The PICA Framework for Performance Analysis of Pattern Recognition Systems and Its Application in Broadcast News Segmentation

Xiangdong Wang[1,2], Meiyin Li[1,2],
Shouxun Lin[1], Yueliang Qian[1], and Qun Liu[1]

[1] Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100080, China
[2] Graduate University of Chinese Academy of Sciences, Beijing 100085, China
{xdwang,limeiyin,sxlin,ylqian,liuqun}@ict.ac.cn

Abstract. In this paper, the performance influencing class analysis (PICA) framework is proposed for performance analysis of pattern recognition systems dealing with data with great variety and diversity. Through the PICA procedure, the population of data is divided into subsets on which the system achieves different performances by means of statistical methods. On basis of the division, performance assessment and analysis are conducted to estimate the system performance on the whole data population. The PICA framework can predict true performance in real application and facilitate comparison of different systems without the same test set. The PICA framework is applied to the analysis of a broadcast news segmentation system. The procedure is presented and experimental results were given, which verified the effectiveness of PICA.

Keywords: performance analysis, PICA, PIFA, pattern recognition, speech recognition, broadcast news segmentation.

1 Introduction

In the field of pattern recognition, the variance and diversity of input data poses great challenge to performance assessment, since a system can 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. As a resolution, the evaluation scheme is popular adopted, in which test database is collected with variability in some basic data properties (e. g. speaker gender for speech data) 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. In most evaluations [1,2,3] and other researchers work, the test data are collected randomly or arbitrarily despite the coverage of some basic data properties, so they are not representative of all data and the performances are not representative, either.

In this paper, we present the Performance Influencing Class Analysis (PICA) framework of performance assessment for applied pattern recognition systems. It

H.G. Okuno and M. Ali (Eds.): IEA/AIE 2007, LNAI 4570, pp. 925–934, 2007.

aims to estimate the overall performance on the set of all possible input data (the population) for a given task, making the performance assessment more reliable and performance comparison between different systems feasible. The main idea of PICA is to divide the data population into some subsets, and estimate the performance on the data population with performances on the subsets and the proportions of each subset in the data population.

The rest of the paper is organized as follows. In Section 2, the basic logic and procedure of PICA are presented. In Section 3, we describe the application of PICA to the performance assessment of a broadcast news segmentation system. Experimental results and analysis for the application in broadcast news segmentation are given in Section 4. Finally, conclusions are drawn in Section 5.

2 The Framework of PICA

2.1 The Basic Logic of PICA

To further explore the PICA framework, some terms are proposed as follows.

Population: the set of all possible input data for a specific task, denoted by Ω. Each element of the population is a *basic unit of data*, which can be decided according to the features of data and the task performed. For example, for a speech recognition task, the basic unit of data may be decided as one utterance.

Data property: the feature or characteristic of each basic unit of data, e. g. the gender of speaker, or the signal-noise ratio for the speech recognition task. For a basic unit of data d, its data properties are denoted as $d.P_i$, where $i = 1, 2, \ldots$

Levels of a data property: values or classes of a data property. In PICA, only discrete levels are used, so levels of data properties with continuous values are decided by dividing the value domain into intervals. We defined that for basic unit of data d, $d.P_i$ refers to the level of the data property instead of the value.

Performance metric: a value as the measurement of performance. For example, word error rate for the continuous speech recognition task.

PIF and PIC. The main idea of PICA is to divide the population into subsets satisfying that performances are significantly different on different subsets and close on data in the same one. These subsets are referred to as performance influencing classes (PICs). To achieve the division, the method of ANOVA (analysis of variance) [4,5] is introduced from statistics, which is a powerful method of hypothesis test. For data with n data properties, when not considering interaction between data properties, the statistical model of ANOVA is written as

$$F_{l_1 l_2 \ldots l_n m} = \mu + \sum_{i=1}^{n} \tau_{l_i} + \varepsilon_{l_1 l_2 \ldots l_n m} \tag{1}$$

where l_i $(i = 1, \ldots, n)$ stands for a level of the i^{th} data property, $F_{l_1 l_2 \ldots l_n m}$ denotes the performance metric value on the m^{th} data with specific data property levels, τ_{l_i} denotes the effect of the l_i level on the performance, and ε denotes

experimental error. The purpose of ANOVA is to test statistically that for a data property with K levels, whether the following hypothesis is accepted or rejected.

$$\tau_1 = \tau_2 = \ldots = \tau_K = 0 \tag{2}$$

Definition 1. *If for a data property, the hypothesis in (2) is rejected through ANOVA, the data property is called a performance influencing factor (PIF).*

Definition 2. *For a PIF P_i, if for each two levels l_1, l_2 of it, $\tau_{l_1} = \tau_{l_2}$ is rejected through ANOVA, then P_i is called a level-complete PIF.*

Only PIFs are considered when dividing the population into PICs, since other data properties bring no significant difference in performance. Though not all PIFs are level-complete ones, in practice, most PIFs can be modified to be level-complete PIFs by adjusting the definition of levels.

Theorem 1. *Let P_i be a level-complete PIF, whose levels are $l_1, l_2, \ldots, l_K$, and set A_j is defined as $A_j = \{d | d.P_i = l_j, d \in \Omega\}$, $j = 1, 2, \ldots, K$, then $S_i = \{A_j, j = 1, 2, \ldots, K\}$ is a partition of Ω.*

Proof. S_i is a partition of Ω because A_j, $j = 1, 2, \ldots, K$, satisfying

$$A_{j_1} \cap A_{j_2} = \Phi, j_1 \neq j_2, \text{ and } \bigcup_{j=1}^{K} A_j = \Omega \tag{3}$$

□

Definition 3. *The partition $S_i = \{A_j, j = 1, 2, \ldots, K\}$ in Theorem 1 is called the performance influencing partition of Ω for P_i, and $A_j \in S_i, j = 1, 2, \ldots, K$ is called a performance influencing class (PIC) for P_i.*

Definition 4. *Let $P_1, P_2, \ldots, P_n$ be level-complete PIFs, whose performance influencing partitions are $S_1, S_2, \ldots, S_n$, then the product of the partitions $S = S_1 \cdot S_2 \cdot \ldots \cdot S_n$ is called the performance influencing partition of Ω for $P_1, P_2, \ldots, P_n$, and each $B \in S$ is called a performance influencing class for $P_1, P_2, \ldots, P_n$.*

It can be seen from the definitions that when many data properties are considered, the performances on different PICS are quite likely to be different due to different levels of PIFs. And when enough data properties are considered and not too few data are used, performances on subsets in the same PIC are likely to be similar for few factors may influence the performance in a PIC.

Estimation of Performance on the Population. Once the PICs are determined, performance metric value on each PIC can be obtained by testing the system using corresponding data. These performances as a whole can give more information than simply test the system using a randomly selected test set. But sometimes, only one metric value is needed to represent the overall performance on population or to compare with other systems. In the following theorem, it is proved that for metrics such as precision or error rate, the metric value on the population equals to a weighted sum of metric values on all PICs.

Theorem 2. *Assume that for a data set D, a performance metric is defined as*

$$R = f(D)/q(D) \tag{4}$$

where f and q are functions of D, satisfying that for two data sets D_1, D_2,

$$f(D_1 \cup D_2) = f(D_1) + f(D_2), q(D_1 \cup D_2) = q(D_1) + q(D_2), \text{ if } D_1 \cap D_2 = \Phi \tag{5}$$

Then for a partition $S = \{A_1, A_2, \ldots, A_n\}$ of the data population Ω, letting R_i be the metric for the subset A_i, the following holds true.

$$R = \sum_{i=1}^{n} c_i R_i, \text{ where } c_i = q(D_i)/q(\Omega), i = 1, 2, \ldots, n \tag{6}$$

Proof. According to (4) , for each subset, we have $R_i = f(D_i)/q(D_i)$, and according to (5), $f(\Omega) = \sum_{i=1}^{n} f(D_i)$, $q(\Omega) = \sum_{i=1}^{n} q(D_i)$. Let $Q = q(\Omega)$, we have $R = f(\Omega)/Q = [\sum_{i=1}^{n} f(D_i)]/Q = [\sum_{i=1}^{n} R_i q(D_i)]/Q = \sum_{i=1}^{n} [q(D_i)/Q]R_i = \sum_{i=1}^{n} c_i R_i$ □

In fact, in most metrics used in the pattern recognition area, the function $q(D)$ in the above theorem usually stand for the amount of data, such as the number of basic units of data or the whole duration of speech. So the proportion $q(D_i)/Q$ stands for the proportion of amount of D_i in the population.

Design of Test Data. For performance metrics that do not satisfy (4) or (5), there is a more direct way for estimating overall performance on the population. That is, to design and collect a test set in which the proportion of each PIC is equal to that in the population. For cases that levels of all PIFs can be controlled when collecting data, this can be easily done. However, for most cases, the data are just collected with little control, so a selection approach is proposed, as described in the following.

Let Ω be the data population for a specified task. Suppose that there are K PICs. If there are n sets of data already collected, denoted by $D_1, D_2, \ldots, D_n$. The data amount of D_i is N_i, the proportion of the j^{th} PIC in D_i is a_{ij}, and the proportion of the j^{th} PIC in the population is b_j. So the problem can be described as forming a test data set D of data amount N by selecting sets from $D_1, D_2, \ldots, D_n$, satisfying that the proportion of PICs are most close between D and Ω. The *Euclid distance* is used as the measurement of similarity between the proportions of PICs in D and Ω, so this can be transformed into the problem of finding $X_0, X_1, \ldots, X_n$ that minimizes

$$d(D, \Omega) = [\sum_{j=1}^{K} (\frac{1}{N} \sum_{i=0}^{n} a_{ij} N_i X_i - b_j)^2)]^{1/2} \tag{7}$$

under the restriction of

$$\sum_{i=0}^{n} N_i X_i = N, \ \mathrm{X}_i \in \{0, 1\}, \ i = 1, \ldots, n \tag{8}$$

This is a problem of *integer programming* and can be solved using classical algorithms such as the *branch and bound method* [10].

2.2 The Procedure of PICA

The whole procedure of PICA is shown in Figure 1.

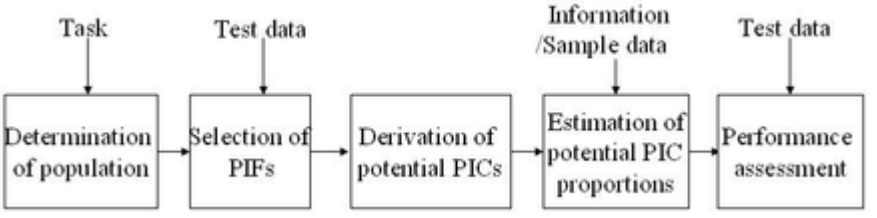

Fig. 1. The procedure of PICA

First, the data population should be determined according to the task, which includes deciding the basic unit of data and the coverage range of the population.

Once the population is fixed, data properties under examination should be decided because it is impossible to study all of them. Then the framework of PIFA [8] is incorporated into PICA to select PIFs from the data properties. The main idea of PIFA is to design the experiments using orthogonal design and test for significant differences of performance between levels using ANOVA.

After the PIFs are chosen, their levels may be slightly adjusted to become level-complete PIFs. Then, PICs for each PIF and all PIFs are derived according to Defination 3 and 4. The product of multiple partitions is computed as [9]

$$S = S_1 \cdot S_2 \cdot \ldots \cdot S_n = \{\bigcap_{i=1}^{n} A_i | \bigcap_{i=1}^{n} A_i \neq \Phi, A_i \in S_i\} \tag{9}$$

To decide whether $\bigcap_{i=1}^{n} A_i \neq \Phi$, all $\bigcap_{i=1}^{n} A_i$ are maintained as *potential PICs*, and after the proportions of all potential PICs are estimated, the sets with proportion less than a threshold will be eliminated as approximate null set.

As for estimating the proportions, information from other researchers may be useful. If such information is not available, the sampling method can be used to draw a sample from the population and compute the proportions in the sample set as substitution. Sampling theory supports that if the sampling method is appropriate, the sample set may be good miniature of the population [6,7].

When PICs are determined and their proportions are know, an overall performance metric is calculated as estimation of performance on the population. There are two ways for such assessment—-estimating the metric on the population using (6) or designing a test set by solving (7) and (8).

3 Application of PICA to Broadcast News Segmentation

To verify the effectiveness of the PICA, we applied it to the performance assessment of a broadcast news segmentation system. The aim of broadcast news segmentation is to segment the audio stream into homogeneous regions (referred to

as *segments*) according to speaker identity, environmental condition and channel condition. It can also be seen as detection of change points which are the positions in the audio where change occurs in speaker identity or recording conditions.

Considering the cost of experiment, the system task is narrowed to segmentation of broadcast news from one radio channel (China National Radio). Segmentation result of the system is compared to the reference to yield performance metrics of rejection rate (RJ) and false alarm rate (FA), as defined in the following, where N_{miss}, N_{fa}, and N_{ref} denote the number of missed, false alarm and reference change points.

$$RJ = N_{miss}/N_{ref} \tag{10}$$

$$FA = N_{fa}/(\text{Duration of the audio stream}) \tag{11}$$

The PICA framework is applied respectively to RJ and FA. Since the procedures are quite similar, only details for the FA metrics are described in this paper.

Determination of Population. First, decision should be made about the basic unit of data. In our work, when investigating false alarm rate, each segment (speech between two change points) is considered as a basic unit of data, since data properties such as recording condition usually keep the same within one segment and varies much between different ones. Then, population is defined as the set of all segments in broadcast news from China National Radio.

Selection of PIFs. Once the basic unit of data is chosen, data properties of a unit are also determined. Since there could be innumerous data properties, only those that may influence system performance and can be measured in practice are involved in the selection of PIFs. In our work, data properties about the speaker, recording condition and channel condition are chosen, and their levels are decided, as listed and explained in table 1.

When selecting PIFs from these data properties, the PIFA (performance influencing factor analysis) framework [8] is adopted. In our work, because the data properties such as speaker gender are only related to speech segments, a hierarchical approach is utilized: The two levels of the data property "content" is analyzed first using a 1-way ANOVA, and the orthogonal design (the orthogonal table $L_8(2^7)$ is adopted) is used for the other data properties. Then a data set (referred to as the *PIFA set*) of 6 hours is collected and used as input of the segmentation system. With the result and reference, statistical data of performance are generated for each group according the method presented in [8], the main idea of which is to divide the group into subgroups and consider metric on each subgroup as one observation. The results of ANOVA shows that content, speech background and speech scene influence system performance significantly ($Pr < 0.05$). So three PIFs are selected, which is content, background, and speech scene.

Derivation of Potential PICs. When PIFs are determined, PICs can be derived according to (9). As mentioned in section 2.2, potential PICs are first derived and tested later for whether it can be eliminated as a null set. For our work, the potential PICs are listed in table 2.

Table 1. Data properties and their levels

Data property	Level	Explanation
Content	Speech	Speech by human
	Non-Speech	Music, noise, etc.
Speaker gender	Male	Male speaker
	Female	Female speaker
Speaker accent	Yes	With dialectal accent
	No	Without dialectal accent
Speech mode	Planned	Reading planned text
	Spontaneous	Speaking spontaneously
Background	Yes	Speech with music or noise
	No	Speech in silent environment
Speech scene	Studio	Speech in studio
	Live	Speech in open environment

Table 2. Potential PICs and their proportions

	Content	Background	Speech scene	Proportion
1	Speech	No	Studio	0.566
2	Speech	Yes	Studio	0.028
3	Speech	No	Live	0.247
4	Speech	Yes	Live	0.083
5	Non-speech	—	—	0.060

Estimation of Potential PIC Proportions. Since no information of the potential PICs is available from other researchers, the approach of sampling is adopted, as explained in section 2.2. In our work, the sampling frame is defined as all broadcast news from China National Radio in 2005, and sampling unit is defined as one section. A procedure similar to stratified sampling [6,7] is performed: the sampling frame is divided into 12 strata according to the month (from Jan. to Dec.) and broadcast news of 2 hour is draw from each stratum using simple random sampling method. Notice that what is needed for these data is the information of the duration of the PICs, so full speech data is unnecessary. So with the sample data of 24 hours, the proportion of a PIC is calculated as the quotient of the duration of all segments in the PIC and the total duration of the sample set. The result is shown in table 2. Since the proportion of the 2nd potential PIC is less that 0.05 which is the threshold in our work, so it is eliminated. Therefore, there are only 4 PICs and the proportions are re-estimated, as shown in part of table 3.

Performance Assessment. Once the PICs are fixed, the system is tested with a test set of 2 hours in which total duration of each PIC is about 30 minutes and the FA metric is calculated for each PIC. The results are shown in Table 3. From the definition of FA, it can be seen that (6) is suitable. So FA metric on the population is estimated using (6), and the result is 3.213.

Table 3. Proportion and FA values of PICs

	Content	Background	Speech scene	Proportion	FA
1	Speech	No	Studio	0.592	2.348
2	Speech	No	Live	0.258	4.313
3	Speech	Yes	Live	0.087	4.944
4	Non-speech	—	—	0.063	4.427
Estimated FA on population = Proportion(i)*FA(i) = 3.213					

The method of test data designing is also used. A data set of 30 hours are divided into 90 clips each with the duration of 10 minutes, from which 12 clips are selected according to (7) and (8) resulting in a test set of 2 hours. The proportions of PICs in the selected data set are shown in Table 4. The system is tested with this set, and the FA metric obtained is 3.115, which is close to the estimation in Table 3.

Table 4. Details of the selected test set

PIC	1	2	3	4
Proportion in population	0.592	0.258	0.088	0.063
Proportion in the selected set	0.586	0.251	0.081	0.060
$d(D, \Omega)$	0.0014			
FA	3.115			

4 Experiments and Analysis

Test of PICs. It is expected that performance be similar when on data within the same PIC, and be significantly different on data from different PICs. So for each PIC, we collected 10 data sets of 20 minutes, and for comparison, ten random sets are also selected randomly without any consideration of PIC. The FA metrics were calculated on all 5 data sets, as shown in Figure 2(a). The result indicates that performance is consistent in one PIC and varies much for different ones. And for random sets with the same size, performance also varies much, which implies that testing the system using one or few random sets is unreliable.

Test of Performance Assessment. The main advantage of PICA is that it can achieve performance approximate that on the population. So we test the system using 5 difference test sets: the test set designed and the sample set described in Section 3, and 3 random selected test sets. Sizes and FAs for those test sets are listed in Table 5, which shows that performance on the test set designed is most similar to that on the sample set of larger size, while metrics on sample sets varies considerably for different sets. It is also favorable that the FA value estimated in Table 3 is quite close to the metric on the sample set, which means the estimation is reliable, too.

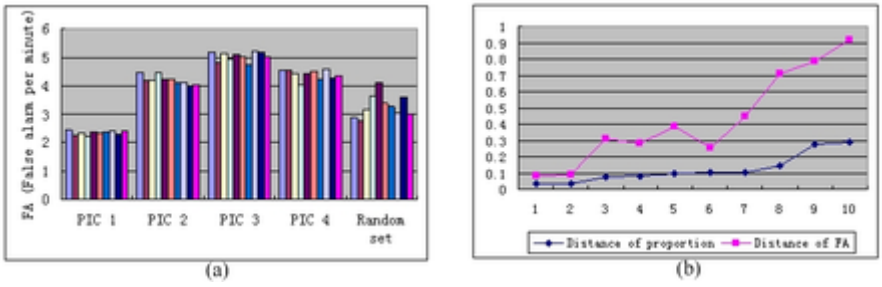

Fig. 2. Experiments results: (a) Result of test of PICs; (b) Result of test of performance assessment

Table 5. Result of experiment on different test sets

	Designed set	Sample set	Random set1	Random set2	Random set3
Size	2 hrs	24 hrs	2 hrs	2 hrs	2 hrs
FA	3.115	3.223	2.933	3.318	3.581

Figure 2(b) shows the result of another experiment: the system was tested using 10 test sets, each is of 2 hours. For each test set D_i, the PIC proportion distance between D_i and Ω is calculated as the Euclid distance, and distance of FA is calculated as

$$\text{FA distance for } D_i = |(\text{FA on } D_i) - (\text{FA on } \Omega)| \tag{12}$$

The figure indicates that the less the distance D_i, the closer the performance on D_i to the performance estimated for the population, implying that when the distance is small enough, the performance on the designed test set is close to that on population.

5 Conclusions

In this paper, the PICA (performance influencing class analysis) 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.

Also in this paper, the application of the PICA framework to the analysis of a broadcast news segmentation system was described. The whole procedure was presented and experimental results were given, which verified the effectiveness of PICA.

Actually, the PICA framework can be applied to any pattern recognition task with complex input data. So in the future, we are planning to apply PICA in other task of speech recognition and wider fields.

References

1. Vandecatseye, A., et al.: The COST278 pan-European Broadcast News Database. In: Procs. LREC 2004, Lisbon, pp. 873–876 (2004)
2. Paul, D.B., Baker, J.M.: The Design for the Wall Street Journal-based CSR Corpus. In: Proceedings of Second International Conference on Spoken Language Processing, pp. 899–902 (1992)
3. Pearce, D., Hirsch, H.-G.: The Aurora Experimental Framework for the Performance Evaluation of Speech Recognition Systems under Noisy Conditions. In: Proceedings of 6th International Conference on Spoken Language Processing, pp. 29–32 (2000)
4. http://www.statsoft.com/textbook/stanman.html
5. Shen, Q.: SAS Statistical Analysis, pp. 84–107. Higher Education Press, Beijing (2005)
6. Renssen, R.: A Course in Sampling Theory. http://www.cs.vu.nl/~stochgrp/aionetwerk/course.doc
7. Friel, C.M.: Sampling Theory. http://www.shsu.edu/~icc_cmf/cj_787/research11.doc
8. Wang, X., Xie, F., et al.: DOE and ANOVA based Performance Influencing Factor Analysis for Evaluation of Speech Recognition Systems. In: ISCSLP, Proceedings (companion volume), pp. 431–442 (2006)
9. Jiao, Z.-y., Hu, Y.-p.: Quotient Set and Fundamental Operation of Quotient Set. Journal of Xi'an University of Science and Technology 24(3), 372–375 (2004)
10. Kaufmann, A.: Pure and Mixed Integer Programming (1997)