Feature Selection Software to Improve Accuracy and Reduce Cost in Automated Recognition Systems

by Petr Somol

A specialized software library that helps identify the most informative measurements used in automated recognition systems has been made available to the public by researchers from the Institute of Information Theory and Automation of the Czech Academy of Sciences.

Pattern recognition systems are becoming increasingly important as the variety of scenarios in which they can help reveal important (but otherwise inaccessible) information grows. The increasing role of such systems is made possible partly by the ever-growing performance and proliferation of computers as well as by advances in theory. Pattern recognition is applicable in a vast variety of fields, including:

- Medicine (eg diagnostic systems, gene search)
- Finance (eg trend evaluation, credit scoring)
- Governmental planning (eg analysis of remote sensing data)
- Text processing (eg keyword extraction, document categorization)
- Security (eg face or fingerprint recognition)
- Military (eg target spotting)
- Industry (eg defect detection).

Traditionally, one of the key issues in pattern recognition system design has been the identification of a set of distinctive pattern properties, referred to as features (also known as variables or attributes), that can be used as a basis for pattern discrimination. Such features are selected from among a set of available measurements by mathematical tools which allow the designer to measure the discriminatory content of a feature set.

When building automatic decision systems the commonly followed practice is to first collect as many types of measurement as possible to ensure that no potentially useful information is omitted. Then a dimensionality reduction technique is usually applied to automatically identify which combination of measurements actually contains the maximum discriminatory information. In many situations only a fraction of the originally considered measurements is identified to contain all the useful information. (Note that even seemingly unimportant measurements may prove important in combination with others.) Restricting the final number of various types of measurement not only saves measurement acquisition cost in application phase, it may even help to improve recognition accuracy as it reduces the influence of noise and other unwanted "curse-of-dimensionality" effects.

The theoretical framework of dimensionality reduction now covers a vast range of approaches, some of which have been implemented as supplemental tools in several machine learning software packages (Weka or PRTools being among the better known ones). Nevertheless, many powerful feature selection techniques haven't been generally available so-far except in research papers.

The recently published Feature Selection Toolbox 3 (FST3) library written in C++ narrows several gaps in this area. It contains a selection of highly efficient feature selection algorithms as well as various supportive tools. It enables application of non-trivial subset search techniques even to commonly encountered very high-dimensional (and thus computationally expensive) problems, for instance, in text categorization or gene searches. To tackle the potentially high computational complexity of a feature selection task, the library provides workarounds of both a technical (parallelization) and conceptual nature (fast deterministic and/or non-deterministic techniques of gradual result improvement, etc.). Various anti-over-fitting techniques help to prevent degradation of final system recognition performance on new, previously unseen data.

The library has been developed within the Pattern Recognition Group at the Institute of Information Theory and Automation as part of long-term research activity in the field of statistical pattern recognition. The current 3rd installment of the software package follows an earlier development started in 1999. In addition to the current FST3, the dedicated web (see Figure 1) now also provides the former FST1 as a tool which is less powerful but more suitable for quick experimenting and educational purposes. In addition to FST software a wealth of related informational resources is provided, with the aim of creating a comprehensive portal of interest to any R&D practitioner dealing with pattern recognition problems.

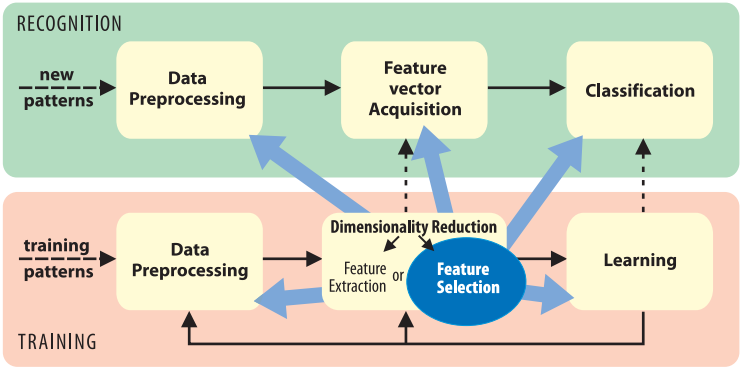

Figure 1: Feature Selection Toolbox 3 software library provides a selection of advanced tools focused primarily on solving the feature selection form of the dimensionality reduction problem, and also addressing and interacting with all other stages of the machine learning and recognition process.

The work has been supported by grants from the Czech Ministry of Education No. 1M0572 DAR and No. 2C06019 ZIMOLEZ.

Link:
http://fst.utia.cz

Please contact:
Petr Somol, CRCIM (UTIA), Czech Republic
Tel: +420 2 6605 2205
E-mail: somol@utia.cas.cz