IBM

Progetto FIRB LIBI

MTA

USER GUIDE

Dipartimento di Informatica,
Università degli Studi di Bari

Laboratorio di Acquisizione
della Conoscenza e
Apprendimento delle Macchine

Authors: C. Loglisci, M. Berardi, D. Malerba

Contratto di Ricerca:

“Problematiche di Knowledge Management e tecniche di Data Mining in ambito Bioinformatico”

INDEX

1. Introduction

This document is intended as a description of main functionalities and user interface of the MTA system. It aims to facilitate the MTA usage in biomedical literature mining tasks.

1.1. Generalities

MeSH Term Associator (MTA) is a data mining tool able to discover association rules on biomedical text corpora. It imports both some MeSH[1] (Medical Subject Headings) taxonomies and a set of abstracts published on Medline and discovers associations at different levels of abstraction (generalized association rules). Both automatic and semiautomatic approaches can be applied to structure the set of discovered rules and filter out uninteresting ones. In the automatic approach rules are filtered out without using user knowledge, while in the semiautomatic approach user domain knowledge is exploited to strongly guide the exploration of the set of discovered rules. Discovered association ruels can be imported/exported in PMML. Similarities between discovered association rules can be visually explored through a multidimensional analysis technique. MTA integrates the IBM Web Services for Life Sciences[2] which allow to directly query the PubMed remote database (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi) and to obtain a set of abstracts of interest for the task at hand. It includes two data mining algorithms for generalized association rules.

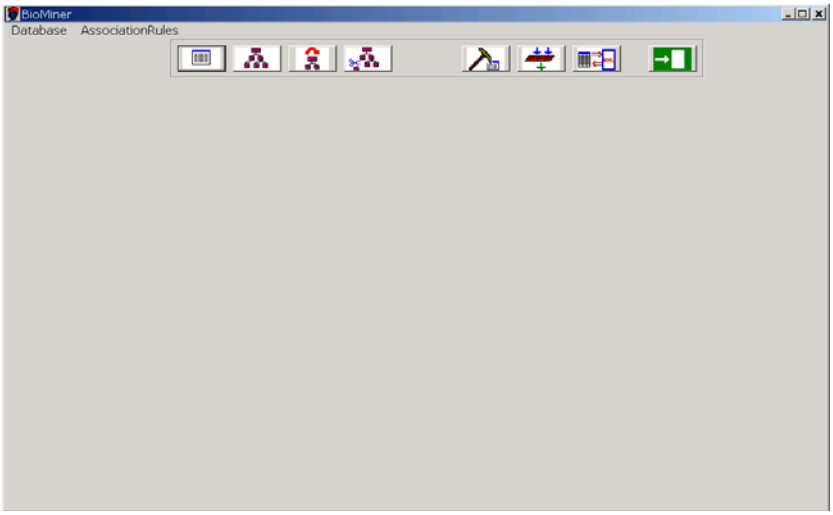


Fig. 1 MTA initial window

1.2. System requirements

- Microsoft® Windows® 2000
- 256 MB of available RAM is needed
- 640 MB of space on hard disk. User has to take into account that the space on hard disk depends on input data size and the size of discovered association rules as well.

[1] http://www.nlm.nih.gov/mesh/

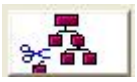
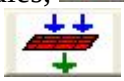
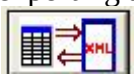
[2] Weil, N. *BioIT World*, URL: http://www.bio-itworld.com/news/042203_report2381.html

2. Installing MTA

MTA requires a DataBase connection to store input text corpora, to perform mining processes and to access discovered knowledge for filtering and visualization steps.
MTA interfaces Microsoft Access (97 or 2000) DataBase Management Systems (DBMS). It requires two different databases, one for input data and mining results and another to manage taxonomical knowledge used during the mining process.

3. Using MTA

MTA supports the following tasks:

- access to input data ;
- creating new input data source ; 
- selecting taxonomical knowledge; 
- loading and restoring taxonomical knowledge; 
- mining generalized association rules; 
- filtering mined association rules; 
- importing/exporting association rules from/to a representation model in PMML standard language. 

4. Access input data

This version of MTA complies the connection to a database only by means of Microsoft Jet Engine that resorts to the DAO access, namely only the Microsoft® Access ® version XX.XX DBMS.

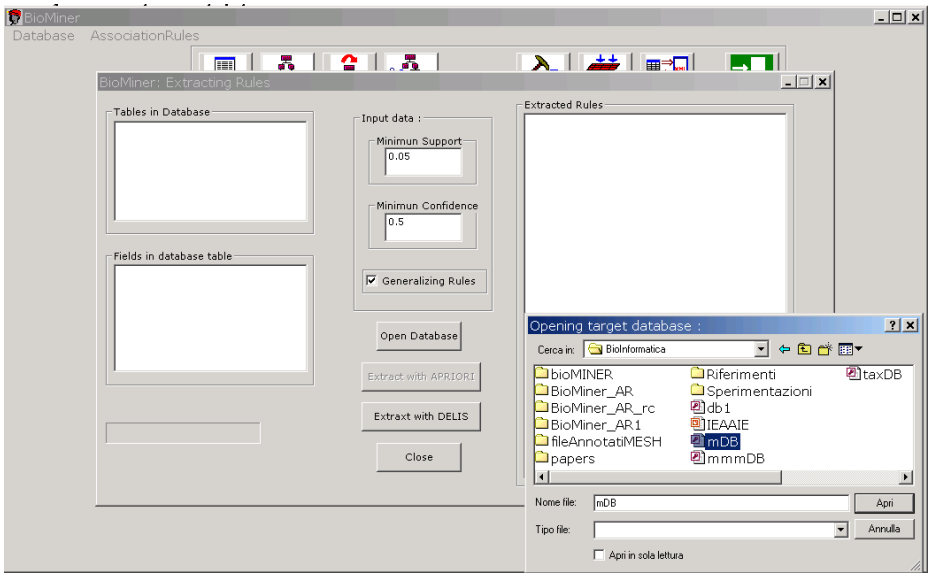


Fig. 2 Database configuration.

In order to select a data source, user should select the menu item *AssociationRules* > *ExtractRules* or to press the button  .
From the next new dialog window by pressing the button *Open Database* the user can access and connect to Microsoft Access target database (Fig. 2).

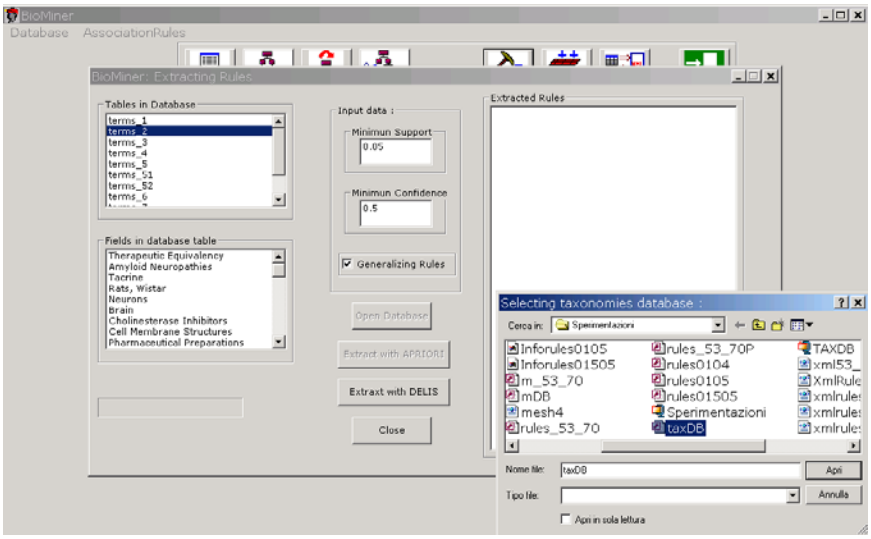


Fig. 3 Input data selection.

Then, table names (dataset) of the selected database is showed in the list window *Tables in Database*: the user can select the target dataset and field names are showed in list window *Fields in database table* (Fig. 3).

5. Create new data source

This functionality allows to create a new data source having to contain target database. More precisely, the input data originally consists of a set of tagged files representing a set of biomedical

scientific paper abstracts. By selecting the menu item *Database> Create MeshTerm DB* or by pressing the button  the window reported in Fig. 4 is showed.

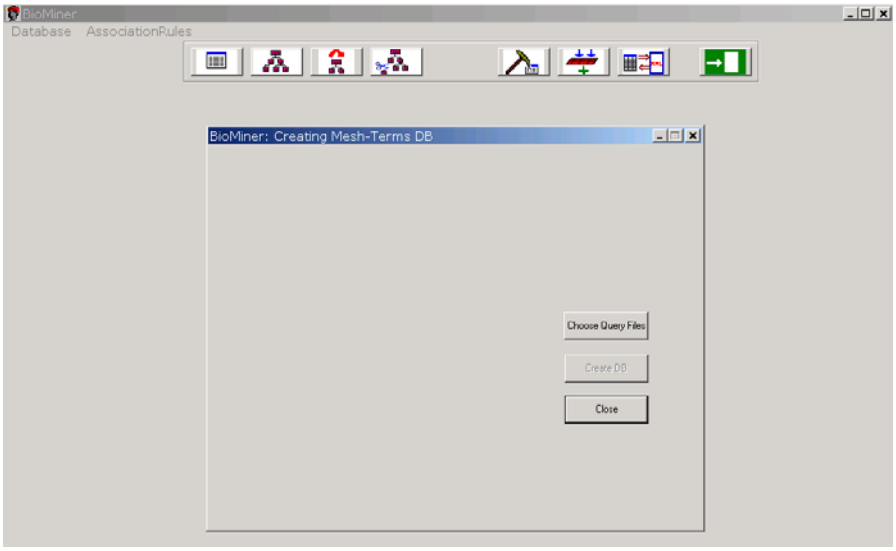


Fig.4 Text data preparation.

This operation allows the user to select the original tagged files (by pressing button *Choose Query Files*) and, subsequently, create a corresponding new Microsoft Access database (by pressing button *Create DB*, see Fig.5).

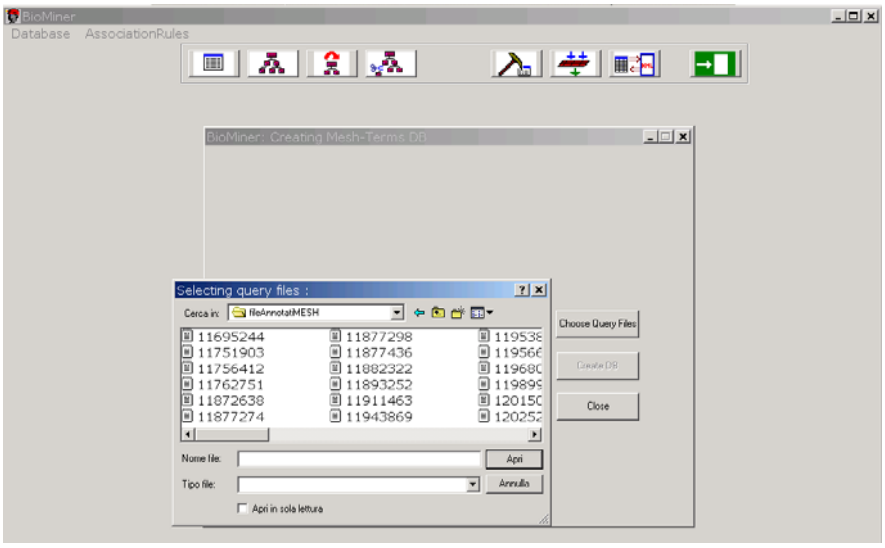
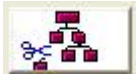


Fig.5 Text corpora DB creation.

6. Select taxonomy of terms

The taxonomical knowledge exploited in MTA consists of a vocabulary composed by a set of “is-a” taxonomies that the user can browse by selecting the menu item *Database > Select Taxonomies* or by pressing the button  .

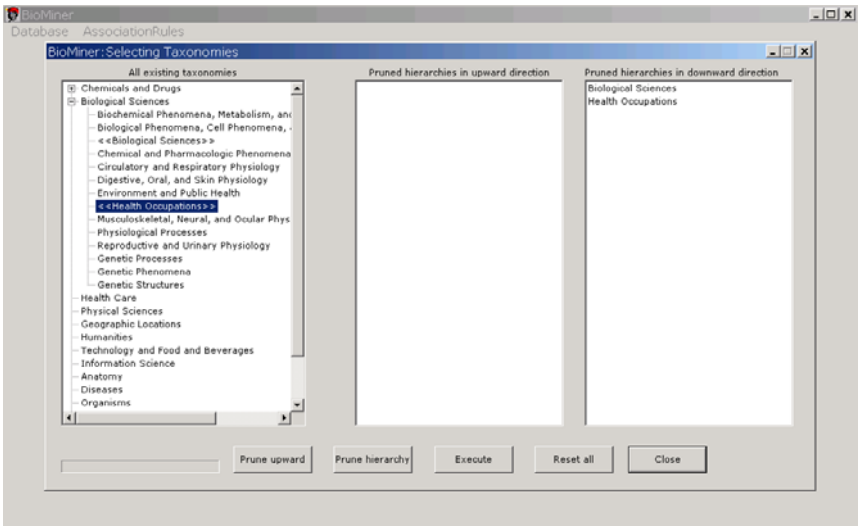


Fig. 6 Taxonomy navigation.

This step allows the user to select concept hierarchies that will play the role of taxonomical knowledge in the mining process. From the window showed in Fig.6 the selection of portions to be pruned is enabled by pressing the *Prune* buttons; the pruning process starts by pressing button *Execute*.

7. Load and Restore taxonomies

Taxonomical knowledge is loaded (Fig. 7) in the database by selecting *Database* > *Load Taxonomies* or by pressing the button  and subsequently, by selecting an XML file representing. The XML file is given in input to a parser and concept hierarchies and attributes are selected and preserved in the database interfaced by MTA. A previously pruned taxonomy can be completely restored by selecting the *Database* > *Restore Taxonomies* operation or by pressing the button  .

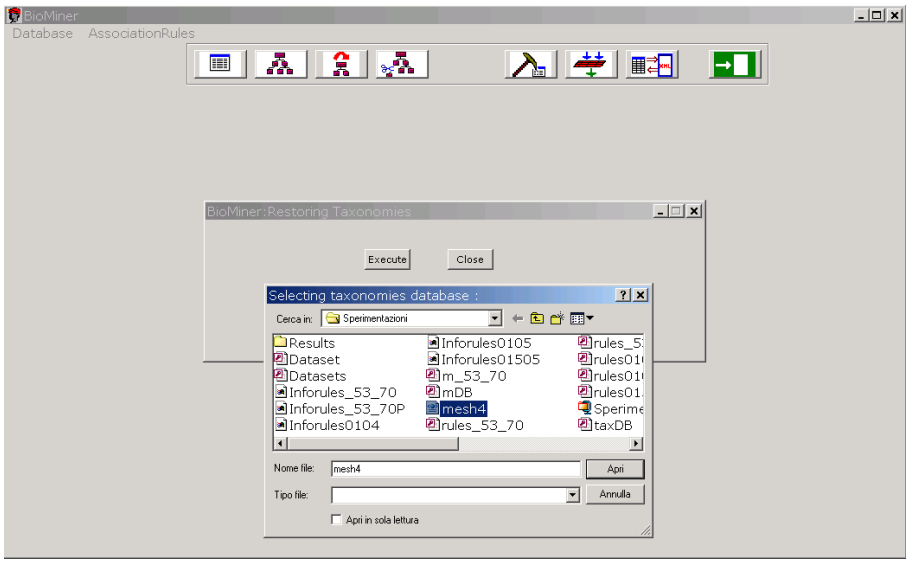


Fig. 7 Taxonomy load/restoration.

8. Mining Generalized Association Rules

MTA includes two algorithms for generalized association rule mining, namely the standard Apriori algorithm[3] and the DELIS (DEscending Levels Increasing Size) algorithm, which allows to discover non-redundant association rules resorting to the concept of closed item sets[4]. To start up the mining process the following steps should be performed:

1. selection of the *AssociationRules >Extract Rules* option or click on the  button;
2. selection of database storing the taxonomical knowledge (Fig. 8);
3. selection of the input data source, selection of the mining method, setting of the input parameters (*minimum support*, *minimum confidence*) (Fig. 9);
4. specification of the database name that will store discovered rules (Fig. 10);
5. discovered rules can be roughly browsed by the interface (Fig. 11).

[3] Agrawal, R., and Srikant, R., "Fast Algorithms for Mining Association Rules", *Proc. of the Twentieth Int.Conf. Very Large Databases*: Santiago, Cile,1994.

[4] Pasquier, N., Bastide, Y., Taouil, R., and Lakhal, L.. "Discovering Frequent Closed Itemsets for Association Rules", *Proc. of the 7th Int. Conference on Database theory*. C. Beeri and P. Buneman, Eds. Lecture Notes In Computer Science, vol. 1540. Springer-Verlag, London, 398-416, 1999.

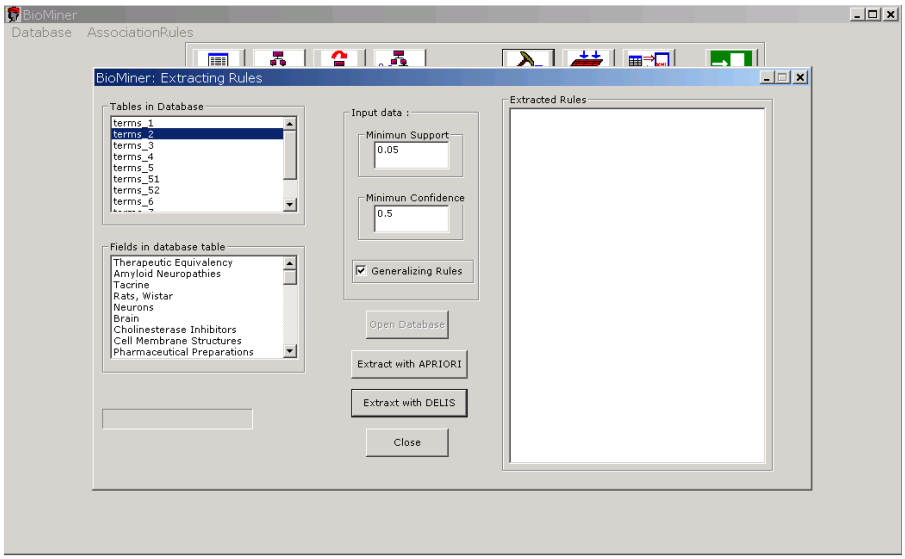


Fig. 9 Window to set input for the mining process.

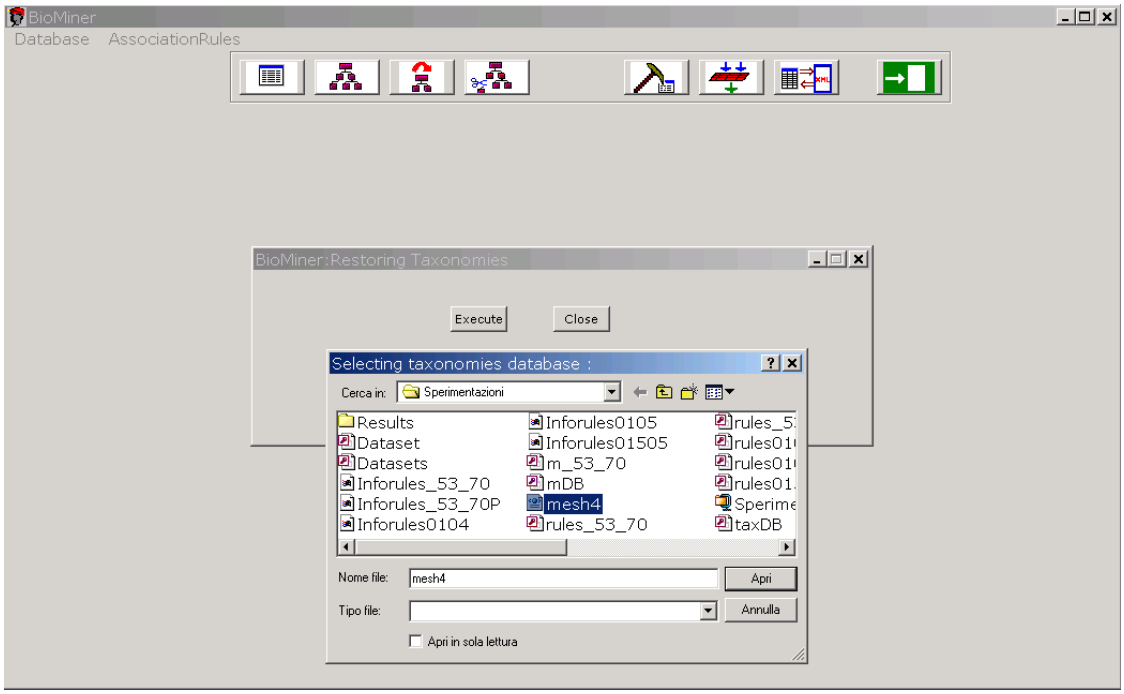


Fig. 8 Taxonomy selection.

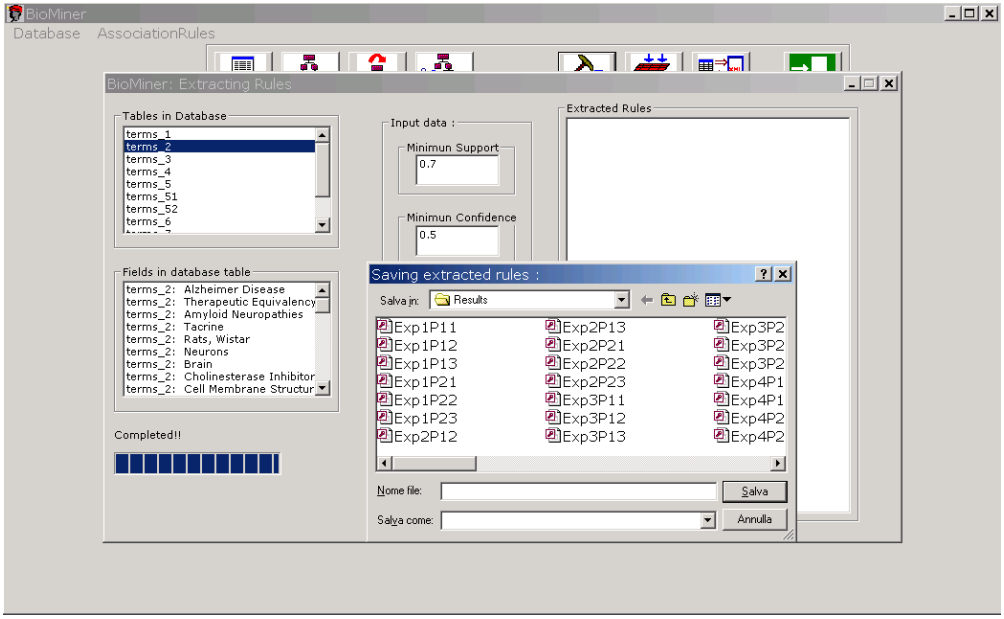


Fig. 10 Mining result storage.

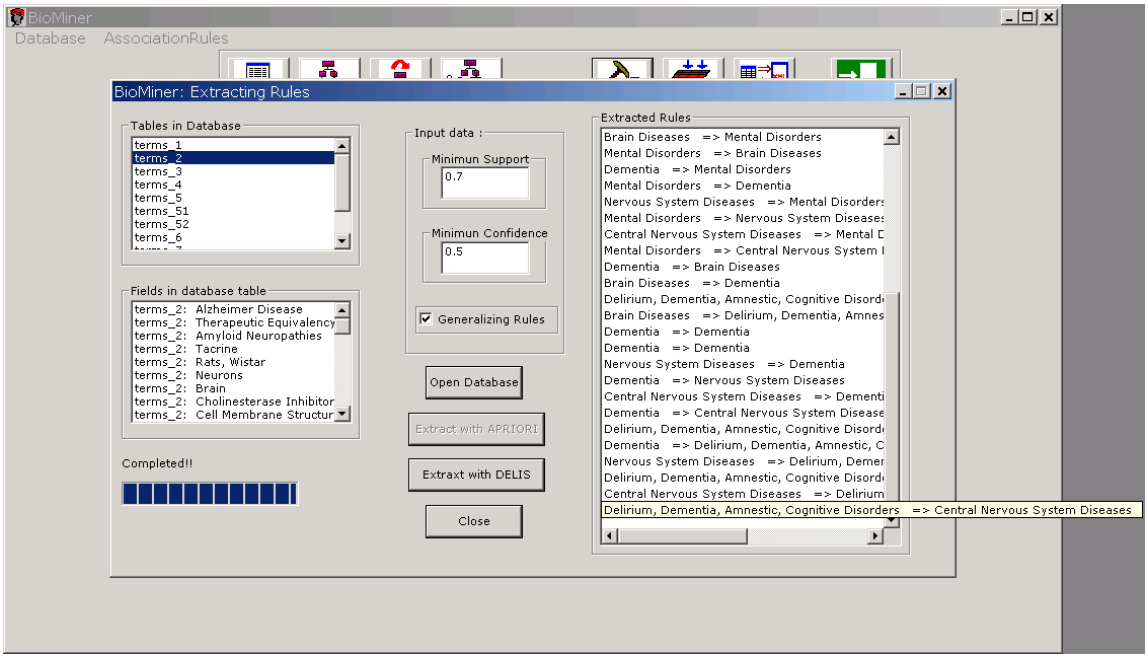


Fig. 11 Window showing the completion of a mining task.

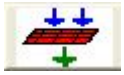
9. Filter Mined Association Rules

As the number of discovered rules can be very high, MTA supports some functionalities for rule filtering with the aim to browse among them looking for interesting ones.

Four modalities have been implemented in MTA:

1. *Template*, where the user can specify the kind of rules of interest;
2. *Cover*, where the rules are synthesized in order to identify the most concise ones;
3. *Rating*, interesting rules are identified on the basis of their statistical behaviour;

4. *Specificity*, where user can explore the rules by means of rule subspaces.

The task is started by selecting the menu item *AssociationRules* > *Filter Rules* or by clicking the  button that loads the window showed in Fig. 12 which allows the user to select the database of rules to be analysed (*Load Rules* button).

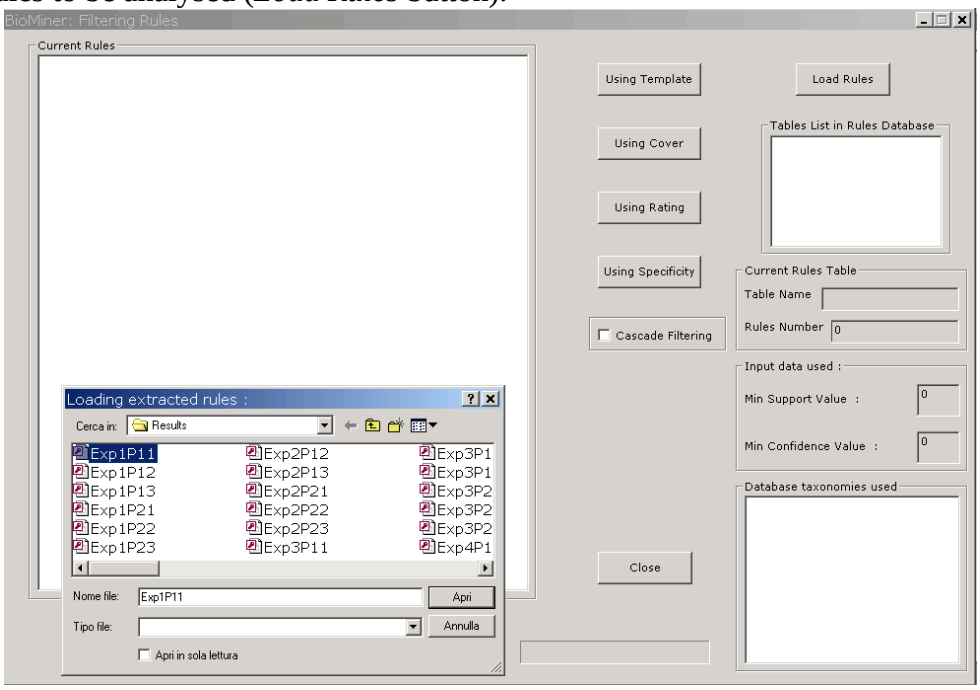


Fig.12 Window for rule filtering operations: rule database selection.

By selecting the table containing rules of interest (*Table List in Rules Database* frame), the system loads rules and some information describing the rules set are provided (number of rules, minimum support and confidence values, employed taxonomies (Fig.13). Then, the filtering technique can be selected and the result is showed to the user in the *Current Rules* frame (Fig.14). An option to combine different filtering techniques in a single filtering workflow that operates in a cascade mode is also provided.

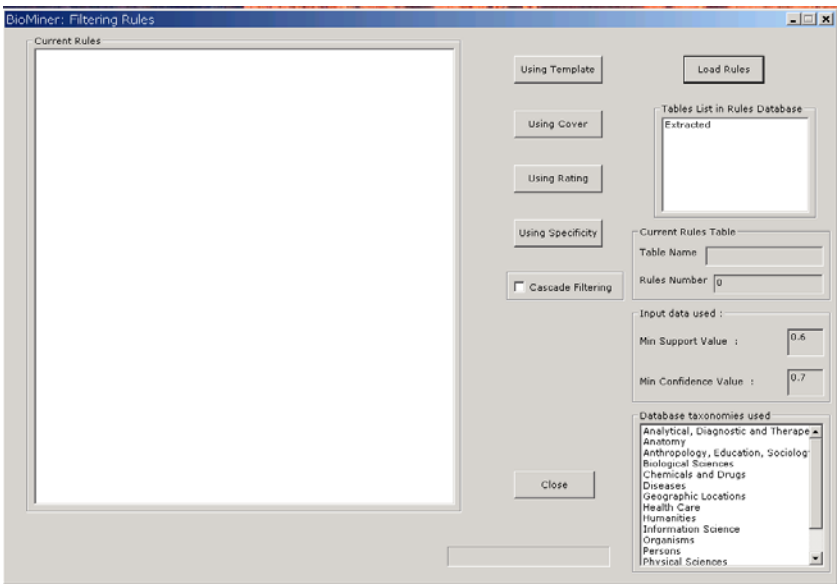


Fig.13 Window for rule filtering operations: rule loading and filtering method selection.

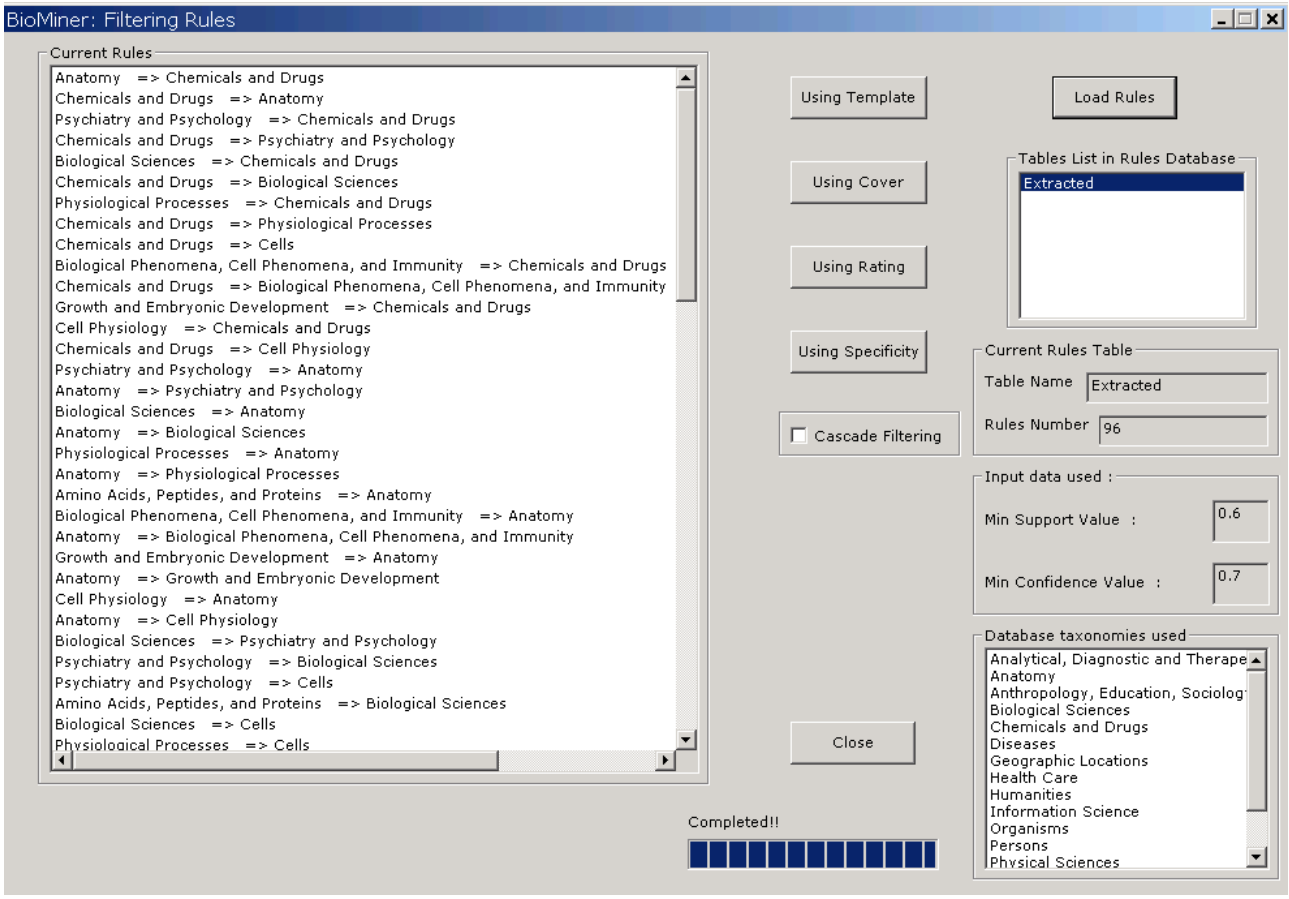


Fig.14 Window for rule filtering operations: example of a filtering task.

9.1 Template Filtering

The Template method allows users to specify the notion of interesting and not-interesting rules. By means of the scroll controls (Items, Taxonomies Classes, Cardinality) the user can specify the criteria that resulting rules should meet (*Inclusive Template*) and/or do not have to meet (*Restrictive Templat*e), see Fig.15. The resulting rules are ordered on the basis of support values (or confidence values), see Fig.16. Filtered rules can be saved (*Save Rules* button) in the same repository of analysed rules.

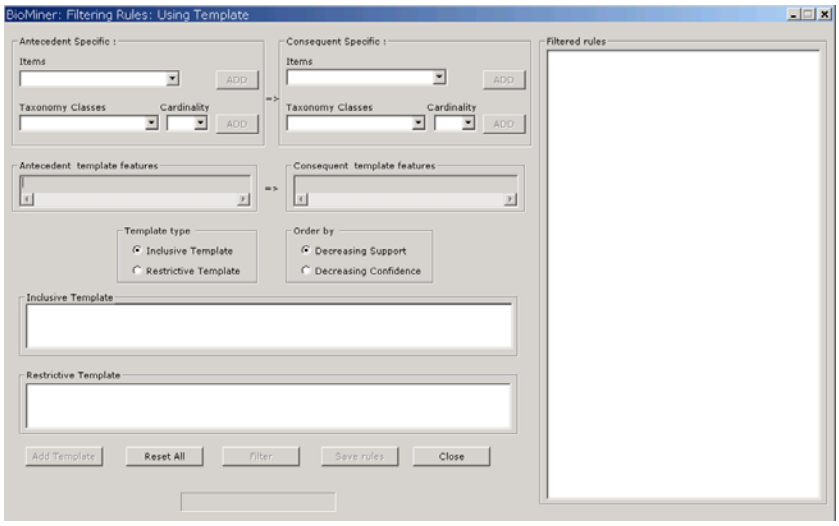


Fig.15 Template filtering set up.

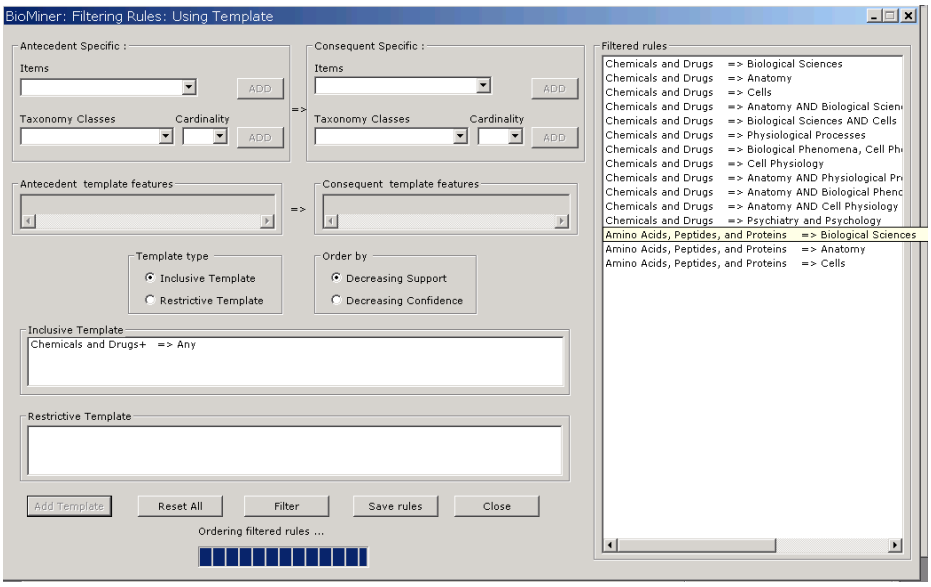


Fig.16 Template filtering results.

9.2 Cover Filtering

The Cover method allows to specify criteria to identify the most concise and non-redundant rules (Fig.17). Covers can be extracted both on the antecedent part and the consequent part of the rules (*Algorithm direction* frame). Moreover, covers can be combined with a clustering method (*Cover algorithm type* frame) and matching rules can be showed in two different modalities (*Order by* frame). Filtered rules can be saved as well (*Save Rules* button).

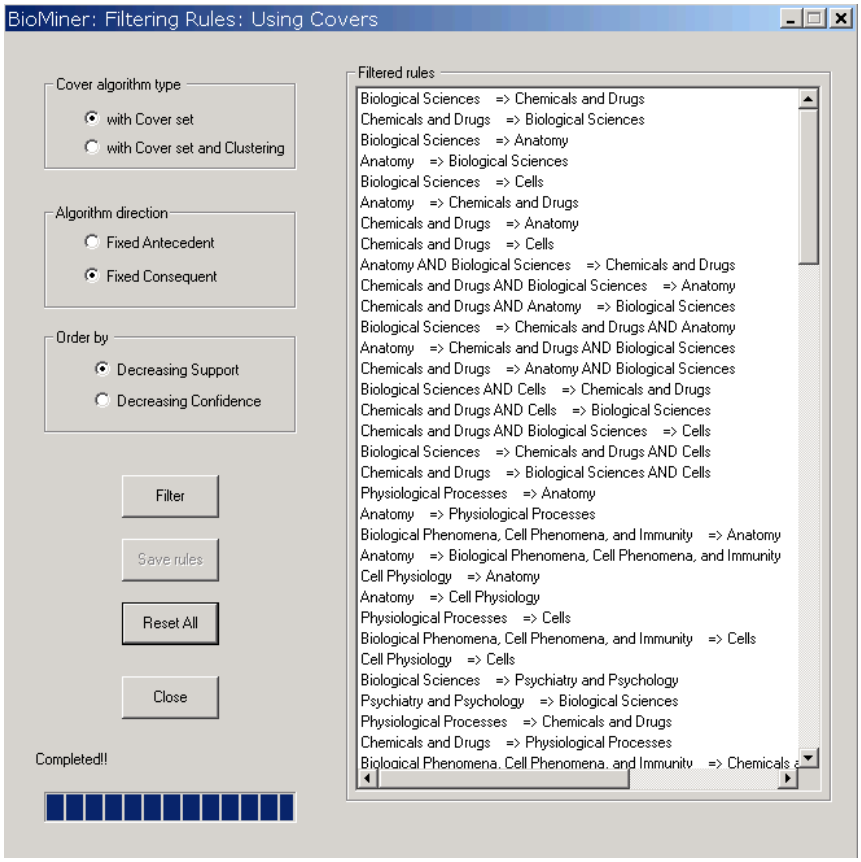


Fig.17 Cover filtering: setting and results.

9.3 Rating Filtering

This method allows to detect interesting rules on the basis of their values with respect to a statistical property. More precisely, the resulting rules are detected on the basis of their "estimated statistical behaviour" (*Statistical behaviour* buttons) with respect to a selected (*Statistical viewpoint* button) statistical measure (see Fig. 18). In particular, a minimum threshold value should be specified for the property Dependency. Filtered rules can be saved as well (*Save Rules* button).

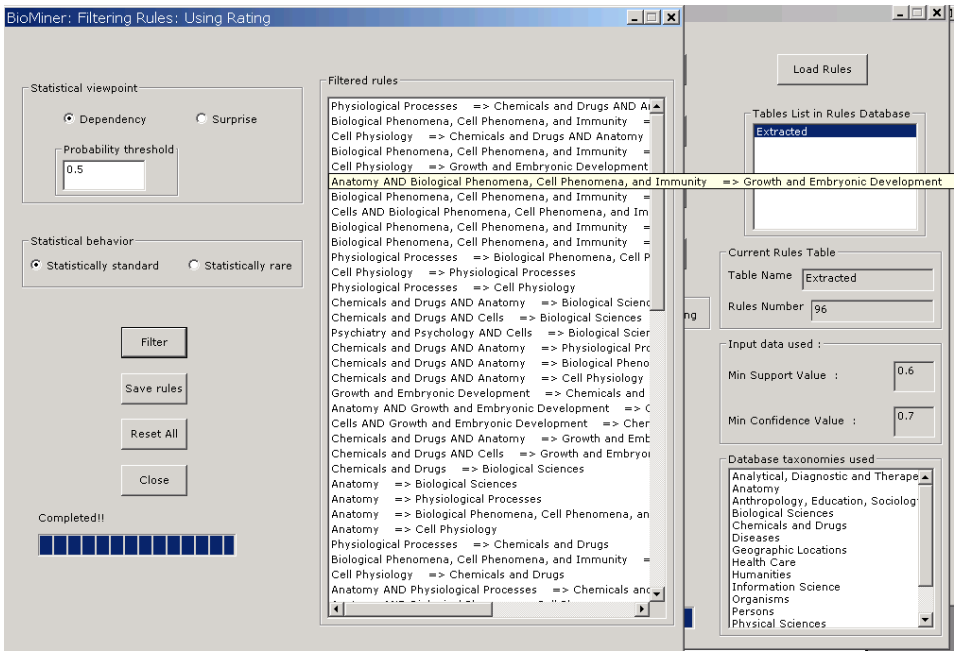


Fig.18 Rating filtering setting and results.

9.4 Specificity Filtering

This method allows to select subspaces of rules satisfying user interest and to explore more and more specific sets of rules. In particular, the user can browse towards subspaces that are the specialization of one of the two sides of a selected rule (that is the representative rule of a certain subspace). Subspaces of interest are identified by defining one or more nodes of the taxonomy (i.e., biomedical concepts) that are relevant for the user (*Ground Items* scroll control) and by selecting preferences on rule parts to be explored (*Antecedent/Consequent* radio button), see Fig.19. Once the initial subspace is obtained, the user pursues his exploration by choosing a single rule (by means of the *Filtered Rules* frame), then the next subspaces are generated by specializing one side (*More specific* button) as well as by specializing the other side (*Enhance other side* button), see Fig.20.

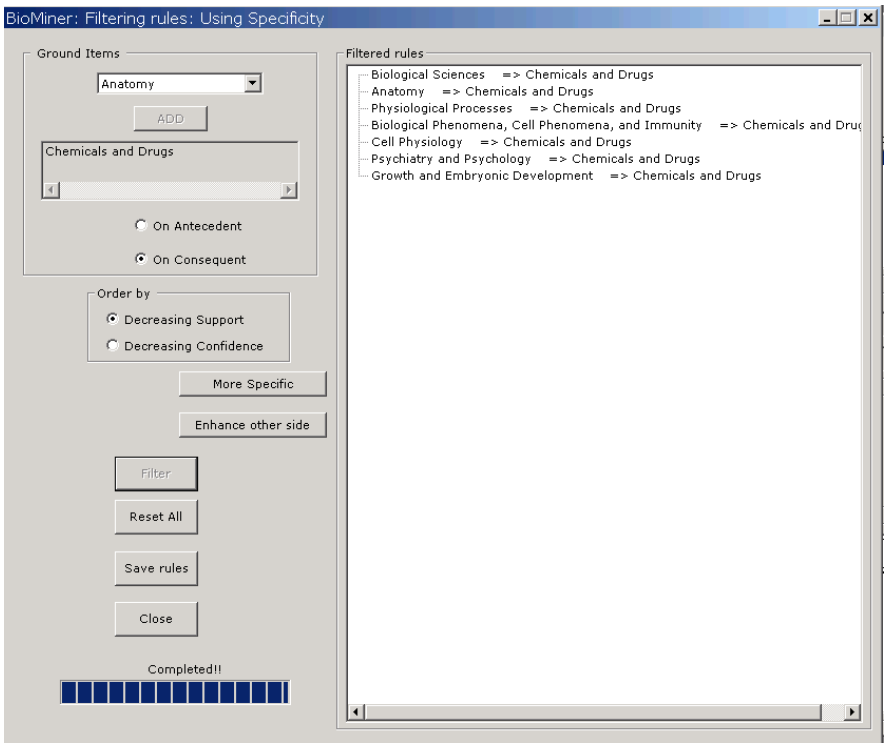


Fig.19 Specificity filtering method: setting and results.

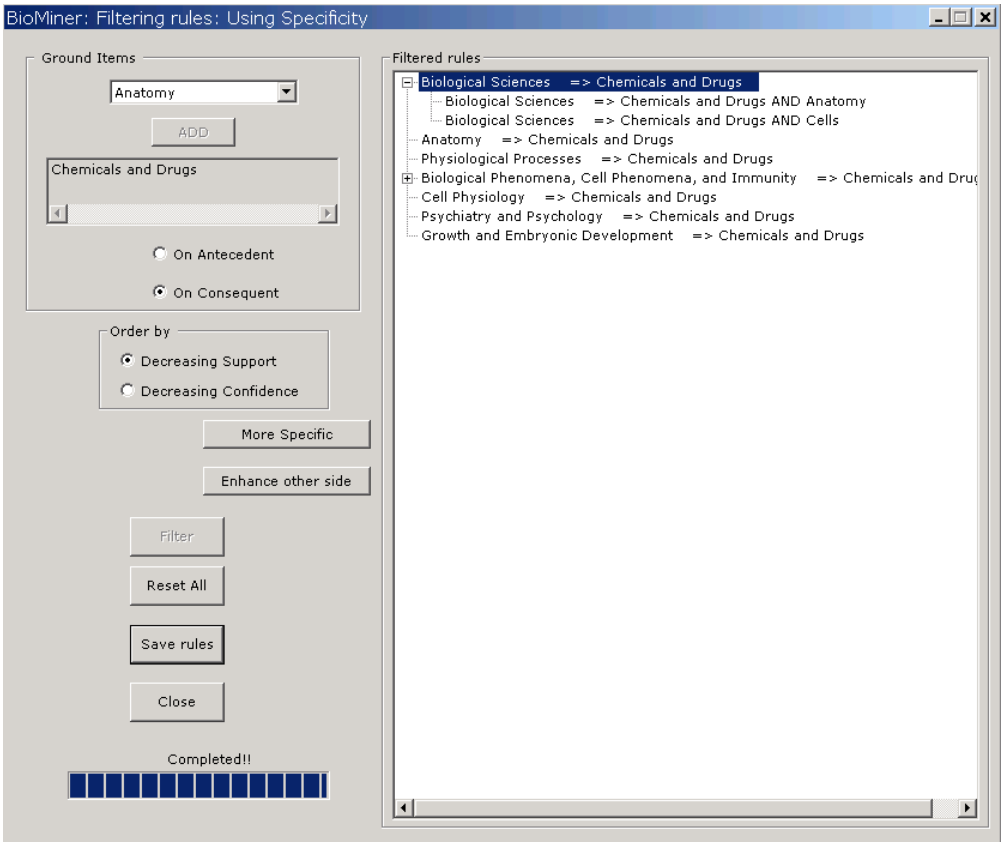
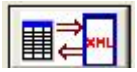


Fig.20 Specificity filtering method: subspace browsing.

10. Import/Export Association Rules

MTA supports the importing/exporting of association rules from/to a data-mining description model compliant with the PMML language[5]. This functionality is accessible by clicking on the  button or by selecting the menu item *Association Rules> Import/Export Rules*. In order to export rules, the user should select the data source (button *Select Database*) containing rules stored as results of previous mining sessions (see Fig. 22). Then, from the list shown in the *Rules Tables in Database* box, a rule set should be selected. The export step is accomplished by specifying the name of target XML file (see Fig. 23), the database storing the taxonomical knowledge and by pressing the button *Export* .

To import rules, the user should select the source XML file (*Select XML file* button). The import step is accomplished by specifying the database storing the taxonomical knowledge, pressing the button *Import* and by specifying the name of database that will store the imported rules.

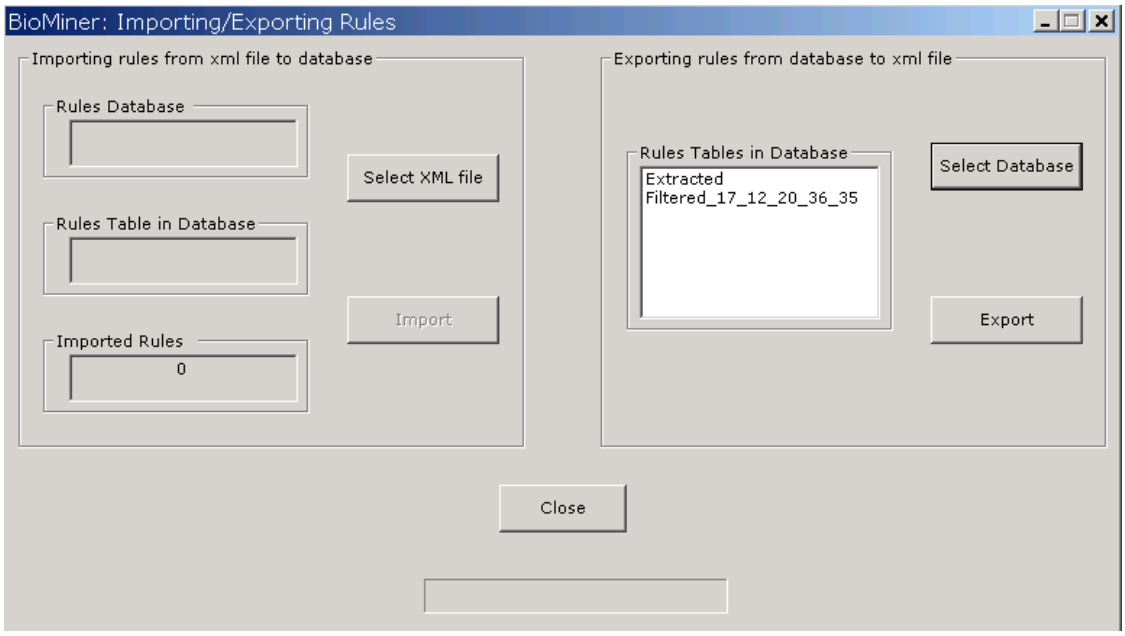


Fig.22 Rule Import/Export: selection of rules.

[5] http://www.oasis-open.org/cover/pmml.html

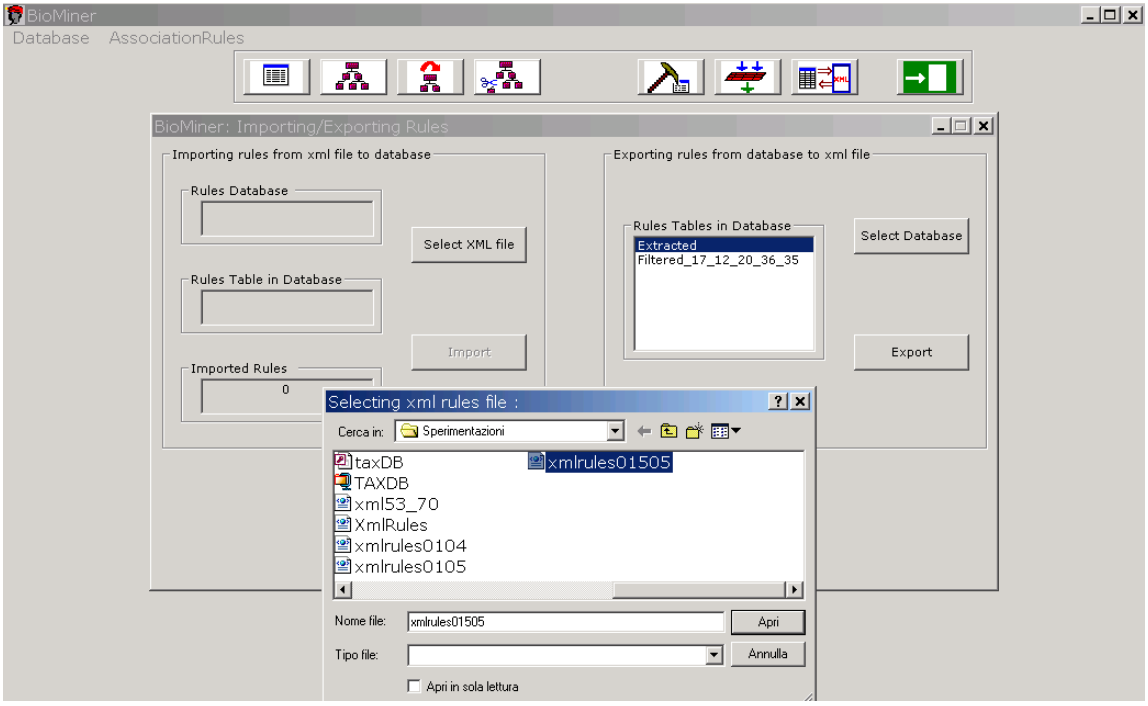


Fig.23 Rule Import/Export: selection of xml file.