Computational Intelligence: Methods and Applications

Lecture 37
Summary & review

Włodzisław Duch
SCE, NTU, Singapore
Google: Duch

Overview

First I was talking about **General Introduction** to computational intelligence, problems and inspirations how to solve them.

Vast fields, it is impossible to cover all of them, for example optimization inspired by evolution, immunological system or swarm behaviors was left out, fuzzy ideas and neural inspirations were only mentioned.

Much of this course was just showing you the basic techniques and only mentioning inspirations of biological or other nature that help to solve problems, for which no effective computational algorithms exist.

All new CI techniques should finally be understood in light of probabilistic foundations, and compared to good standard algorithms in the field.

Ex: there are many global optimization methods that are used for parameter optimization, but new methods based on evolutionary algorithms, swarm, ant or other methods are almost never compared to standard mathematical algorithms ... makes you suspicious;
either people do not know the basics or their results are poor.

Rough sketch

- General introduction.
- Exploratory data analysis – visualization, direct and linear projections.
- Bayesian theory of adaptive systems: decisions, risks, bias-variance and model selection, evaluation of results.
- Decision trees in WEKA and Ghostminer.
- Discrimination algorithms, linear, kernel-based SVM, kernel PCA.
- Density estimation, expectation maximization, parametric and non-parametric probability density.
- Basis function networks, RBF, separable functions, fuzzy rules, combinatorial reasoning.
- Nearest neighbor methods.
- Information theory, selection of information.
- Metalearning, committees and boosting techniques.

Visualization for data understanding

We have spent some time on **Exploratory Data Analysis.**

Feature Space model has been introduced, used in most lectures, with only similarity-based models that could deal not with the object description, but with similarity between the objects.

Basic probability concept were introduced using histograms, and a few direct visualization methods followed:

- multidimensional histograms
- scatterplots, or 2D projections
- multiple scatterplots showing all pairs of data
- Grand Tour, searching for good projection angle
- star plots, and clusterized version of multiple scatterplots & starplots
- parallel coordinates
- Chernoff faces

Linear projections

First the effect of data pre-processing was considered:

- standardization of data makes it scale-free, Mahalanobis distances show relations that are invariant to any linear transformation.
- Covariance matrices measure variance of features and correlation (co-variance) between them; info about classes is not used.
- Eigenvectors of d-dimensional covariance matrix transform data to principal components that are decorrelated.
- First principal component corresponds to the largest eigenvalue of the covariance matrix, and thus to the direction of the largest variance.
- PCA gives interesting projections and may be used for dimensionality reduction but for some data strongly overlapping clusters are shown.
- FDA maximizes separation between classes, using supervised method to find best projection, usually showing better separation.

Non-linear mappings

- Independent Component Analysis (ICA) finds linear combinations making data components independent, and more general Projection Pursuit methods find more interesting views on the data.
- Self Organized Mapping (SOM) shows distance relations between clusters of data in topographically correct way on a mesh in 1, 2 or 3D.
- Growing Cell Structures creates new nodes and is able to represent the data more accurately.
- Multidimensional Scaling (MDS) tries to represent distances between the original objects in the high-dimensional space.
- PCA needs to diagonalize dxd matrix,
  MDS needs to minimize stress function with (n-1)n/2 parameters.
- Later Kernel PCA and maximization of mutual information were introduced as visualization methods.
- Many other visualization methods exist, this is an active research field.

Bayesian foundations

The best foundation for adaptive systems is afforded by Bayesian formulation, all learning methods may be seen as approximations.

Many ad-hoc methods have been invented but only those with good foundations will remain ...

Bayes MAP rule is used to compute posterior probabilities:
posterior is equal to the likelihood x prior / evidence.
Confusion matrices were introduced and risk of decision procedures formulated, with minimum risk classifiers giving best discrimination.

Bayesian approach requires density estimations.

Two Gaussian distributions in d-dim. lead to two approximations:

1) linear discriminants for identical covariance matrices, related to perceptrons and neural mappings;
2) nearest prototypes, if also prior probabilities are identical

Density estimations

On the density estimation side we have followed:

- Naive Bayes approach, assuming that all features are independent and thus the total density is a product of densities for single features.
- Basis set expansions for parametric density estimation, using RBF networks or FSM networks based on separable functions.
- Expectation maximization to find generative parametric data models.
- Non-parametric density estimations using histograms and Parzen windows.

Probability density appeared to be useful for:

- classification and approximation;
- generation of fuzzy rules;
- finding missing values;
- combinatorial reasoning for solving complex problems.

Other approximations

Approximation using linear discrimination with direct parameter evaluation led us to:

- linear DA, logistic DA, and linear SVM methods;
- kernel-based SVM that attempts linear discrimination in high-d spaces, and kernel-based methods;
- decision trees based on very simple discrimination but powerful idea of divide-and-conquer, or hierarchical partitioning of data.
- Nearest prototype approach leads to nearest neighbor methods, that may also be justified from the density estimation point of view.
- On the theory side model selection based on bias-variance decomposition provides understanding of what it takes to make a good model of appropriate complexity.

Remaining themes

Filter and wrapper approaches to select relevant information were defined, and information theory has proved to be very useful in feature selection for dimensionality reduction.

Discretization of continuous features that preserves information in the distribution of data and may be used to improve histograms was introduced; this is important to calculate information indices and density estimations accurately.

Finally meta-learning techniques were introduced to improve quality of results: committees, bootstrapping, bagging, boosting and stacking.

All these methods are very useful in practical applications.
Some methods are very popular and good software is easy to find, other methods are found only in papers and textbooks.
Much remains to be done, especially for complex domains, objects with structure, integration with reasoning, hierarchical data, but most new methods that people introduce vanishes without a trace ...

The end is near ...

We have covered a lot of ground but ...

- it is still rather a small portion of computational intelligence;
- many “natural computing” inspirations have not been explored;
- once the problem is understood good mathematical foundations may be found;
- this course has focused on practical methods, implemented in WEKA and GM packages; although there are more models there they do not differ significantly from those we have covered.
- Key idea: set the goal, learn from data, adapt – don’t solve directly.
- Many algorithms have only been sketched, and some pointers given; this allowed us to cover more topics.
  Those that will find these methods useful will learn them anyway, and those that will have no need for them will save some memory for other things ...

So …

What should one do to make progress in science,
and become rich and famous?

Should I add more reading material after each lecture?

More step-by-step exercises using software packages?

Perhaps lab/tutorial would be a good idea?

More assignments? Just joking …

More references to biological/cognitive systems?

What else would be helpful to make this course useful ?