The PIC-TDD Framework of Test Data Design for Pattern Recognition Systems

Xiangdong Wang, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China

Ying Yang, College of Information and Electrical Engineering, China Agricultural University, Beijing, China

Hong Liu, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China

Yueliang Qian, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China

ABSTRACT

In this paper, a new approach is proposed for the design of test data for pattern recognition systems. In the theoretical framework put forward, performance on the population of data is viewed as expectation of a random variable, and the purpose of test is to estimate the parameter. While the most popular method of test data design is random sampling, a novel approach based on performance influencing classes is proposed, which can achieve unbiased estimation and the variance of estimation is much lower than that from random sample. The method is applied to the evaluation of systems for broadcasting news segmentation, and experimental results show the advantages over the random sampling approach.

Keywords: Broadcast News Segmentation, Data Population, Performance Influencing Class, PIC-TDD, Test Data Design

1. INTRODUCTION

In the field of pattern recognition, it is well known that the variance and diversity of input data poses great challenge to performance assessment, since a system may achieve quite different performances on different test data. This makes it difficult to assess the overall performance of a system considering all possible test data and to compare the performances between different systems. Recently, while many pattern recognition systems, e.g. speech recognition systems achieve real-world application, the evaluation of performance on all possible...
data in real-world application becomes a major challenge, and the key problem of the challenge is the design and collection of test data. Currently, a popular approach for assessing the performance of a system is the evaluation scheme (Ajot and Fiscus, 2009; Vandecatseye et al., 2004; Deshmukh et al., 1996; Douglas and Janet, 1992), in which test database is collected and all systems are tested with the same data. Though this makes comparison possible, the performances obtained are still highly dependant on the test data. Most test data are collected randomly or arbitrarily, so they are not representative of all possible data and the performances are not representative, either.

Some researchers noted that data properties (e.g., speaker gender, age, or dialect for speech data) may affect system performance and tried to arrange some basic data properties. For example, Deshmukh et al. (1996) and Douglas and Janet (1992) added noises to the test data at SNRs of 10dB, 16dB and 22dB. Pearce and Hirsch (2000) and Nakamura et al. (2003) added 8 types of noise at 5 SNR levels. However, the data properties and their values are all decided arbitrarily, which cannot ensure the test data are representative of data in real-world application. Some researchers (Steeneken et al., 1989; Nagorski et al., 2002) automatically select test data from a large data set using acoustic features. Although the selected data is representative, they are highly dependent on the large data set used. If the large data set is biased, the selected data will be biased, too.

In this paper, the PIC-TDD (Performance Influencing Class based Test Data Designing) framework is presented. Under the framework, performance on the population of all possible data is estimated to analyze the system. By means of the analysis, difference in performance caused by different test data can be avoided, performance in real application can be predicted, and comparison between different systems tested with different data can be easily realized. The PIC-TDD framework is applied to the analysis of a broadcast news segmentation system. Experimental results were given, which verified the effectiveness of PIC-TDD.

The rest of the paper is organized as follows. In Section 2, statistical description of evaluation are presented, which is the basis of the theory of PIC-TDD. In Section 3, the theory and procedure of PIC-TDD are presented. In Section 4, we describe the application of PIC-TDD to test data designing of a broadcast news segmentation system. Experimental results and analysis for the application in broadcast news segmentation are also given in Section 4. Finally, in Section 5, conclusions are drawn and future work is discussed.

2. STATISTICAL DESCRIPTION OF EVALUATION

2.1. Performance on Data Population

First, some basic concepts need to be explained:

- **Quantity of data**: The amount of data, e.g., the number of images, the duration of speech, or the number of words in speech. The quantity of a data set $D$ is denoted by $q(D)$;
- **Basic data unit (BDU)**: Data of unit quantity, e.g., one image, speech lasting for one second, or speech containing one word. Therefore, a data set $D$ can be viewed as a set of BDUs;
- **Data population**: The set of all data, denoted by $\Omega$. As explained above, $\Omega$ can be viewed as the set of all BDUs. Though data population seems to be infinite in some cases, in this paper, to make description simple, we assume that the data population is a finite set whose size can be as large as needed;
- **Result function**: When evaluating the performance of a pattern recognition system, the system output is compared to the reference to give a value concerning the number of correctness or errors. For a given system, this value can be seen as a function of the data and is referred to as result function;
Reinforcement and Non-Reinforcement Machine Learning Classifiers for User Movement Prediction
[www.igi-global.com/chapter/reinforcement-non-reinforcement-machine-learning/76790?camid=4v1a](www.igi-global.com/chapter/reinforcement-non-reinforcement-machine-learning/76790?camid=4v1a)

Designing a Ubiquitous Audio-Based Memory Aid
[www.igi-global.com/chapter/designing-ubiquitous-audio-based-memory/37797?camid=4v1a](www.igi-global.com/chapter/designing-ubiquitous-audio-based-memory/37797?camid=4v1a)

"Neomillennial" Learning Styles Propagated by Wireless Handheld Devices
Edward Dieterle, Chris Dede and Karen Schrier (2007). *Ubiquitous and Pervasive Knowledge and Learning Management: Semantics, Social Networking and New Media to Their Full Potential* (pp. 35-66).
[www.igi-global.com/chapter/neomillennial-learning-styles-propagated-wireless/30475?camid=4v1a](www.igi-global.com/chapter/neomillennial-learning-styles-propagated-wireless/30475?camid=4v1a)