

eXiTCBR version 4

Tutorial (getting started)

1 Targeted users

This tool is designed for conducting experiments under the Case-based Reasoning methodology. Do not expect to have a prototype nor a final product with it.

Target users: Health care and Medical staff. You can use the tool to compare different results regarding the inclusion or exclusion of some variable, learning the importance of the variables, changing the population, etc. Task: diagnosis/prognosis (+/-), or classifying over a set of n possible class values (or labels).

Other users:

- Applied sciences: Conduct experiments as in the health care staff.
- Engineers: As a result of your experiments, you can determine the parameters and methods that are more suitable for building your final CBR software.
- Students: Learn about CBR.

2 The tool

The following Figure shows the main Graphical User Interface (GUI) of eXiTCBR. Basic functionalities, explained in this tutorial, are highlighted.

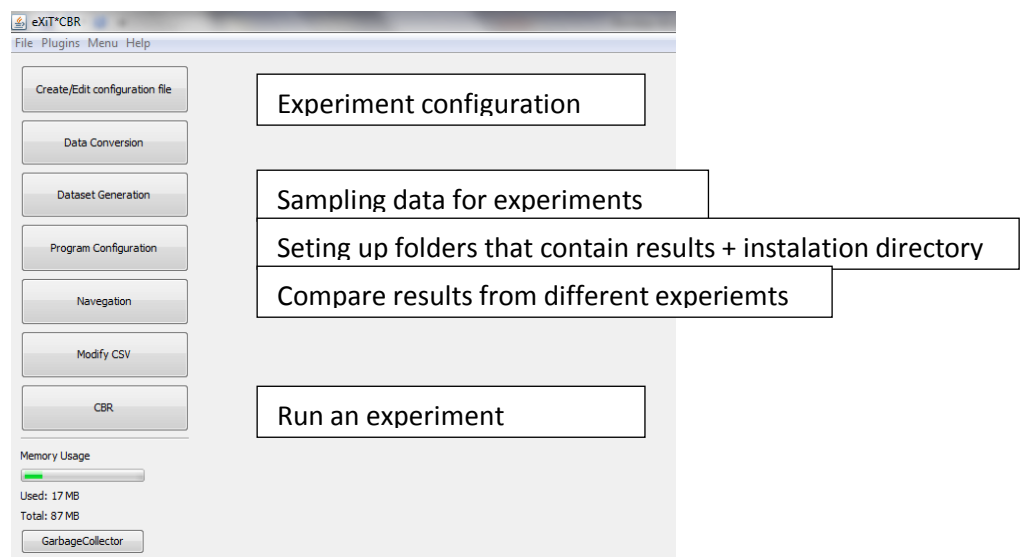


Figure 1. Main window of the tool.

3 Download and installation

1. Follow the instructions in <http://exitcbr.udg.edu/> to download the tool
2. Execute the installation file.
3. The GUI (Figure 1) will be prompted.
4. Select the “Program Configuration” button. The following window should be prompted:

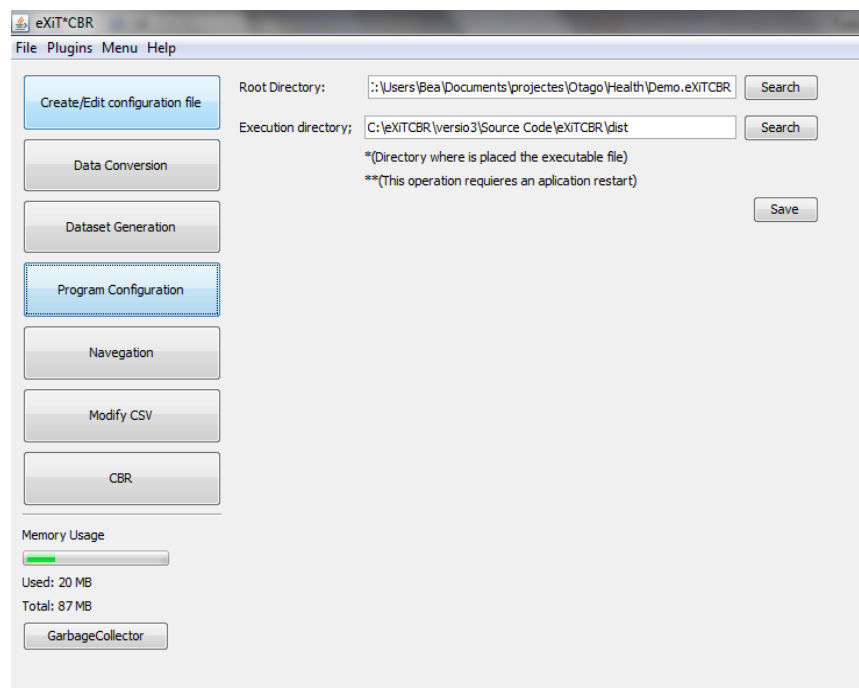


Figure 2. Program Configuration window.

5. Root directory: search in your file system the place you want the results be stored
6. Execution directory: if you have answered the default installation options, the directory should be as the one shown in Figure 2. Otherwise, you need to indicate here the directory you provided in the installation process.

4 Basic procedure

1. Adequate your data according to the admitted input files (see Section 4)
2. Generate sampling datasets to carry out your experiments (see Section 5)
3. Define your experiment (see Section 6).
4. Run your experiment (see Section 7)
5. Compare different experiment results (see Section 8)

5 Data files

Data files admitted by eXiTCBR are csv files (editable with excel). Some file conversions are available in the Data Conversion button, but they are out of the scope of this tutorial.

The files looks like the one provided in Figure 3. Rows are cases (records or examples). Columns are attributes.

	A	B	C	D	E	F	G	H	I
1	Id	Temperature	Occurrence	Lumbar pain	Urine pushing	Micturition	Burning of ur	Inflammatory	Nephritis of n
2	Codigo	Temp	Nausea	LumbarPain	Urine	Micturition	Burning	Inflammatory	Nephritis
3	-1	1	0	0	0	0	0	-1	-1
4	0	1	1	1	1	1	1	0	0
5	1	35,5	no	yes	no	no	no	0	0
6	2	35,9	no	no	yes	yes	yes	1	0
7	3	35,9	no	yes	no	no	no	0	0
8	4	36	no	no	yes	yes	yes	1	0
9	5	36	no	yes	no	no	no	0	0
10	6	36	no	yes	no	no	no	0	0
11	7	36,2	no	no	yes	yes	yes	1	0
12	8	36,2	no	yes	no	no	no	0	0
13	9	36,3	no	no	yes	yes	yes	1	0
14	10	36,6	no	no	yes	yes	yes	1	0
15	11	36,6	no	no	yes	yes	yes	1	0
16	12	36,6	no	yes	no	no	no	0	0
17	13	36,6	no	yes	no	no	no	0	0
18	14	36,7	no	no	yes	yes	yes	1	0
19	15	36,7	no	yes	no	no	no	0	0
20	16	36,7	no	yes	no	no	no	0	0
21	17	36,8	no	no	yes	yes	yes	1	0
22	18	36,8	no	no	yes	yes	yes	1	0
23	19	36,9	no	no	yes	yes	yes	1	0
24	20	36,9	no	yes	no	no	no	0	0
25	21	37	no	no	yes	yes	no	1	0

Figure 3. Input file example (data from Acute Inflammation dataset of UCI Repository)

Requirements:

- First row: description of the attributes
- Second row: short name of the attributes, in a single word
- Third row: attribute type:
 - o 0: discrete (as Nausea, LumbarPain, Urine, Micturition, Burning)
 - o 1: numeric (as Temp)
 - o 2: text
 - o 3: date
 - o 4: sequence
 - o -1: do not take into account (as identifier, or the classes).
- Fourth row: attribute weight
- First Column: case identifier. There should be a different identifier for all of the cases
- Class attribute: it should be numeric or discrete. In the example, Inflammatory is a diagnosis (class attribute), as well as nephritis.
- No empty lines at the end of the file
- Decimal numbers expressed with dots (39.0). Do not use dots for thousands.

Checking your input file:

- If you edit the file with an ASCII (simple) editor, each column should be separated by the token “;”, as in the Figure below.

```
1 Id;Temperature of patient ;Occurrence of nausea ;Lumbar pain ;Urine pushing ;Micturition
2 Codigo;Temp;Nausea;LumbarPain;Urine;Micturition;Burning;Inflammatory;Nephritis
3 -1;1;0;0;0;0;0;-1;-1|
4 0;1;1;1;1;1;1;0;0
5 1;35,5;no;yes;no;no;no;0;0
6 2;35,9;no;no;yes;yes;yes;1;0
7 3;35,9;no;yes;no;no;no;0;0
8 4;36;no;no;yes;yes;yes;1;0
9 5;36;no;yes;no;no;no;0;0
10 6;36;no;yes;no;no;no;0;0
```

Figure 4. The tokens used in the csv should be “;”.

Finally: you can have a single input file or multiple input files. The latest case is when you are using a multi-agent approach (out of the scope of this tutorial).

6 Data set generation

Cross validation is a very well-known technique as to repeat your experiments by a given number of times, and obtain unbiased results.

Click the “Dataset Generation” button of the left panel to obtain, from a single csv input file, several folders to follow a cross validation approach in the experiment. The window is shown in Figure 5.

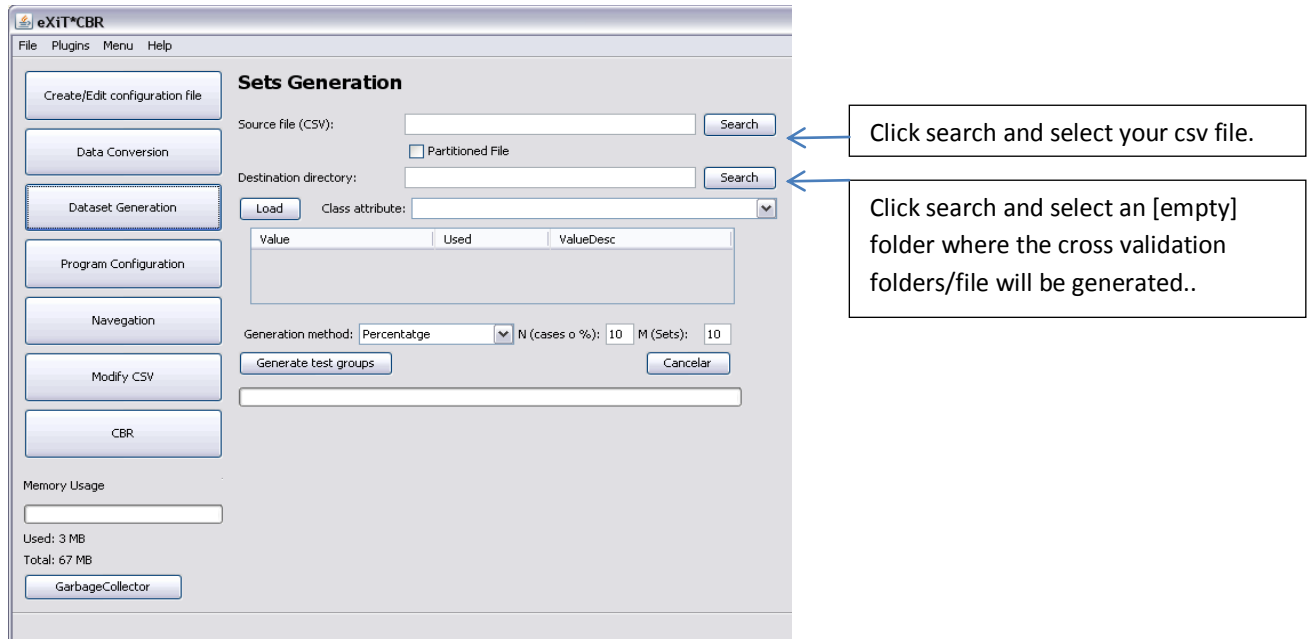


Figure 5. Dataset Generation.

Steps:

1. Select the source file as explained in the Figure
2. Select the destination directory as explained in the Figure.
3. Click on the “load” button. The different attributes of your file will be loaded as option in the class attribute pop-up menu.
4. Select the attribute in the class attribute pop-up menu that corresponds to the class.
5. Select, in “Used” column, the labels of the class that you want to take into account. Individuals with unselected class label will be ignored.
6. Change (if wanted) the name of the generated test files for each value, in the “ValueDesc” column.
7. Generation method: you can provide a given number of cases for control (15 cases) or express them in a percentage. By default, the tool offers the percentage. Use it.
8. Click the “Generate Tests Groups” button.

Following the “acute inflammatory” example, the window should look like the one displayed below:

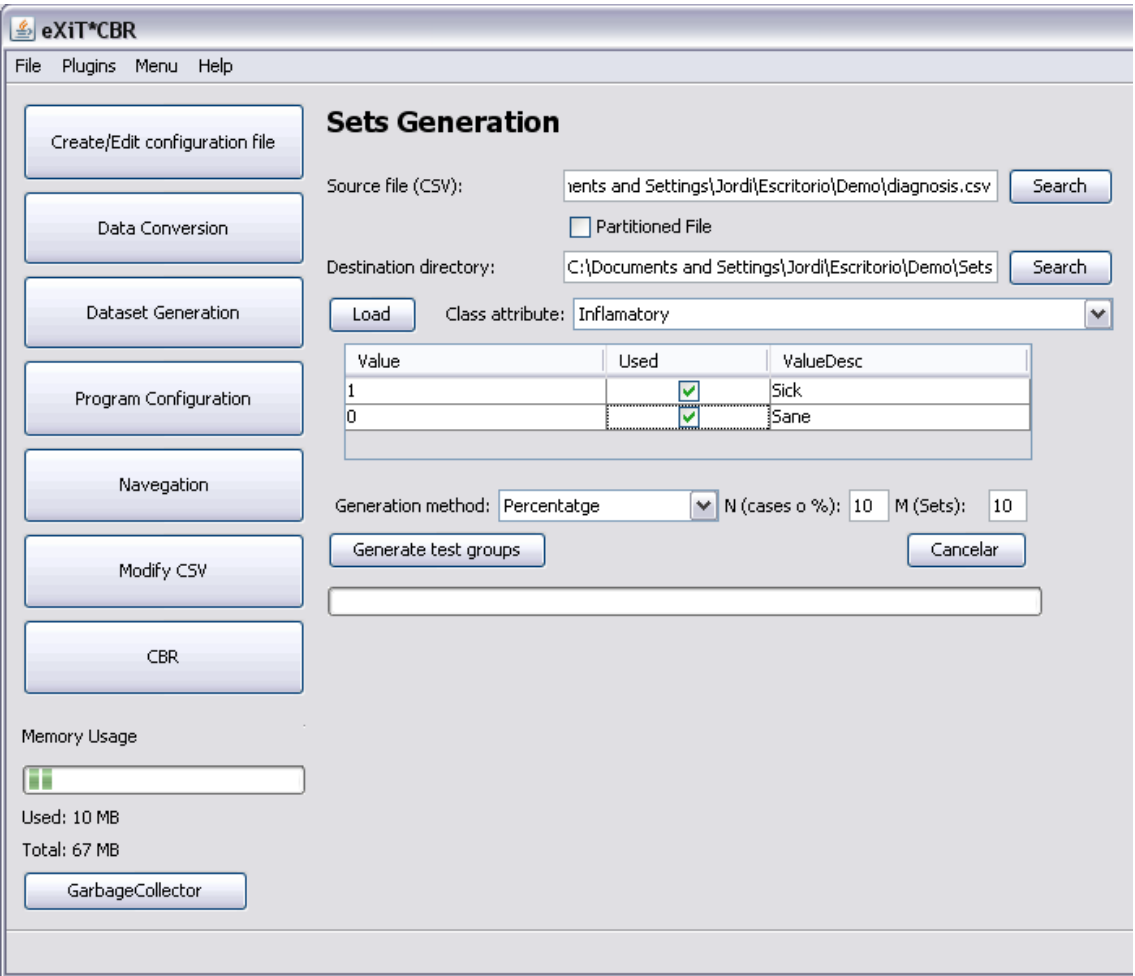


Figure 6. Dataset Generation example

7 Experiment configuration

Click the first button of the left panel (“create/edit configuration file”) to configure an experiment, you need to fill up the information of the window shown in Figure 6.

The screenshot shows the 'eXIT*CBR' Configuration File window. The left sidebar has buttons for 'Create/Edit configuration file', 'Data Conversion', 'Dataset Generation', 'Program Configuration', 'Navigation', 'Modify CSV', 'CBR', and 'GarbageCollector'. The main area is titled 'Configuration File:' and contains several sections: 'Experimentation' (Measures: 1, Visualization: 2, Method: MAS_Multiple_Experiment), 'Data' (Destination folder, Sets directory, Class, Positive value, Negative value), 'Pre-proces' (Discretization: 3, Normalization: 3, Attribute Selection: null), 'Model Generation' (Retrieve, Unknown, Local Distance, Numeric, Discrete: 4, Text, Date, Sequence, Global Distance, Selection), and 'Multiagent' (Weigh assignment, Directory, Coordination strategy). The bottom section includes 'Saving directory' (5) and 'Load Configuration File'.

Figure 7. Experiment configuration.

7.1 Step 1. Defining the kind of experiment.

Measures: all possible measures for your experiment are provided by default, so you do not need to specify none of them here. The measures included are the following:

- Successful and failure rates
- Confusion matrices
- Binary confusion matrices
- ROC curves according to the parameters selected in step 4 (CBR methodology)
- AUC (averages and deviations)

Visualization: all measures will be visually provided. No options available here yet to select which ones you wish to watch.

Method. Options available are the following:

- MAS_Multiple_Experiment: multiagent system, cross validation
- MAS_Eingle_Experiment: multiagent system, single run
- Multiple_Experiment: single input csv file, cross validation.

- Single_Experiment: single input csv file, single run
- Multiple_Experiment_Probabilistic: single input csv file, cross validation. Complementary experiments are made for every test dataset. **Choose this.**
- Single_Experiment_Probabilistic: single input csv file, single run. Complementary experiments are made for every test dataset.

7.2 Step 2. Specifying the dataset.

Destination folder: write here a name so as you can recognize your experiment later on (for example "Test1").

Sets directory: click the "search" button and select the dataset generated (see previous Section 6).

Class: enter the attribute name corresponding to the class in the csv file (for example "inflammatory").

Positive value: enter positive value of the class. This step is only needed in experiments using classification methods for binary classes (i.e. classes having exactly two possible labels in which one of them is considered positive and the other one is considered negative). Probabilistic experiments are one of them. **For inflammatory, enter "1".**

Negative value: enter negative value of the class. Same use case as positive value. **For inflammatory, enter "0".**

7.3 Step 3. Pre-process.

Discretization: Here you have available different discretization methods, to change your numeric variables to discrete values. This is not mandatory, it is an option. **Choose null** (so as you will not discretize your numeric variables).

Normalization: Numeric variables should be normalized so that they can be comparable. Choose any methods of the available one, or just leave the default method.

Attribute selection: This is an advanced feature, out of the scope of this tutorial.

7.4 Step 4. Case-Based Reasoning methods.

You should be familiar with the Case-Based Reasoning methodology to configure your experiment. Basically, it consists in four main phases:

- 1) Retrieve: recovering cases from memory, following a given similarity metrics and a selection method.
- 2) Reuse: reuse a case according to some reuse method.
- 3) Revise
- 4) Retain

Current eXITCBR version provides support for the first two steps.

Regarding the Retrieve phase, you need to specify several methods:

1. Unknown. This method deals with the unknown attribute values. **Leave the one by default.**
2. Local Distance. Numeric. Here you need to specify the distance measure used to compare two numeric attribute values. By default, all local distances are provided. There is only one option available for the moment.
 - DistanciaLocalEuclidiana. Implements the Euclidean distance between to numeric values. Choose one as you first trial of the tool.
3. Local Distance. Discrete. Here you are required to select one distance measure for dealing with discrete values. There is only one option available at this moment.
 - DistanciaLocalHamming. Implements the Hamming distance between to discrete values.
4. Local Distance. Text. Here you are required to select one distance measure for dealing with text values. There is only one option available at this moment.
 - DistanciaLocalHamming. Implements the Hamming distance between to discrete values.
5. Local Distance. Date. Here you are required to select one distance measure for dealing with dates. There is only one option available at this moment.
 - DistanciaDataStandard. Calculates the distance between two dates as the difference of milliseconds between them.
6. Local Distance. Sequences. Here you are required to select one distance measure for dealing with sequences. **No action is needed in this tutorial since we are not dealing with sequences.**
 - DistanciaLocalSequencesDIST4MADCAPPablo. To deal with sequences¹.
 - DistanciaLocalSequencesLaneAndBroadly. To deal with sequences².
 - DistanciaLocalSequencesMarcCompta. To deal with sequences³
7. Global Distance. The distance measure here is used to combine the different individual distances of attributes. Options available are the following:

¹ P. Gay, B. López, J. Meléndez. Sequential learning for case-based pattern recognition in complex event domains. Proceedings of the 16th UK Workshop on Case-Based Reasoning. 13th December 2011 Edited by Miltos Petridis, pages 46-55.

² Lane, T., Brodley, C.E.: Temporal Sequence Learning and Data Reduction for Anomaly Detection. ACM Transactions on Information and System Security 2(3),295–331 (1999)

³ Marc Compta, Beatriz López. Integration of Sequence Learning and CBR for Complex Equipment Failure Prediction. A. Ram and N. Wiratunga (Eds.): ICCBR 2011, LNAI 6880, pp. 408–422, Springer 2011.

- DistanciaGlobalEuclidiana. This method is the Euclidean average. **Choose this one as your first attempt to use the tool.**
 - DistanciaGlobalMitjana. This method is the mean.
 - DistanciaGlobalNCM. Use this method if you have selected the corresponding local distance.
 - DistanciaGlobalNCM2. Use this method if you have selected the corresponding local distance.
 - DistanciaGlobalSequences. Use this method if you are dealing with sequences.
 - DistanciaGlobalWM. This method is the weighted mean.
8. Selection. Among the retrieved cases, the selection methods selects the best cases from with the solution of a new case is provided at the reuse phase. The selection methods available are the following:
- Seleccio1K. The best / most similar case will only be used.
 - SleccioNK. The N best similar cases will be used.
 - SeleccioThreshold. The cases close to the new case will be used. “Close” is modelled by a threshold that needs to be specified at the right box. **Use this option as a first attempt to use the tool, indicating in the right box “0.7”**. This option illustrates in a nice way the plots that result after running the experiment.
 - Null. No method is applied. All the cases are used.

Regarding the Reuse phase, you need to specify a single method:

Reuse: The method indicated here build the solution for a case from the cases retrieved and selected in the retrieve phase. The methods available are the following:

- ReuseCopia: Copy the solution from the best retrieved case to the new case.
- ReuseProbabilisticBliskaWolakAdaptat⁴. Only for binary classes.
- ReuseProbabilisticPous⁵. Only for binary classes. **Select that one since it provides good results in diagnosis environments**. It requires two parameters. Use “0.1, 0.7”, as shown in Figure 8 (see more information on the reference provided).

⁴ An adaptation of: Biliska-Wolak, A.O., Floyd Jr., C.E.: Development and evaluation of a case-based reasoning classifier for prediction of breast biopsy outcome with bi-radstm lexicon. Medical Physics 29(9), 2090–2100 (2002)

⁵ Pous, C., Gay, P., Pla, A., Brunet, J., Sanz, J., Ramon y Cajal, T., López, B.: Modeling Reuse on Case-Based Reasoning with Application to Breast Cancer Diagnosis. In: Dochev, D., Pistore, M., Traverso, P. (eds.) AIMSA 2008. LNCS (LNAI), vol. 5253, pp. 322–332. Springer, Heidelberg (2008)

- ReuseMajoriaK: the solution is determined according to the majority of votes regarding the selected cases. If this method or one of the four following methods is selected, SeleccionK will be forced as the selection method.
- ReuseMajoriaKPablo: this is a variation of the previous one.
- ReuseMajoriaKRandom: Same as ReuseMajoriaK, but ties are solved randomly.
- ReuseProbabilisticMulticlassKBliskaWolakAdaptat: Calculates a similarity value in the range [0,1] for each label of the class, in the same way that ReuseProbabilisticBliskaWolakAdaptat does, and the label with the highest similarity is selected. Two parameters are needed.
- ReuseProbabilisticMulticlassKPous: Calculates a similarity value in the range [0,1] for each label of the class, in the same way that ReuseProbabilisticPous does, and the label with the highest similarity is selected. Two parameters are needed.

7.5 Step 5. Saving the configuration

1. Click on the search button to indicate the folder where the configuration file will be stored.
2. Give a name to it at the end of the path.
3. Click the save button.

The following figure shown how the window looks like after entering the data for the “acute inflammatory” experiment:

The screenshot shows the eXIT*CBR configuration window with the following settings:

- Configuration File:**
 - Create/Edit configuration file: [Button]
 - Data Conversion: [Button]
 - Dataset Generation: [Button]
 - Program Configuration: [Button]
 - Navigation: [Button]
 - Modify CSV: [Button]
 - CBR: [Button]
 - Memory Usage: [Bar chart showing 10 MB used, 67 MB total]
 - GarbageCollector: [Button]
- Experimentation:**
 - Measures: NotSupportedYet
 - Visualization: Tries to visualize all the implemented methods
 - Method: Multiple_Experiment_Probabilistic
- Data:**
 - Destination folder: Test1
 - Sets directory: \Dataset [Search]
 - Class: Inflammatory
 - Positive value: 1
 - Negative value: 0
- Pre-proces:**
 - Discretization: null
 - Normalization: NormalizacioMaxMin
 - Attribute Selection: null
- Model Generation:**
 - Unknown: DesconegutsProva
 - Local Distance:
 - Numeric: DistanciaLocalEuclidiana
 - Discrete: DistanciaLocalHamming
 - Text: DistanciaLocalHamming
 - Date: DistanciaDataStandard
 - Sequence: DistanciaLocalSequencesDIST1MADCAPPablo
 - Global Distance: DistanciaGlobalEuclidiana
 - Selection: SeleccioThreshold 0.7
 - Reuse: ReuseProbabilisticPous 0.1,0.7
- Multiagent:**
 - Weigh assignment: CalcPesosFromFile
 - Directory: [Search]
 - Coordination strategy: CollaborationFunctionalBinary
- Bottom Section:**
 - Saving directory: C:\Documents and Settings\Jordi\Escritorio\Demo\test1.conf [Search] [Load Configuration File]
 - [Search] [Save]

Figure 8. Configuration File filled for the “acute inflammatory” experiment.

8 Running an experiment

Click on the CBR button, at the left panel. The following window should be prompted to you:

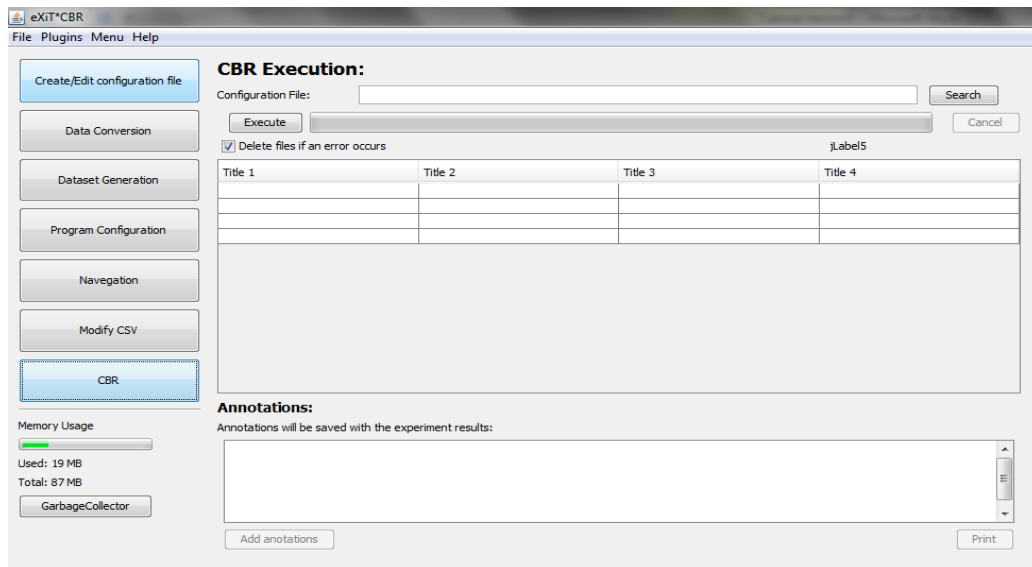


Figure 9. CBR Execution window.

To run your experiment:

1. Click on the search button and select the configuration file defined previously

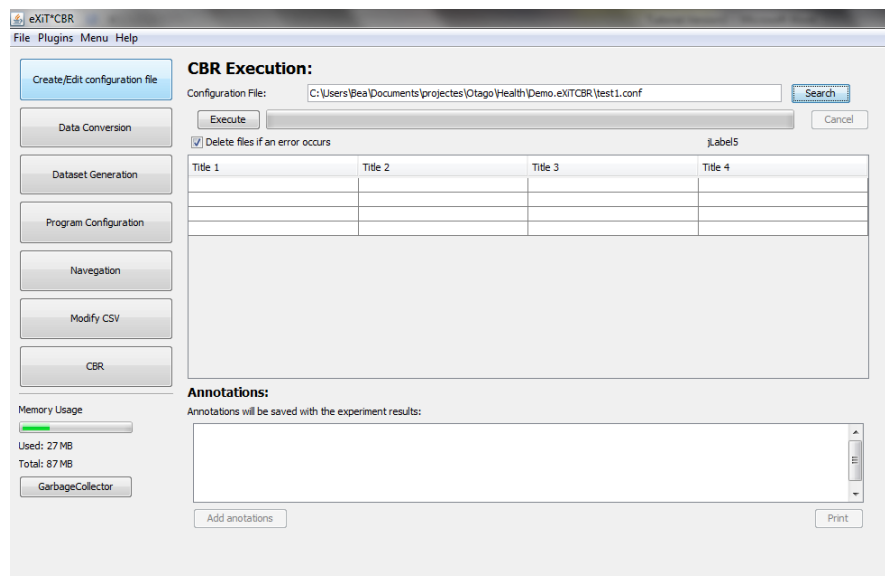


Figure 10. CBR Execution window.

2. Click on the “execute” button.

When the system finishes running the experiment, the different plots with the results will be provided to you. However, the Navigation facility facilitates the analysis of the results.

9 Experiment navigation

Click the “Navigation” button to analyse your results. The GUI should be like in the following Figure:

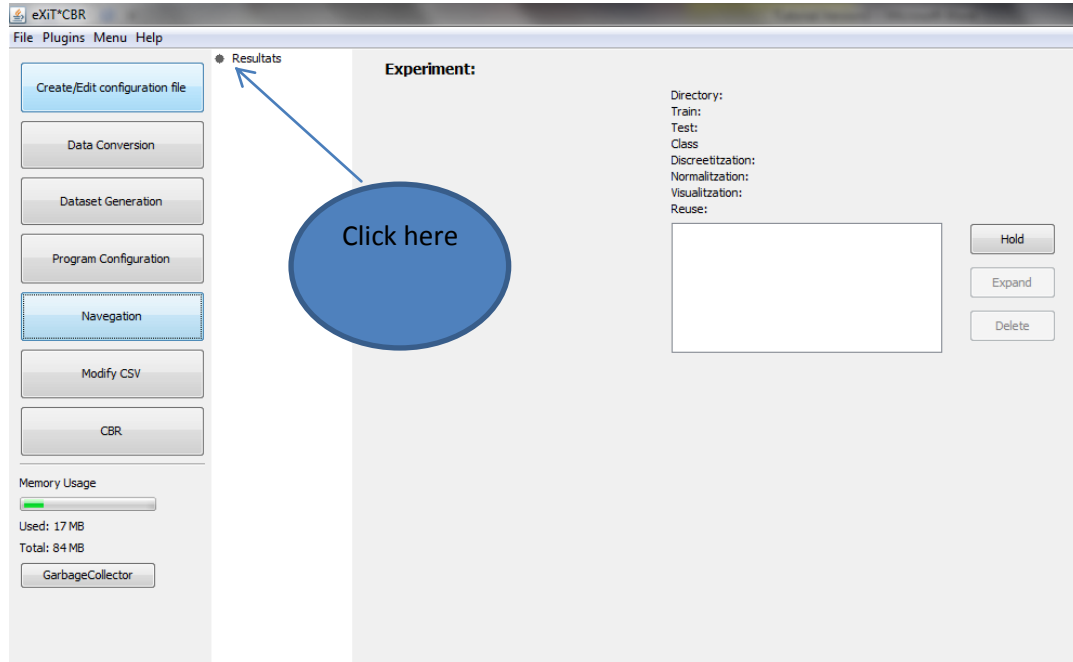


Figure 11. Navigation window.

Click on the first name that appears on the window, as shown in the Figure. Then, all the experiments carried out as so far will be prompted to you:

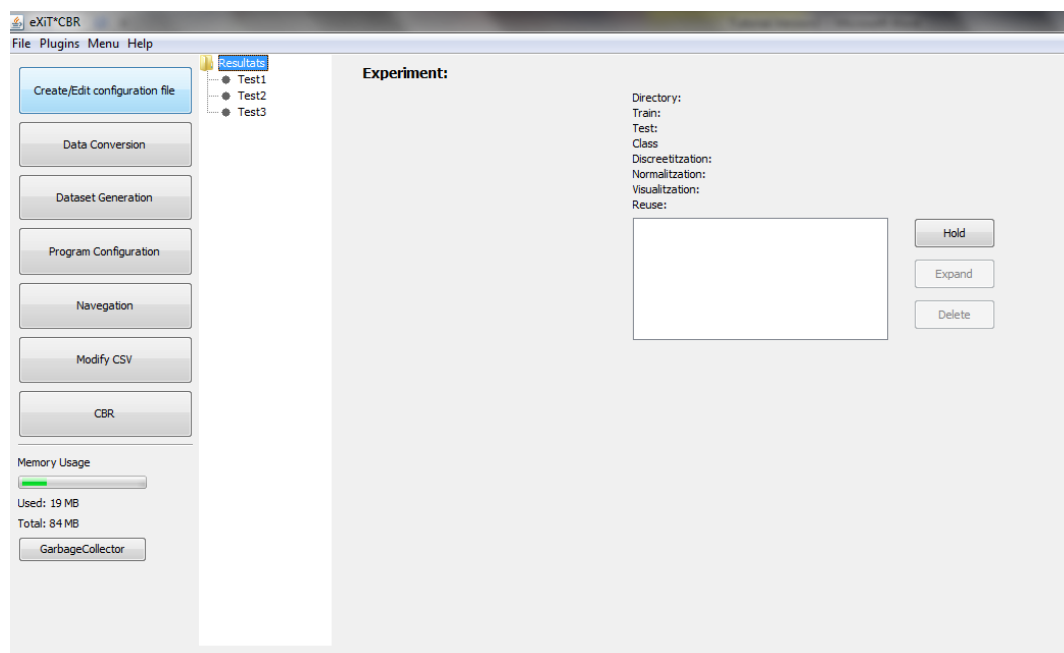


Figure 12. Navigation window.

Since the same experiment could be repeated several times, you can have under the same name different results, labelled according to a timestamp, as shown in the following Figure:

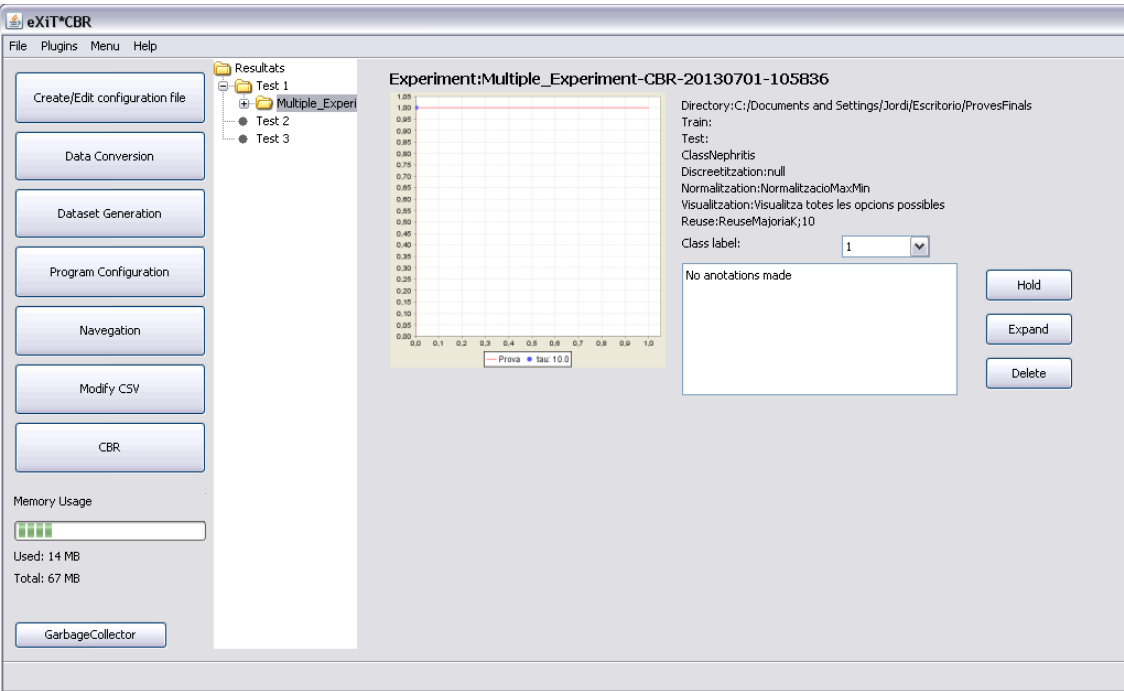
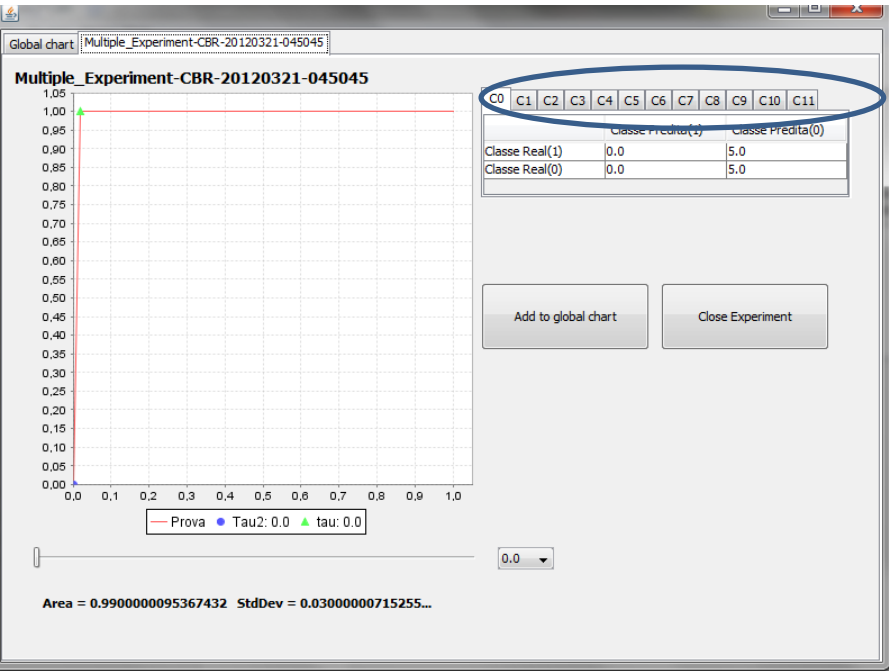


Figure 13. Navigation window. The ROC curves are displayed.

In the *Class Label* combo box, you can select the class label that is considered positive when calculating true positive, false positive, true negative, and false negative rates.

If you want to compare several experiments (test1 and test3 in the Figure, for example), you need to expand the plot, by clicking on the “expand” button.



Click on this tabs if you one to visualize the results individually, for each of the cross validation folder.
C0 contains the mean values.

Figure 14. Plot expanding

Do that for each of the experiments to be compared. As a result, a new window will appear. There will be a tab for each of the experiments under comparison:

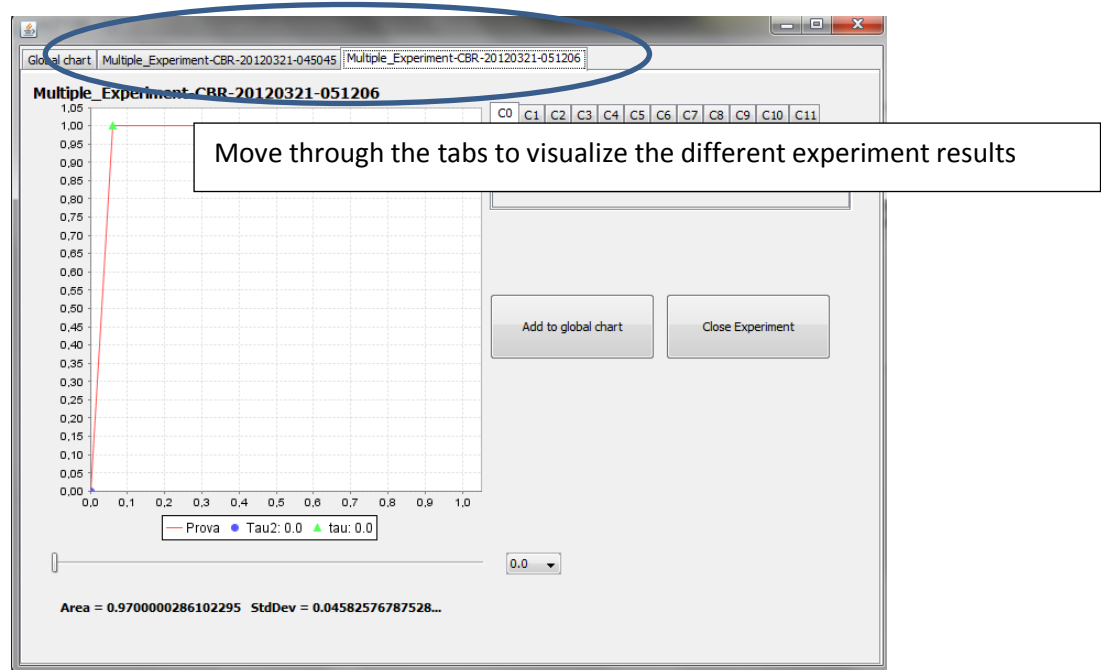


Figure 15. Plot expanding

The experiments aren't done only with the chosen tau for reuse method, but they are repeated several times using a sequence of tau values. The results of the sequence are displayed in a chart by means of a ROC curve, as can be seen in figure 15.

There are three labels under the chart.

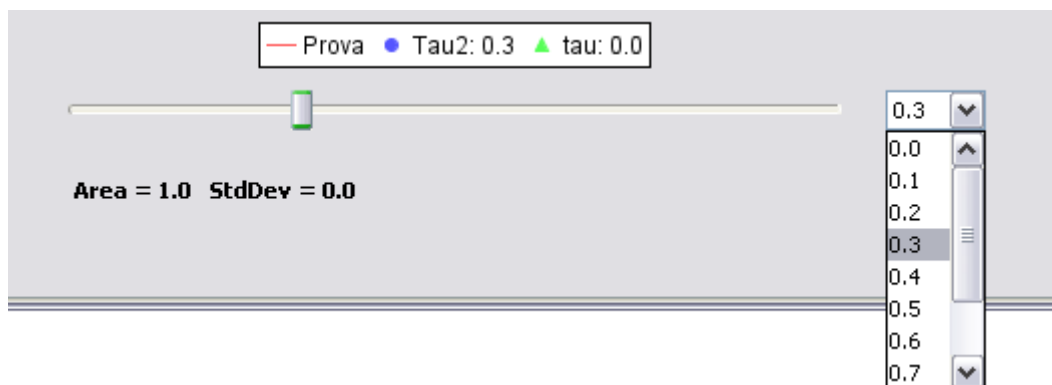


Figure 16. Tau selection

- tau: it's the value that was entered to be the reuse's tau when creating the configuration file. It's displayed in the chart with a green triangle.
- Tau2: it's a tau among all the tau used for generating the ROC curve, and is selected with the combo box or the slider shown in figure 16. It's displayed in the chart with a blue spot.

To have all of the plots in a global plot, click the “Add to global chart” in the experiments you want to select. Then, in the “Global chart” tab, you will have all of the plots together, as in the following Figure:

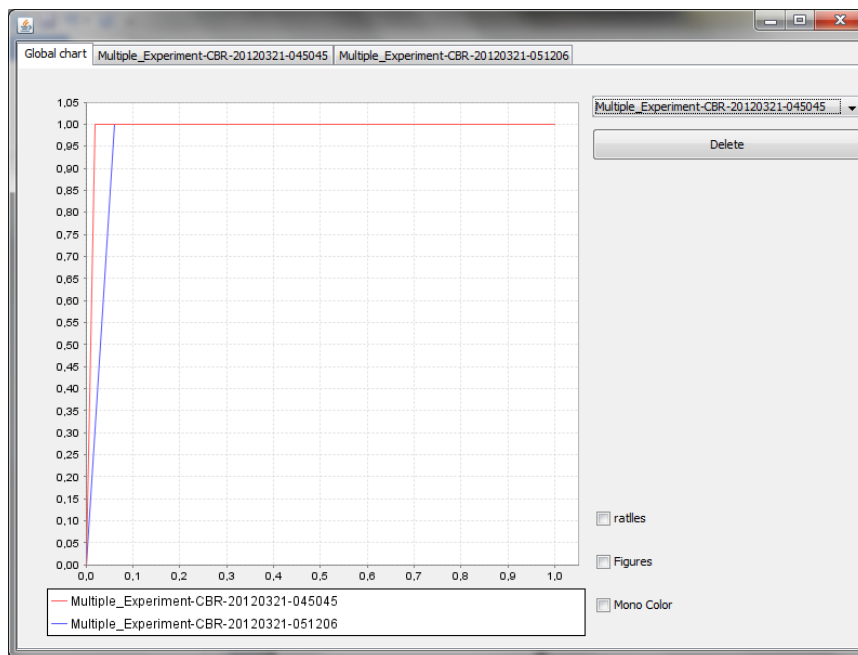


Figure 17. Global chart

You can play with the different options provided in this window in order to have the plots in the style that you want.