikora^{1,2} (marek.sikora '@' polsl.pl), Lukasz Wrobel^{1} (lukasz.wrobel '@' polsl.pl)
(1) Institute of Computer Science, Silesian University of Technology, 44-100 Gliwice, Poland
(2) Institute of Innovative Technologies EMAG, 40-189 Katowice, Poland

Description :
L'activité minière a été et est toujours liée à l'apparition de dangers qui sont communément appelé les risques miniers. Un cas particulier de cette menace est un aléa sismique qui se produit fréquemment dans de nombreux mines souterraines. L'aléa sismique est la plus difficile détectable et prévisible des aléas naturels et en cet égard, il est comparable à un tremblement de terre. De plus en plus les systèmes de surveillance permettent de mieux comprendre les processus de la masse rocheuse et la définition de l'aléa sismique par des méthodes de prédiction. La précision des méthodes jusqu'ici créés est cependant loin d'être parfaite. La complexité du processus sismique et la grande disproportion entre le nombre d'événements sismiques de faible énergie et le nombre des phénomènes de haute énergie provoqués rendent les techniques statistiques insuffisantes pour prédire l'aléa sismique. Par conséquent, il est essentiel de rechercher de nouvelles possibilités d'une meilleure prévision des risques, également en utilisant des méthodes d'apprentissage automatique. Dans l'aléa sismique les techniques de clustering de données peuvent être appliquées.

Les attributs sont numériques et se présentent dans la liste ci-dessous :

1. seismic: result of shift seismic hazard assessment in the mine working obtained by the seismic method (a - lack of hazard, b - low hazard, c - high hazard, d - danger state);
2. seismoacoustic: result of shift seismic hazard assessment in the mine working obtained by the seismoacoustic method;
3. shift: information about type of a shift (W - coal-getting, N -preparation shift);
4. genergy: seismic energy recorded within previous shift by the most active geophone (GMax) out of geophones monitoring the longwall;
5. gpuls: a number of pulses recorded within previous shift by GMax;
6. gdenergy: a deviation of energy recorded within previous shift by GMax from average energy recorded during eight previous shifts;

7. gdpuls: a deviation of a number of pulses recorded within previous shift by GMax from average number of pulses recorded during eight previous shifts;
8. ghazard: result of shift seismic hazard assessment in the mine working obtained by the seismoacoustic method based on registration coming form GMax only;
9. nbumps: the number of seismic bumps recorded within previous shift;
10. nbumps2: the number of seismic bumps (in energy range [10^2,10^3)) registered within previous shift;
11. nbumps3: the number of seismic bumps (in energy range [10^3,10^4)) registered within previous shift;
12. nbumps4: the number of seismic bumps (in energy range [10^4,10^5)) registered within previous shift;
13. nbumps5: the number of seismic bumps (in energy range [10^5,10^6)) registered within the last shift;
14. nbumps6: the number of seismic bumps (in energy range [10^6,10^7)) registered within previous shift;
15. nbumps7: the number of seismic bumps (in energy range [10^7,10^8)) registered within previous shift;
16. nbumps89: the number of seismic bumps (in energy range [10^8,10^10)) registered within previous shift;
17. energy: total energy of seismic bumps registered within previous shift;
18. maxenergy: the maximum energy of the seismic bumps registered within previous shift;
19. class: the decision attribute - '1' means that high energy seismic bump occurred in the next shift ('hazardous state'), '0' means that no high energy seismic bumps occurred in the next shift ('non-hazardous state').