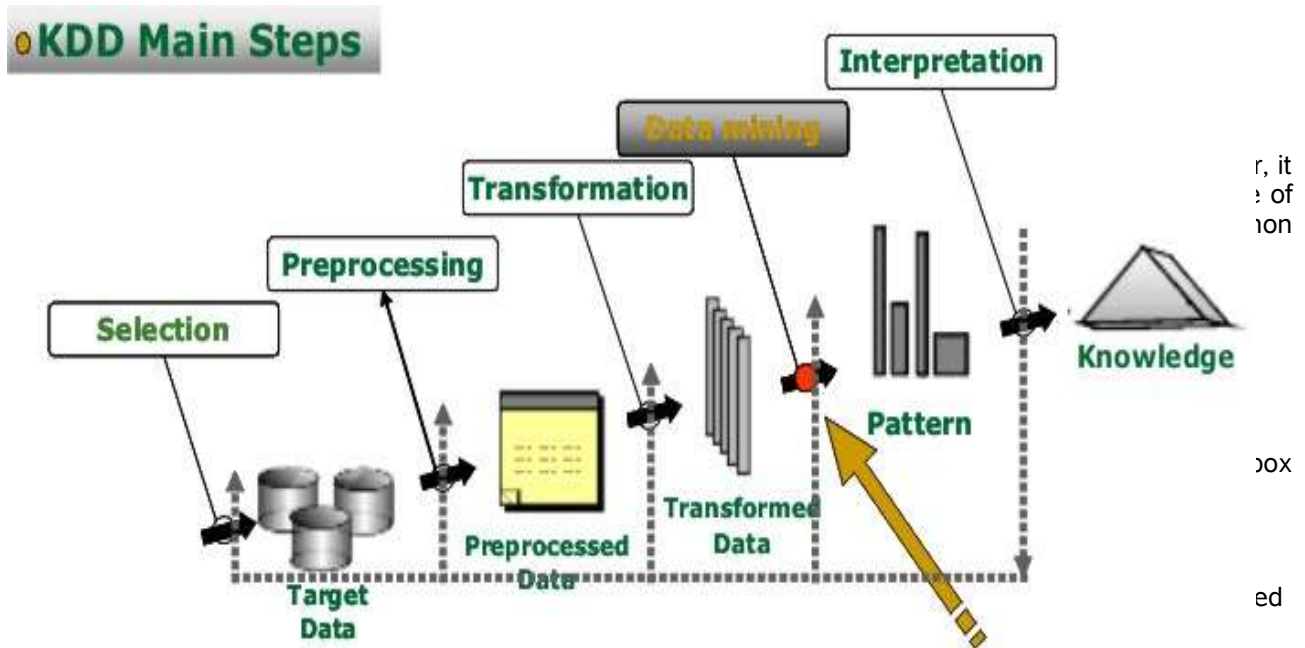


To the VO-Tech Science Committee members

1.0 What is Astro Neural

ASTRONEURAL was initially conceived as a package of MATLAB routines aimed at performing a variety of tasks in the field of "Knowledge Discovery in Databases" or KDD which is outlined in the following figure.



- iii) help on line. It is divided into two parts, the manual and a data mining tutorial. Both are half way through (a preliminary version will be ready for the September meeting)
- iv) integration with VISIVO. Done
- v) access to VO-table: done

In very general terms, most DM tasks can be reconduced to clustering applications, *id est* to the capability to group together objects (in our case strings of data) having similar properties or matching a specific pattern. For instance, we could be interested in finding which objects in a given astronomical database have photometric properties undistinguishable from those of QSO's, or in finding in a gene expression database which genes share the same temporal trends, or to identifying in any DB the groups of objects which share the same properties. All these examples can be considered as clustering problems.

In a very broad classification scheme, we may consider two families of techniques:

1. **Supervised techniques** work using "a priori" knowledge. In the case of Neural methods, this means that the network learns to classify or to recognize patterns using a priori knowledge provided by the user for a relatively large set of cases. This methods split the objects for which apriori knowledge is available in three subsets (namely training, validation and test sets) which are used, respectively, to train the network, to make sure of its generalization capabilities and to evaluate the errors.
2. **Unsupervised techniques** do not require a priori knowledge and exploit the statistical properties of the data. Errors are evaluated using the statistical properties of the data themselves and some a priori knowledge might be required "a posteriori" to understand the results (usually known as "labeling" of the data).

In general, supervised techniques are more accurate but can be applied only in a restricted number of cases (they depend on the availability of a large set of data for which the a priori knowledge is present) while unsupervised methods are less accurate but much more general. A proper choice of the method or a proper combination of different methods very often leads to surprisingly accurate results.

The main DM tasks which can be performed with [ASTRONEURAL](#) are:

1. Supervised clustering (MLP)
2. Unsupervised clustering (SOM, PPS)
3. Time series analysis (PCA, ICA)
4. Feature selection (PCA, ICA)

Artificial Neural Networks (NNs) were originally introduced as simplified models of the brain (processing nodes instead of neurons, multiple connections instead of dendrites and axons). Such a brain analogy, however, may mislead non experienced users who tend to overlook the fine details requested by the proper application of NNs, thus obtaining discouraging results.

In fact, as stressed for instance in [\[1\]](#), even though it is likely that the logic of the processing of the signals in both the brain and NNs is very similar, the scale is enormously different: the human brain consists of 10^{11} neurons each connected to many others with a complex topology for a minimum of 10^{14} synaptic connections, while even the most complex NN consists of a few hundred neurons and achieves a level of complexity at least 10 orders of magnitude smaller.

As a matter of fact, however, in most individual processes, the brain uses only a very small subset of its resources and therefore, in spite of the lack of complexity, some specific tasks may be emulated by NNs. In fact, they offer the advantage of being objective, relatively fast and, what is more relevant, not necessarily biased by some human limitations such as, for instance, the lack of capability in dealing with high dimensionality spaces. This last feature makes NNs appealing to all these fields where there is a need to organize and extract the relevant information from multiparametric spaces of high dimensionality. This is particularly true for astronomy and bioinformatics where the recent technological advances have produced a true explosion in both the quality and the amount of data available to any user [\[2\]](#).

Furthermore, due to the ongoing efforts for the implementation of the International Virtual Organizations (such as the International Virtual Observatory) , most of these data will become available to the community via the network (cf. [3], [4] and references therein). These huge and heterogeneous data sets will open possibilities which so far are just unthinkable, but it is already clear that their scientific exploitation will require the implementation of automatic tools

August 27 2005 -03

capable to perform a large fraction of the routine data reduction, data mining and data analysis work.

In its most general definition, a NN is a software which learns about a problem through relationship which are intrinsic to the data rather than through a set of predetermined rules. A NN is usually structured into an input layer of neurons, one or more hidden layers and one output layer. Neurons belonging to adjacent layers are usually fully connected and the various types and architectures are identified both by the different topologies adopted for the connections and by the choice of the activation function. The values of the functions associated to the connections are called "weights" and the whole game of NN's is in the fact that, in order for the network to yield appropriate outputs for given inputs, the weights must be set to a suitable combination of values. The way this is obtained leads to the first important difference among modes of operations, namely between "supervised" and "unsupervised" methods.

In supervised methods, in order to teach to the NN how to provide the correct output, the user needs to know the correct output value for a fair subsample of the input data. This set is further divided in other three subsets named, respectively, training, validation and test sets. The first one is used to fine tune the weights, the second one to check whether the network has achieved an acceptable generalization capability and, finally, the third subset is used to evaluate the performances.

In unsupervised methods, instead, the input data are clustered on the basis of their statistical properties only. Whether the obtained clusters are or are not significant to a specific problem and which meaning has to be attributed to a given cluster, is not obvious and requires an additional phase, the so called "labeling". The labeling requires that the user knows the characteristics of a small sample of input vectors (labeled set). It needs to be stressed that the labeled set needs to be much less conspicuous than the training, validation and test sets necessary to supervised methods.

A further distinction among different NN's can be based on the way the information propagates across the network: either feedforward (i.e. the information propagates only from the layer K to the layer $K + 1$), or recurrent (i.e. the information may propagate in loops).

The optimal choice of the architecture of the network and of its operating modes depends strongly on the intrinsic nature of the specific problem to be solved and, since no well defined recipe exists, the user has often to rely on a trial and error procedure. It has therefore to be stressed that, in order to be effective, all neural techniques require a lengthy procedure to be selected and optimized, and an extensive testing to evaluate their robustness against noise and inaccuracy of the input data. This also restricts their application to those data intensive or computational intensive problems where the work required by the implementation of a network may prove to be advantageous with respect to more traditional methods.

REFERENCES

- [1] Bailer-Jones, C.A.L., Gupta, R., Singh, H.P. (2001). Automated Data Analysis in Astronomy. Gupta et al. eds., [astro-ph/0102224](#)
- [2] Brunner J.R., Djorgovski S.G., Prince T., Szalay A.S.(2002). Massive data sets in Astronomy, Handbook of Massive Datasets, Ch. 27, J. Abello, P. Pardalos, and M. Resende Eds., Kluwer:Dordrecht
- [3] Brunner J.R., Djorgovski S.G., Szalay A.S. editors (2000), Proceed. of the Int. Workshop: Virtual observatories of the future, Astron. Soc. of the Pacific Conf. Series, n.225
- [4] Banday A.J., Zaroubi S., Bartelmann M. editors (2001), Proceed. of the Int. Workshop: Mining the sky, Springer:Heidelberg