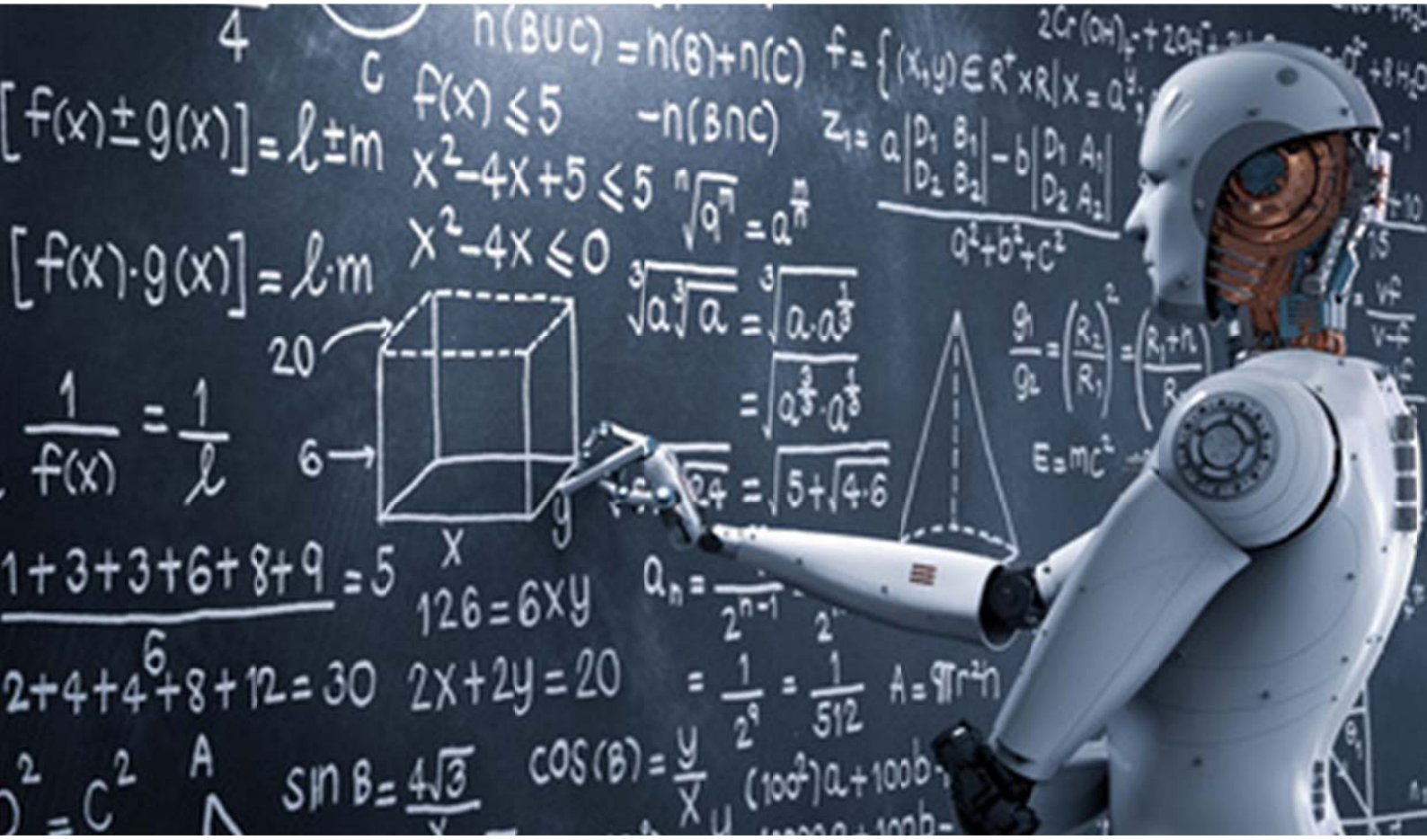
intellitech
INTELLIGENT TECHNOLOGIES

xtractis®
BY INTELLITECH

# White Paper

## xtractis® Approach

for Robust Predictive Modelling
and Multi-Objective Optimization
of Complex Processes




Authors : Prof. Z. Zalila, J. Cuquemelle, C. Penet, A. Chikh, B. Lorentz, D. Deschamps,
C. Assemat, S. Marbach, G. Gueydan, C. Leroux
xtractis@intellitech.fr



First publication: 02/2008
Revision : v2.8 – 05/2014



**INTELLITECH** – 14 rue du Fonds Pernant – Technopolis III – 60200 Compiègne, FRANCE
Tel: +33 (0)3 44 23 48 – Web: www.intellitech.ai – www.xtractis.ai
A Private Limited Company with 206 400 Euros capital – R.C.S. Compiègne B 419 510 128 – SIRET 419 510 128 00020 – Code APE 73 1Z

# Warning

The entirety of this document is protected by copyright. Reproduction rights are reserved. Quotations from any part of the document must necessarily include the following reference:

**Zalila, Z., Cuquemelle, J., Penet, C., Chikh, A., Lorentz, B., Deschamps, D., Assemat, C., Marbach. S., Gueydan, G., Leroux, C.** (2008-2014) x**tractis**® Approach for Robust Predictive Modelling and Multi-Objective Optimization of Complex Processes, White Paper, v2.8, **intelli**tech, Compiègne, France, May 2014, 12p, http://xtractis.ai/en/a-unique-mathematical-and-algorithmic-approach/


*Copyright © **intelli**tech [intelligent technologies] 2002-2014. All rights reserved.*

# Contents





*Copyright © **intelli**tech [intelligent technologies] 2002-2014. All rights reserved.*

# 1. INTRODUCTION

Fuzzy inference systems allow to easily and intuitively model any decision making process, whether it represents a physical measurement, a mathematical computation or a human evaluation. The decision making process is modelled as a deterministic relationship between inputs – available knowledge about the situation – and an output – the decision to be taken – implicitly expressed by linguistic rules.

The classical fuzzy modelling method, derived from Artificial Intelligence, builds the fuzzy decision rules thanks to a linguistic expression of the available knowledge about the studied process. However, in many situations (subjective evaluation, high complexity of the decision process), it is impossible to *a priori* make explicit the linguistic rules that explain the process.

In those cases, the x**tractis**® approach proposes to automatically extract the rules through an automatic learning process, performed on a database composed of different reference situations. This approach is similar to neural network training with a learning base, with the advantages of the fuzzy model paradigm over neural networks.

Databases used by x**tractis**® are divided into three different types: objective data (**O**) – analysis results or physico-chemical measurements, demographic or financial data … –, subjective data (**S**) – consumer liking / preference – and subjective objectivised data (**SO**) given by human experts – sensory panels – that represent the most objective evaluation of subjective attributes.

The classical issues arising in any learning process (overfitting[1], noisy data, limited amount of learning points), especially for a high dimensionality decision space, require implementing methods to analyse the quality of the learning base, supervise the learning process, and estimate the generalization capacity of the model to new unknown points.

# 2. FEATURES AND PROPERTIES OF THE FUZZY THEORY

## 2.1 Generalities

The fuzzy theory is a set of mathematical concepts allowing the representation and management of fuzzy data that is uncertain, imprecise or subjective. The fuzzy theory consists in a generalization of several classical theories:

Sets theory ➡ Fuzzy sets theory

Binary logic ➡ Fuzzy logic

Theory of numbers and intervals ➡ Fuzzy arithmetic

[1] See §I.4 Generalization capacity of generated models


*Copyright © **intelli**tech [intelligent technologies] 2002-2014. All rights reserved.*

Probability theory ➡ Possibility theory

From a formal point of view, the fuzzy theory can be considered as an interface between qualitative data or symbolic concepts, and quantitative values. Thus, this natural capacity of the fuzzy theory to handle heterogeneous data (quantitative / qualitative, precise / vague, objective / subjective) makes it particularly suitable to handle real life problems. Actually, a fuzzy inference is an approximate reasoning process, mimicking a human reasoning making a decision according to qualitative concepts or imprecise information.

Moreover, in cases of incomplete, sparse or low quality data, the possibility theory offers a better representation of impreciseness and uncertainty than the probability theory.

### 2.2 Interpretability

Unlike a neural network model[2], or a mathematical function that can be difficult to interpret in a qualitative point of view, a fuzzy system is a collection of linguistic rules[3] that can be intuitively interpreted. For example:

Rule 1: "**If** [CLime] is **high** **and** [Csugar] is **low** **Then** [Acidity] is **strong**".

Rule 2: "**If** [CLime] is **low** **and** [Csugar] is **high** **Then** [Acidity] is **mild**".

**low** and **high** are fuzzy sets on a quantitative scale characterising the fact that a concentration is considered as low or high

**strong** and **mild** are characterizations, on a sensory scale, of the intensity of the acidity as if evaluated by an assessor.

This fuzzy rules collection implicitly defines a multidimensional non-linear function, with the benefit of being far more easily interpreted than its functional form. This ability to hold a sufficient interpretability even for multidimensional models, usually allows us to model a process with the only use of its original input variables, without performing an *a priori* dimensionality reduction[4]. This allows us to keep a maximal connection between the model and the real process.

### 2.3 Locality

Each fuzzy rule defines an expertise domain that can be considered as a local model. For each domain, the good decision is the conclusion of the rule. A fuzzy system is a collection of rules that combines several expertise domains on the decision space. Thus, modifying a fuzzy system's parameter is akin to modify a local behaviour, unlike a neural network or a

---

[2] A neural network is a "black box" non-linear model in which the influence of a given parameter or the whole decision making process is very difficult to interpret.

[3] According to the model "*If Premise Then Conclusion*" derived from the cognitive psychology.

[4] PCA (Principal Component Analysis) is often used to reduce the dimensionality of a problem in order to simplify its modelling. The modelling is then simpler, but at the expense of losing the interpretability of the model, as it is often difficult to explicit a link between a principal component and the real process. Furthermore, even if the dimensionality of the problem is reduced, the knowledge of all the original variables is necessary to compute the principal components of the points that must be predicted. A PLS (Partial Least Square) model has the same drawbacks.


*Copyright © **intelli**tech [intelligent technologies] 2002-2014. All rights reserved.*

polynomial model for which tuning a parameter will have a global influence on the response of the model.

This property allows stabilizing the learning process, as it avoids having antagonist influences of very different learning points on the same model parameters.

### 2.4 Traceability

The decision taken by a fuzzy system is traceable, i.e. every decision can be explained by the firing degree of each rule that will lead to the conclusion of the system.

### 2.5 Universality

A fuzzy system is a universal non-linear function approximator[5], which guaranties the existence of an ad hoc fuzzy system to model any given decision process.

## 3. DATABASES QUALITY AND REPRESENTATIVENESS

The quality of the available data is of utmost importance. In fact, what will be modelled during the learning process is the knowledge that is embedded in the data, which must thus be representative of the process to be modelled.

It is then important, in the case of very imprecise data (human sensors data, **SO** and **S**, or data from very noisy sensors) to have several repetitions of the same measure for each point of the learning base. This allows us to perform an analysis and a filtering of absurd or atypical points.

A property of fuzzy systems that is interesting to mention, is their ability to perform a prediction even with missing inputs. The result of this partial inference will be the most possible[6] output value assuming the ignorance of one or several inputs. This property allows us to get rid of absurd points or take account of missing information, without the need to replace these missing values by estimations[7].

When possible, we usually use non-aggregated data (individual assessors' estimations and repetitions, noisy sensors repetition) for the learning process, which will constraint the model to take this variability into account. Thus, the learning process is not performed on aggregated data, which improves the robustness of the generated models. In fact, these models must be accurate and robust on all individual learning points and not only on their aggregation.

---

[5] This property is also true for neural networks

[6] According to the Possibility theory

[7] It is common to replace a missing value by an estimate, computed from the definition range of the variables or from the values taken by other learning points on this variable (method known as imputation). We consider that this approach may introduce false information on the database.


*Copyright © **intelli**tech [intelligent technologies] 2002-2014. All rights reserved.*

# 4. GENERALIZATION CAPACITY OF GENERATED MODELS

## 4.1 Overfitting

Every automatic learning or training process is prone to a risk of overfitting (or overtraining). It is in fact really easy to obtain a model that is able to exactly predict the points of a learning database, but that has no generalization ability to unknown points. Thus, a fuzzy system composed by as many rules as learning points could easily give exact predictions (the conclusion of the rule would be the known output value of each point), but a prediction on any other points would have no sense. The same behaviour happens when trying to fit a statistical model with too many parameters on a small data sample.

It is then required to implement several means of minimizing the risk of overfitting during the learning process (regulation methods) and to check the generalization capacity of the generated models (validation methods).

## 4.2 Regulation methods

The first method we use is a noise injection on the input variables. This prevents a rule to become specific to a learning point during the learning process. From a more analytic point of view, this technique allows smoothing the non-linear function implicitly defined by the fuzzy model.

We have also developed a fuzzy-specific method to reduce the complexity of a model without altering its accuracy. This is achieved by comparing and merging similar fuzzy subsets to use one premise for several rules, and by deleting uninformative classes. By reducing the complexity of a fuzzy system (i.e. reducing the number of descriptive parameters of the model), we augment its capacity of generalization and at the same time its interpretability.

Lastly, for a given modelling problem, we always perform several fuzzy system generations with different number of rules[8] and select the most relevant variables to predict the output. Among these different models, the most compact (low variables and rules count) will be chosen. As explained earlier, a simpler structure is less prone to overfitting and more interpretable than a complex one. The models generally selected have typically between 2 and 15 rules.

These regulation methods allow a reduction of the overfitting risk, but they do not give any information about the actual robustness of the generated models. It is thus necessary to check their generalization capacity with the help of validation techniques.

[8] A rule represents a partial knowledge of the process to be modelled. With a low number of rules, each rule represents a rather generic knowledge of the process, whereas with a higher number of rules, the knowledge modelled by each rule is more specific. Typically, a model with a high number of rules is more precise but may have a lower capacity of generalization, whereas conversely with a lower number of rules the robustness of the model will be higher at the expense of a somewhat lower accuracy.


*Copyright © intellitech [intelligent technologies] 2002-2014. All rights reserved.*

## 4.3 Validation methods

***DEFINITION OF A VALIDATION SAMPLE***

The most used validation method consists in partitioning the available data in a learning sample and a validation sample, usually keeping 2/3 of the data for learning and 1/3 for validation. Models are generated using the learning sample only, and are then tested on the validation sample to give an estimate of their capacity of generalization. This technique works well when the available amount of data is large regarding the dimensionality and the complexity of the process to be modelled. When this condition is not verified, the robustness estimation is strongly dependant on the way the data have been partitioned, which questions the validity of the estimation. This first limitation is illustrated by Figure 1.

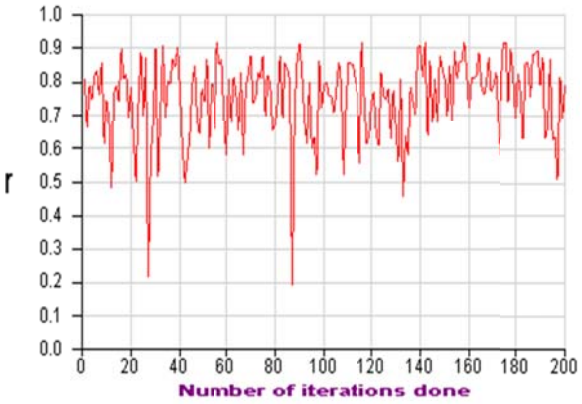



*Figure 1 - Variability due to the choice of a learning sample*

This plot shows the outcome of 200 repetitions of the same experiment on the same database: random draw of 30% of the database as validation sample, model generation (with a fixed learning strategy[9]) and prediction of the validation points by the generated model. Figure 1 shows the correlation between actual and predicted values on the validation sample of each experiment.

Using a validation sample to assess the validity of a model is akin to randomly choose one single experiment among the 200 represented here (with a correlation ranging from 0.2 to 0.92) to draw conclusions, and thus mask the strong variability of this estimator.

Moreover, the validation data is not used during the modelling process, what may lead to the impossibility of generating a model representative of the process in the case of very few available data.

[9] Using a fixed learning strategy means using the same learning algorithm with the same learning parameters, as well as using a fixed structure for the model that is extracted, *i.e.* same input variables and same number of rules.


*Copyright © **intelli**tech [intelligent technologies] 2002-2014. All rights reserved.*

Practically (especially for sensory evaluation problems), there is seldom enough available data for this method to be valid. Then, we use methods known as Cross Validation (CV) techniques that allow getting rid of the sensibility to a particular data split, and generating models using the entire dataset while still providing an estimate of the generalization capacity of the models. This estimate is computed by realizing several different splits of the initial dataset into one learning sub-sample and one validation sub-sample. For each split, a test model is generated[10] with the help of the learning dataset only, and its predictions on the validation set are compared to the actual values to compute a performance index. The aggregation of these performance indexes computed on several different splits gives an estimate of the performance index of the model generated with the whole database.

### *LEAVE-ONE-OUT (LOO)*

This is the most straightforward validation method. For a given initial dataset of N points, N splits will be generated, each composed by a learning sample of N-1 points and a single validation point. For each split, a model is generated with the N-1 learning points and tested on the validation point. The resulting scatter of points represents, for each point of the initial base, the prediction of a model generated without this point. This gives an estimate of the capacity of generalization of the model computed on the whole available dataset.

In the case of a non-deterministic learning algorithm, the whole process can be repeated several times to take into account the learning algorithm uncertainty in the robustness estimation.

### *MONTE-CARLO P-CROSS VALIDATION*

This method allows exploring more thoroughly the influence of the distribution of the available sample by using splits with more validation points. The LOO estimator is indeed optimistic as only one point is not used for the learning process. Moreover, the number of different splits is limited which may lead to a large variance of this estimator.

To overcome these two limitations, it is useful to use splits of the initial dataset with several validation points. The ideal estimator, known as Exhaustive P-Cross Validation uses all the possible splits with P validation points to generate test models with N-P learning points and compute their generalization error with these P points. The generalization error of all the test models is then aggregated to give an estimation of the robustness of the model generated with the complete database. It is practically impossible to generate a test model for each possible split of the database (there are $C_N^P$ possible splits), but a statistical approximation of this exhaustive estimation is available thanks to the Monte Carlo method: instead of generating a test model on each possible split, the robustness estimation will be

[10] In order to be relevant, the test models generated on each data partition for the robustness estimation must have the same structure than the model generated with the whole learning database. For a fuzzy system, this means that the same variables must be used as inputs, and that the number of rules must be the same.


*Copyright © intellitech [intelligent technologies] 2002-2014. All rights reserved.*

computed by aggregating the generalization error of several test models generated with random drawn splits. Thus, after a certain number of draws (typically between 150 and 500), the aggregation of the generalization error of the test models converges towards the estimation of the robustness of the global model.

A specificity of our approach concerning the use of cross validation techniques is to keep the predictions of each test model to build a scatter of points of every prediction and to compute the performance indexes on the whole predictions, instead of aggregating the performance indexes of each test model. This latest aggregation allows keeping more information for the evaluation of the robustness of the model generated with the whole database. Figure 2 shows the convergence of this method: for each number of iterations, the plotted correlation is the one of the scatter of points aggregating all predictions computed before.

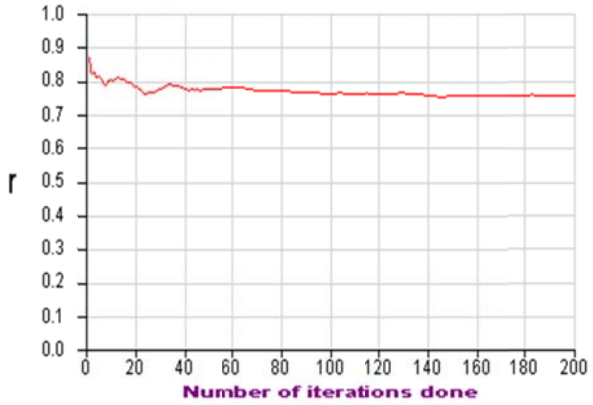



*Figure 2 - Convergence of the Monte Carlo CV estimate*

All these cross validation techniques (including the choice of the number P of validation points) have different characteristics of variance and bias on the estimate they give. For this reason, we always compute several robustness indexes in order to select the models that have satisfactory values on several different indexes.

# 5. EXPLOITING XTRACTIS® MODELS

## 5.1 Interpretation of the model

Each generated models is, as mentioned before, a collection of linguistic rules that have been induced from the examples held in the database. This allows an expert of the field to compare the knowledge extracted from the data with his/her own knowledge on the processes to model, to validate or discuss the relevance of the generated models.

Furthermore, this technique is able to linguistically synthesise the behaviour of a completely unknown process.


*Copyright © **intelli**tech [intelligent technologies] 2002-2014. All rights reserved.*

During the generation process, we perform a selection of the most relevant variables to predict the output. This variable selection takes into account the synergies that may exist between groups of variables[11]. For a given model, the individual influence of each variable on the quality of predictions given by the model is evaluated.

### 5.2 Topographical exploration

Fuzzy models implicitly define multidimensional non-linear functions. x**tractis**® implements several tools to represent graphics of cross-sections of the analytical decision surface[12] in 2D or 3D.

Another surface, named the mapping surface represents the influence areas of each rule of the fuzzy system, along with non-covered areas (usually because of a lack of data in these areas). This helps the modeller to target interesting zones for a future research: providing new learning points in non-mapped areas will extend the prediction domain of the model.

### 5.3 Direct prediction

The generated models can be used to make predictions (of a sensory profile, of a consumer liking, of the early diagnosis of a disease…) to evaluate the influence of a new situation (product formulation, patient…).

Furthermore, the x**tractis**® inference engine is able to give real-time predictions (online monitoring) or predictions on very large databases (CRM decisions …) thanks to an efficient CPU/GPU multi-threading implementation.

### 5.4 Optimization by model inversion

This functionality proposes to search among the possible input values, which ones will allow to obtain a desired output, for example a maximum consumer liking, a target sensory profile, the minimal toxicity, the maximal efficiency…

The desired output can be described with crisp information (scalar values or intervals) or with fuzzy information which is equivalent to giving flexible research constraints. x**tractis**® is thus able to give satisfactory solutions, even when an exact solution to a crisp request cannot be found.

The optimal solution search can also take into account constraints (crisp or fuzzy) on the inputs of the model. For example: how to maximize [Acidity] assuming that [CLime] should be lower than 8%?

---

[11] It happens frequently that some variables are irrelevant when used alone (low individual influence), whereas when used together the group of variables gives a lot of useful information on the process.

[12] As the generated models are multidimensional, it is impossible to represent their decision surface. It is however possible to display cross sections around user-given points.


*Copyright © **intelli**tech [intelligent technologies] 2002-2014. All rights reserved.*

It is also possible to search for optimal solutions satisfying a multi-criteria request by simultaneously inverting multiple x**tractis**® non-linear fuzzy models as well as analytical models. For example: "Which formulation will maximize [Acidity] and reach a value of about 4 for [Sweet] given the fact that [CSugar] must equal 2 x [Clime]".

The optimal solutions can be displayed on cross sections of the decision surface, with the possibility to display the areas holding all the satisfactory solutions for the given cross section.

# 6. APPLICATION FIELDS

## 6.1 A universal approach

The modelling by extracting the knowledge implicitly embedded in a database is universal and can be applied to any problem for which an input/output database of a process is available. No specific knowledge concerning the problem is needed to be able to generate efficient and robust models.

Thanks to its high performance, the x**tractis**® approach has been successfully used on numerous sensory engineering problems (modelling of sensory profile, modelling of liking for different consumer segments), evaluation problems (assessment of financial assets, estimation of real estate assets, classification of flaws in steel tubes/plates, assessment of toxicity / efficiency / biodegradability of molecules, prediction of Parkinson's disease score…), risk analysis (investment risk, solvency risk, fraud risk…), descriptive analysis (modelling of urban density, modelling of chaotic time series…), or diagnostic (detection and characterisation of damage in composite materials, early diagnosis of cancers…).

x**tractis**® may also be useful in any domain for which experts need help to exploit the knowledge embedded in complex databases, for example in oil exploration (seismic analysis) and pharmaceutical research (virtual screening).

According to its policy of excellence and innovation, **intelli**tech constantly promotes research to design and develop new knowledge extraction algorithms and methods that would be more suited to specific application domains, characterised by a very high complexity that may involve millions of data points or thousands of dimensions.

## 6.2 Interface between model and expert

The interpretability of the generated models allows the modelling team to compare and discuss the modelling results with experts of the application field. The structure of the models as a collection of fuzzy rules makes this task far easier than with other types of models. This leads in most of case to a confirmation of the generated model relevance and sometimes gives clues to improve the knowledge of the considered application field.


*Copyright © **intelli**tech [intelligent technologies] 2002-2014. All rights reserved.*