



COST ACTION IC0702

Combining Soft Computing Techniques and Statistical Methods to Improve Data Analysis Solutions

Christian Borgelt, Raul del Coso
European Centre for Soft Computing (Asturias, SPAIN)

*Kick-off meeting
Brussels 31 March 2008*

Soft Computing, as an engineering science, and Statistics, as a classical branch of mathematics, emphasize different aspects of data analysis:

Soft Computing

focuses on quickly obtaining working solutions that meet the needs arising in applications.

Statistics

focuses on establishing objective conclusions by rigorously analyzing all possible situations.



Soft Computing: Strengths

Soft Computing

- designs intelligent systems that are tolerant to imprecision, uncertainty, partial truth and approximation,
- tries to achieve tractable, comprehensible, robust, and low cost solutions to real-world problems,
- is very open to new and unconventional approaches,
- provides the flexibility and swiftness that are needed to **generate models** in industrial applications.



Soft Computing: Weaknesses

Soft Computing Methods

- often lack sound mathematical foundations,
- rarely make their underlying assumptions explicit, thus impeding a reliable transfer to new applications.

Soft Computing Models

- are seldom checked rigorously and monitored w.r.t. performance and robustness,
- can rarely be generalized or easily transferred.

Statistics

- has sound mathematical foundations,
- makes all underlying assumptions explicit, thus easing generalization and solution transfer,
- guarantees performance by analyzing the behavior in all possible situations that could arise,
- provides the mathematical means to rigorously **validate models** and establish objective conclusions.

Statistical Methods

- **tend to focus on models, the mathematical properties of which are easy to analyze,**
- **constrain the set of eligible models, thus perhaps ruling out the most suitable or most promising ones.**

Statistical Models

- **can be difficult to understand and to apply for a non-mathematician, thus hindering plausibility checks.**

The existing interaction is very limited:

- **Descriptive and explanatory statistics are sometimes used in parallel with soft computing techniques.**
- **Mathematical and inductive statistics are applied only very scarcely, if they are used at all.**
- **The importance of the strengths of soft computing for applications is underestimated by statisticians.**
- **The numerous successful applications are not known.**



Vision: New Research Areas

Prototype: **Support Vector Machines**

- resulted from merging statistical learning theory and (artificial) neural networks,
- are among the most intensely researched topics,
- provide an enticing vision of the potential of joining statistical and soft computing methods.

Candidates for achieving a similar development are **fuzzy systems** and **evolutionary algorithms**.

- **Statistical Validation and Monitoring of Soft Computing Models**
- **Model Selection and Validation for Neural Networks**
- **Evolutionary Algorithms as Estimators**
- **Estimation of Distribution Algorithms**
- **Statistics with Fuzzy Data**
- **Psychological versus Statistical Complexity**



Statistical Validation and Monitoring

- **Current focus is on very simple statistical techniques.**
- **Objective, statistical validation is desirable.**
- **Specific topics/areas:**
 - **Cope with overfitting by considering data generation processes governed by Soft Computing models.**
 - **Study resampling for assessment and comparison.**
 - **Statistical monitoring of deployed models (system failure, change points, loss of fit etc.)**
 - **Change mining with Soft Computing methods.**



Model Selection/Validation for NNs

- **Only part of the power and flexibility of NNs is used in combination with statistical techniques.**
- **Restrictions of Support Vector Machines:**
 - **Kernel functions are fixed to data points.**
 - **Parameters are not adapted during training.**
- **NN Training is usually much more flexible.**
- **Statistical methods for designing and training NNs.**
 - **Example: comparison of network structures in terms of statistical hypothesis tests.**



Evolutionary Algos. as Estimators



- **Evolutionary algorithms can be seen as iterative estimation procedures (for the mode of the fitness function).**
- **Convergence corresponds to estimator consistency.**
- **Other statistical properties can also be transferred (unbiasedness, efficiency, sufficiency etc.)**
- **Some seminal work exists, but is limited:**
 - **small populations,**
 - **special fitness functions,**
 - **very restricted genetic operations etc.**



Estimation of Distribution Algorithms

- **EDAs are a newer type of evolutionary algorithms:**
 - **Current population is used to estimate a distribution describing where good solutions may be located.**
 - **Next generation is obtained by random sampling.**
- **Clear connection to statistics due to the use of estimation and sampling procedures.**
- **(Deeper) investigation of statistical properties needed:**
 - **expected number of populations**
 - **expected solution quality**

Statistics with Fuzzy Data

- **In many applications data is derived from a human perception or estimation of some quantity, because**
 - **an objective measurement device is not available,**
 - **a precise physical measurement would be too costly.**
- **Classical statistics has few other possibilities than treating such data as if it were crisp (usually infeasible: families of distribution functions).**
- **Specialized statistical techniques are needed, geared towards the fuzzy character of such data.**



Psychological vs. Statistical Complexity



- **Statistical complexity of a model:**
 - number of free parameters
 - description length of the model
- **Psychological simplicity of a model:**
 - How intuitive and easy to understand is a model?
(for a human, who is not a data analysis expert)
 - Focus is on qualitative characteristics.
 - Essential in (industrial) applications.
- **Local models versus global model.**

- **MC meetings every six months**
- **Working Group Meetings to collect and distribute information and to explore new opportunities**
- **MC & WG online meeting every 2 months for day-to-day management**
- **STSM (Short term scientific missions) coordinator**
- **Administrative coordinator**

- **Workshops and Seminars** with a broad range of experts to address specific areas
- Tailored and structured **scientific exchanges** between research groups
- Large annual **Conferences**
- Two **training schools** to educate a new generation of multidisciplinary researchers

Deliverables

- > 60 **short term scientific missions** (at least one per signatory country per year).
- 4 large (>50 participants) and **open conferences** (one per year)
- 2 **training schools**
- 16 **workshops and seminars**
- > 10 **research proposals** that will generate innovative soft computing techniques.
- > 10 **reports** summarising the results of the Action (at least two per Working Group and one per year for the whole Action)
- Basic scientific knowledge and improved technical developments by opening new areas of exploration. This will lead to a number of **publications** and presentations in conferences.

Dissemination plan

- **Workshops, seminars and conferences** organised by the WGs and the MC
 - **Talks** in other national and international conferences
 - **Papers** in peer-reviewed scientific and technical journals
 - **COST Action reports** (Progress, analysis and reviews)
 - **Online tools:**
 - Webpage (Action activities, related events, new developments in the area, open positions, successful industrial applications, etc.
 - Monthly newsletter

Year 1: April 2008 – March 2009

- **Starting Conference** (overlap with organised conference)
- **3 workshops**, one per WG
- **10 short term scientific missions** (one per signing country?)
- **Plan activities for Year 2** (conference, workshops, seminars, training school, scientific missions, research proposals, etc.)

Thank you!

