Objectif

Montrer qu'il est possible de charger directement un fichier au format WEKA (.ARFF) dans TANAGRA.

Le traitement des données manquantes est très sommaire dans ce module. Si l'on veut bénéficier d'options étendues, le logiciel DATANAMORF est conseillé.

Fichier

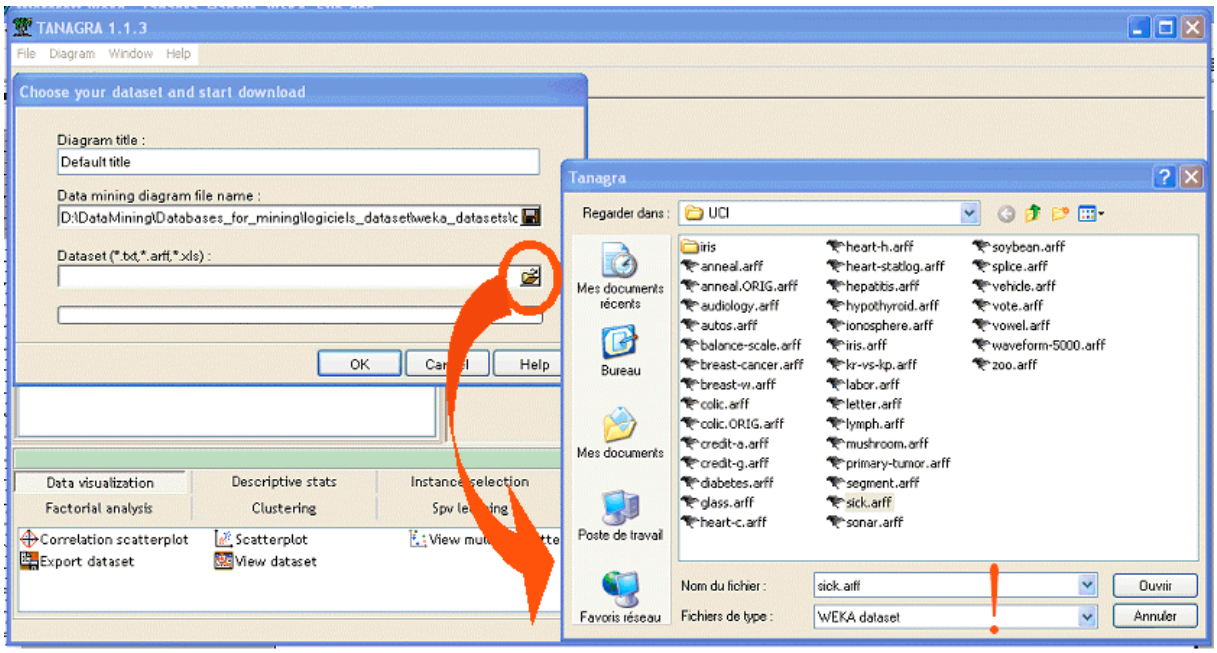
Nous utiliserons le fichier SICK.ARFF.

Manipuler un fichier ARFF

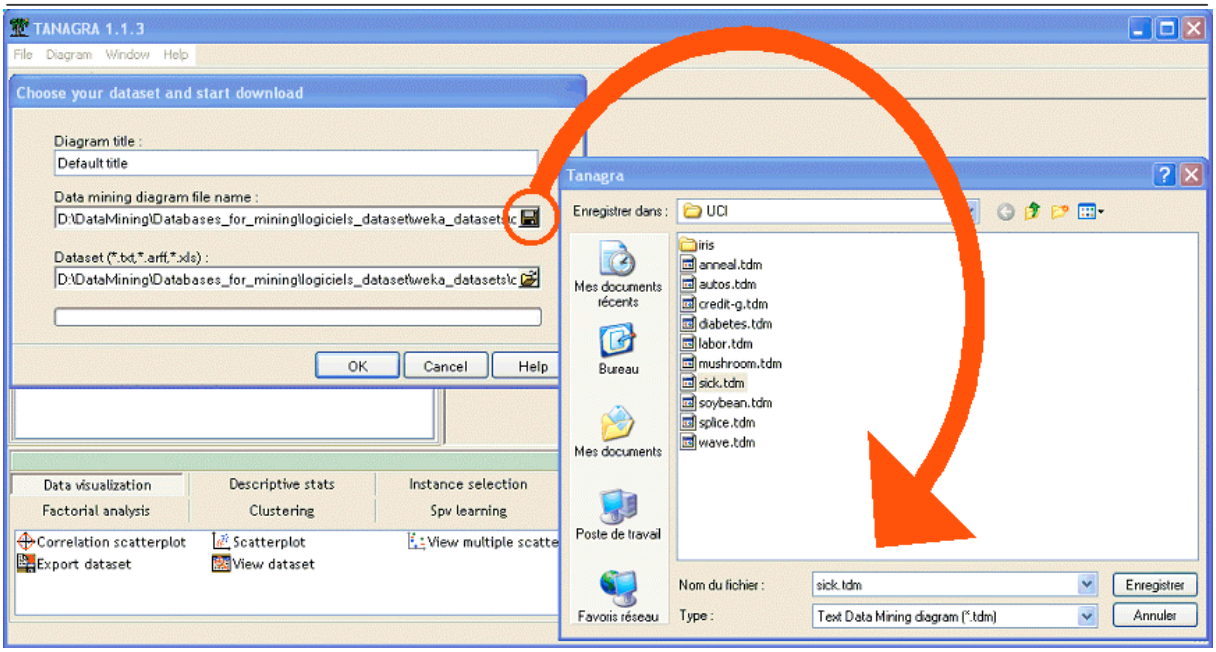
Importer le fichier .ARFF

Le première étape consiste à importer les données et créer dans la foulée le diagramme de traitements. La procédure est identique à tous les formats de fichiers traiter, il faut activer le menu « FILE / NEW ».

Dans la boîte de dialogue d'importation, sélectionner le fichier source, le format WEKA (.ARFF) est maintenant disponible.



Définir alors le nom du diagramme (utiliser SICK.TDM), puis valider en cliquant sur OK.

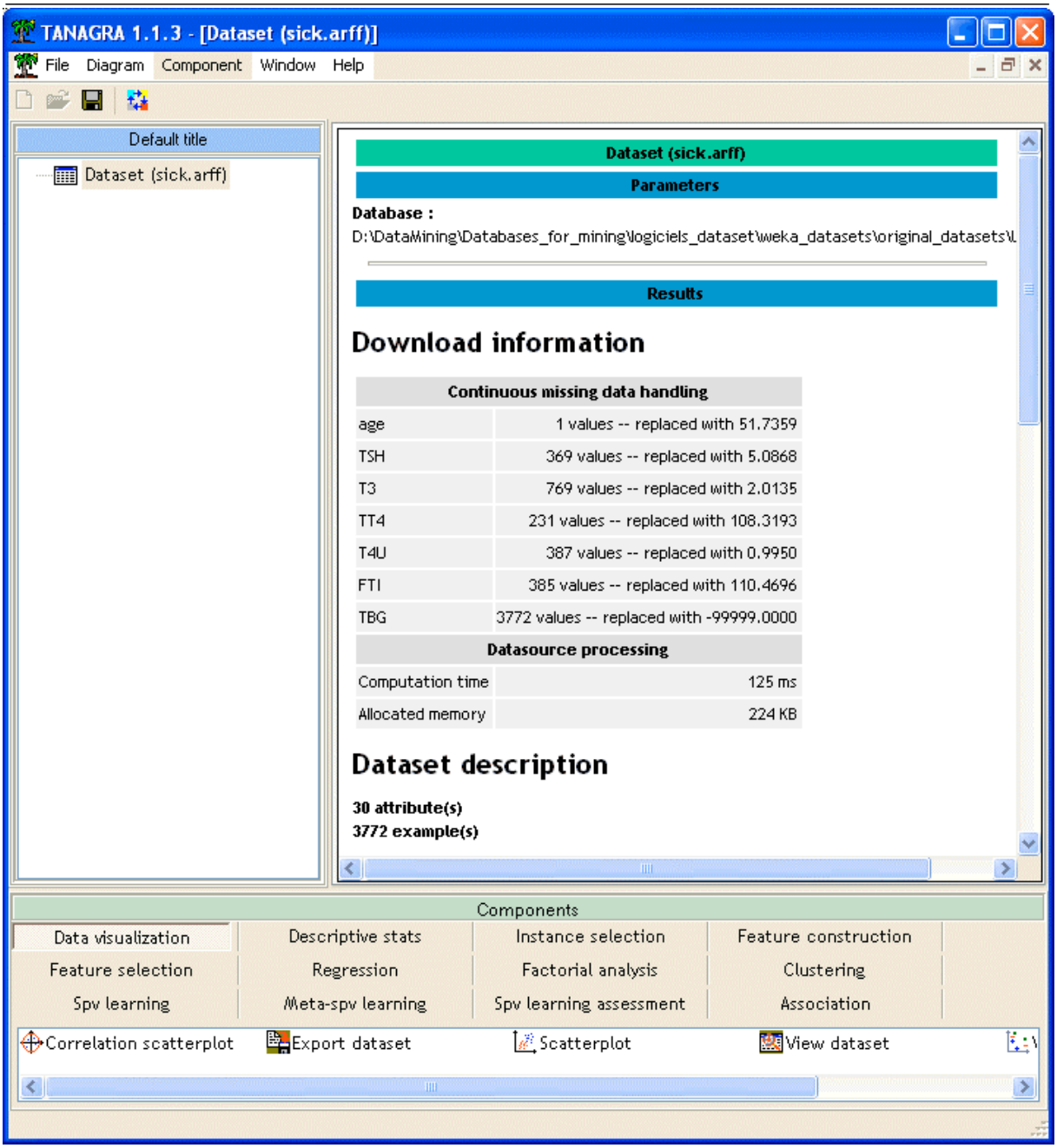


Rapport d'importation

Dans la fenêtre de description des données, de nouvelles informations vont apparaître, elles décrivent les pré-traitements réalisés sur le fichier ARFF. Ceci concerne essentiellement le traitement des données manquantes. Les opérations réalisées sont les suivantes :

- Pour un attribut discret, les valeurs manquantes deviennent une nouvelle modalité. Dans les fichiers ARFF, ces valeurs sont codés « ? » ou « ' ?' », cette modalité est donc automatiquement créée dans le dictionnaire des données de l'attribut. Comme il n'y a pas de calcul de valeur de remplacement à proprement parler, cette opération n'est pas signalée, le nombre de modalités de la variable est tout simplement augmentée d'une unité.

- Pour un attribut continu, les valeurs manquantes sont remplacées par la moyenne des observations disponibles. Dans ce cas, il y a bien un traitement, les modifications sont recensées avec 2 types d'informations : le nombre de données manquantes dans la colonne, la valeur de remplacement utilisée.



Il y a 30 attributs dans le fichier et 3772 observations. Parmi les attributs continus, 7 présentent des valeurs manquantes.

AGE par exemple possède 1 valeur manquante, elle a été remplacée par la moyenne calculée sur les individus correctement renseignés (51.7359).

Le cas de TBG est particulier. Il n'y a aucune donnée valide dans la colonne (3772 valeurs manquantes), dans ce cas, nous utilisons la valeur de remplacement par défaut (-99999).

Ouverture d'un diagramme de traitement

Si vous avez choisi le format TDM, seule la référence du fichier source a été inscrite dans le fichier de sauvegarde. L'avantage est de pouvoir appliquer le même diagramme de traitements sur des données éventuellement mis à jour.

Ainsi, à la prochaine ouverture du diagramme (SICK.TDM), le fichier source SICK.ARFF est chargé et les données manquantes sont de nouveau traités. Vous verrez mieux les différences si vous avez modifié le fichier de données entre temps.