Information Analysis using Softcomputing

The Applications to Character Recognition, Meteorological Prediction, and Bioinformatics Problems

Anto Satriyo Nugroho
asnugroho@ieee.org

Doctoral Dissertation Grad.School of Engineering, Dept. of Electrical & Computer Engineering, Nagoya Institute of Technology, Japan, January 2003

1. Softcomputing and Its Practical Applications

The rapid progress in computer hardware industry has become a great contribution to the progress of research in various fields. Many computational problems with complex calculations can be solved within a reasonable time, opening many inventions in a broad spectrum. Part of these problems have clearly specified formulas and procedures to obtain the solutions. The algorithms can be described very clear, from one to the next steps, towards the final solution. In such cases, computer will give very accurate result within very short computational time. It is no doubt to conclude that in this sense, the performance of computer is much superior to that of human brain.

A solution of a computational task, however, can not always be stated clearly. Many problems in real-life domain are always accompanied by imprecision or uncertainty factors. Sometimes the information is far from complete, but a precise decision is required. In such situation, human can make a correct decision, while computer requires complex calculation to make a mathematical model of the problem. One example of problems with such properties is handwriting character recognition. According to the conventional methods described in prior, we have to collect almost unlimited amount of representative handwriting samples, and derive a certain formula or rule to make a decision to which character that a handwriting pattern should be classified. However, this kind of effort is not practical and the formula will not tolerate to any deviation might occur in the handwriting pattern of different persons. The character recognition system should also tolerate the imprecision and uncertainty as common characteristics of the handwriting characters. The solution to problems with these properties is a family of algorithms named *softcomputing*.

Softcomputing is one of the most popular field in the computer science with very long and attractive history. Many descriptions have been addressed to describe the characteristics of this field. L. A. Zadeh, a pioneer in this field, provided a definition as follows:

> *Soft computing is an emerging approach to computing which parallels the remarkable ability of the human mind to reason and learn in environment of uncertainty and imprecision.*

In the homepage of BISC (Berkeley Initiative Soft Computing), Zadeh explains softcomputing as follows:

> Soft computing differs from conventional (hard) computing in that, unlike hard computing, it is tolerant of imprecision, uncertainty and partial truth. In effect, the role model for soft computing is the human mind. The guiding principle of soft computing is: Exploit the tolerance for imprecision, uncertainty and partial truth to achieve tractability, robustness and low solution cost. The basic ideas underlying soft computing in its current incarnation have links to many earlier influences, among them my 1965 paper on fuzzy sets; the 1973 paper on the analysis of complex systems and decision processes; and the 1979 report (1981 paper) on possibility theory and soft data analysis. The inclusion of neural network theory in soft computing came at a later point. At this juncture, the principal constituents of soft computing (SC) are fuzzy logic (FL), neural network theory (NN) and probabilistic reasoning (PR), with the latter subsuming

belief networks, genetic algorithms, chaos theory and parts of learning theory. What is important to note is that SC is not a melange of FL, NN and PR. Rather, it is a partnership in which each of the partners contributes a distinct methodology for addressing problems in its domain. In this perspective, the principal contributions of FL, NN and PR are complementary rather than competitive.

We can simply summarize the paradigms in softcomputing as follows:

1. **Fuzzy Logic**
   Fuzzy Logic is knowledge representation constructed by if-then rules.
2. **Artificial Neural Networks**
   An artifical neural network is an information processing system that has certain performance characteristics in common with biological neural networks.
3. **Genetic Algorithm (GA)**
   Genetic algorithm is a population-based search method, as a model of machine learning that derives its behavior from a metaphor of the processes of evolution in the nature.

Since human mind is known to be tolerant to imprecision, uncertainty and partial truth, it has become a great inspiration to the development of methods in sofcomputing.

The applications of softcomputing are found in various disciplines. We can find fuzzy logic applications in many control applications. Artificial neural networks are popular solutions to the problems of character recognition, voice recognition, prediction, medical data analysis, among others. Genetic Algorithm is one of the well-known methods in optimization problems. Although softcomputing has shown as an interesting and promising discipline from both theoretical and practical perspectives, the practical application of softcomputing methods, however, still leaves many problems for further investigation.

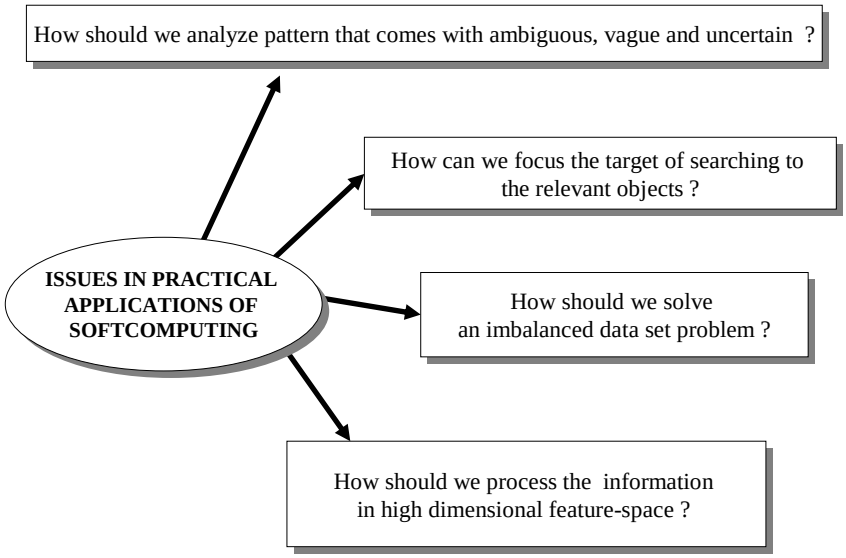

Figure 1 Topic of interests in softcomputing

To have a clear description, an illustration is depicted in Fig.1 showing four of such issues to be discussed in this dissertation. The first issue concerns with the analysis of information by which the problem is posed by the vagueness, uncertainty and ambiguousness as characteristics of the features. The popular example of such problem is handwriting character recognition. Various solutions involving those of statistical pattern recognition, neural networks or hybrid system have been proposed. Although satisfied performance of the method has been achieved, the necessity of developing an accurate character recognition system with low memory requirement is still highly demanding problem. These factors should be taken into consideration when the character recognition system is turned into hardware implementation.

The second issue concerns with the problem of finding a target in a searching area. Let us take human visual system as example, and text located in front of the eyes is assumed as the target of searching. We can categorize the objects inside the picture captured by our eyes into two classes.

The first class is termed as *relevant* information while the remainings are belonging to *irrelevant* ones. To find text from this domain, there should be a cooperation between visual attention and knowledge retrieval in human mind. The knowledge of characteristics of the character is utilized to focus the visual attention to find the target in searching area, while the irrelevant objects are ignored. The development a mathematical model to interpret this mechanism will contribute many advantages in practice. The application of this technology can be found in robot vision, auto-navigation for cars, and developing support system for the visually handicapped.

The third issue is termed as imbalanced data sets problem. This is defined as a problem whereas one of the class is heavily under represented compared to the other. Problem with such properties are often found when we are analyzing data obtained from real-life domain. This condition will create several difficulties for algorithms that assume the balanced condition of the classes. Some examples of real world applications with this kind of feature are the detection of fraudulent telephone calls, spotting unreliable telecommunications customers, rare medical diagnosis such as the thyroid disease in the UCI repository, and travel demand forecasting.

The forth problem is posed by the large number of features that generate a complex topological space. Knowledge discovery in this high-dimensional feature space is challenging theme and highly demanding, especially because of the rapid progress in computational molecular biology (*bioinformatics*)
.
Bioinformatics is a new discipline with motivation to organize, analyze, and distribute biological information in order to answer complex biological questions. It involves the solution of complex biological problems using computational tools and systems. It also includes the collection, organization, storage and retrieval of biological information from
databases. In particular, many efforts are currently dedicated to clarify the function of genes and their relation with disease. The rapid progress in this work is because of the invention of DNA microarray that enables us to measure the expression patterns of thousands of genes in parallel, hence enables a comprehensive analysis to reveal their functions. The number of genes in human body is estimated around 35,000 genes arranged on 23 chromosomes forming a complex topological feature space. In such situation, designing an appropriate pattern recognition method will encounter various problems, such as the curse of dimensionality and computational time. It has opened the opportunity to investigating novel method that is capable to work properly in such complex space within a reasonable computational time.

2. Objective of the Research

In relation to the prior description, we have conducted various experiments covering a broad area and also involving many disciplines: from character recognition, meteorological prediction through medical informatics. In summary, the objectives of this study are described as follows. In the first part of this dissertation, we developed a classification system that works properly to solve a problem with main difficulties in the vagueness, uncertainty and ambiguous properties of the features. As case of the study, we conducted experiments on handwriting character recognition, with numeric, roman alphabet and Japanese Katakana characters as target of classification. This work was our collaboration with Sanyo Electric Inc. to develop a high-end interface that enables handwriting numeral characters as entry information, instead of pushing the buttons to dial a destination number. As significant part of this system, we evaluate the performance of large scale neural network CombNET-II in numeral character recognition problems. The target of classification is then extended by incorporating the Japanese Katakana handwriting characters in the training sets of the neural network. This modification will help the user not to memorize the dial number. What he

should do is just to write, in the header of the draft, the name of the person to whom the facsimile will be sent. The neural network inside the facsimile works to read and to interpret the characters into dial number based on the look-up table of name-fax number that has been registered in prior.

Still in the workframe of character recognition, the second objective of this dissertation is developing an automatic system to find characters in color image. It is a well-known issue in computer vision, and various algorithms have been proposed. However, most of the proposed algorithms assume that text should be composed by a string of characters which makes distinctive features to the non-text regions. Consequently, the algorithms are appropriate if the target of searching is limited to roman alphabet characters. In this study, we propose a system that works appropriately to extract the Japanese Kanji characters in color image. The particular problem posed by Kanji is that it can exist as single character to express a certain word, such as the name of station or restaurants. In such situation, the appropriate solution can be achieved by building a system which is capable to retrieve the knowledge of character to select the relevant targets from the visual domain.

The third objective of this dissertation is proposing a strategy to deal with a classification problem of which one class is heavily under-represented to the other. This is termed as imbalanced data sets problem which is often found in data obtained from real-life domain. As case of the study, we evaluate the algorithm to solve fog forecasting problem whereas the appearance of fog is very rare.

The last objective of this dissertation is developing a system to analyze information in high-dimensional feature space, in bioinformatics problem. The rapid progress of hardware and software development open the possibility to begin an analysis of complex-presented problem such as the DNA gene-expression. It is known from the latest progress of this discipline, that the human body has around 35,000 genes arranged on 23 chromosomes forming a complex topological feature space. A comprehensive analysis of the pattern of gene-expression obtained from human cancer cell lines was conducted to reveal underlying mechanism of many diseases, and invention of many new medical therapeutics. This work is our collaboration with Division of Cancer-Related Genes, Institute for Genetic Medicine, Hokkaido University, Japan.

3. Organization of this Dissertation

To provide a clear description of our work, this dissertation is organized as follows. Chapter 2 provides the general background of theoretical foundation of the methods used throughout this thesis. We provide an introduction to neural network and genetic algorithm as two paradigms of softcomputing. A special section is dedicated to provide theoretical foundation of CombNET-II, since this model is used in many parts of our study.

In Chapter 3, we will describe the application of CombNET-II in handwriting character recognition problem. To accomodate the complex properties of the handwriting patterns, we modified self-growing algorithm in order to work in supervised fashion, which is explained in Section 3.2. Two experiments to evaluate the performance of CombNET-II are reported. In Section 3.3, we discuss the experiments on handwriting numeral character recognition using CombNET-II. In Section 3.4, the target of recognition is extended by the inclusion of katakana and roman alphabets. The conclusion of this chapter is provided in Section 3.5.

In Chapter 4, the discussion is focused on character segmentation as a preprocessing of character recognition system. Section 4.1 introduces the motivation of this study, followed by overview of the proposed system in Section 4.2. Section 4.3 describes the experimental results of this study, and the conclusion of this chapter is drawn in section 4.4.

In Chapter 5, we will discuss the imbalanced data sets problem and its application to fog forecasting. A new solution is proposed by modifying large-scale neural network CombNET-II which is

explained in Section 5.2. Section 5.3 reports the experimental results of model evaluation in fog forecasting problem. The conclusion of this chapter is drawn in Section 5.4.

Chapter 6 describes our present work in exploring the information inherited in gene-expression profiles of mRNA taken from cancer patients. This chapter is started by reviewing essential concepts in molecular biology, provided in Section 6.1. Section 6.2 contains an overview of tumour-suppressor gene p53 as topic of interest. In Section 6.3, we explain the feature subset selection methods which will be used in our analysis. Section 6.4 describes the experimental results of our study in discrimination of three types of TP53 tumour-suppressor gene status alterations, and the conclusion is provided in Section 6.5.

The rest of this dissertation contains conclusions, future works and bibliography of this study.