RANDOM SEQUENCE RECOGNITION BASED ON SAMPLING DISTRIBUTIONS

Kvitashvili Avtandil
International Black Sea University, Tbilissi

Abstract

We considered small and large sample cases, or $n < 30$ or $n \geq 30$, where samples are presented as sequences of random numbers or vectors. Each vector of random components is transformed into one number – the mean of random components. Then the sampling distribution is constructed for the classes of sequences using these means of vector components with a corresponding confidence interval. If confidence intervals of classes are overlapped, or classes are not separable, that practically does not happen in most cases, it is the subject of other consideration. Such confidence spaces are constructed for two or more classes using the training sequences. Then the distance between each pairs of sampling distribution means or centers are determined that defines a real confidence level or recognition accuracy. Afterwards, any new unknown coming sequence or vector is classified depending on to which class subspace it falls in, that in general case presents a hyper-ellipsoid.

Key words: sampling distribution, random series, confidence interval.

1. Introduction

Random sequence recognition (classification) is one of the actual problems in a wide range of technical, medical, economic, military and many other fields. The present approach uses a property of sampling distribution of random numbers as the confidence interval. This concept makes it possible to construct a series of sampling distribution means as $\eta_1, \eta_2, ..., \mu_\omega$ that are for ω – number of classes somehow distant from each other measured by a desirable confidence interval. In other words, using available data or sequences of known belonging to the appropriate classes we perform several steps of averaging and transformations to find a reasonable classification rule to ease classification of new unknown sequences with a high confidence that would depend on means of those sequences. This idea comes from properties of live neural networks where independently on a type of input signals information undergoes several steps of averaging of previous averaging with a next simple transformations [1]. As we will see later this approach creates substantial benefits due to its simplicity and quality of classification.

2. Decision of the problem

We use training random sequences of equal number of elements with a known assignment to one of the ω classes. If sequences have a different length we can add zeros for simplicity. Suppose we have m training n-dimensional vectors distributed along ω classes. Analogously, the input pulses coming to the synapses of neural cell or neuron are accumulated or simply averaged, or in other words, at each neuron takes place so called a spatial and temporal summing up of signals. Therefore, at the next stage of neurons those averaged signals are passing similar averaging or squeezing dimension of input signals. Here we omit discussion about such a non-linear transform as the threshold [2] that actually is the main averaging cutter. As we see living organisms are mostly making very complex behavioral decisions reflected by a simple answer – either “yes” or “no” or by a categorical assessment of an external influence.

Following the abovementioned averaging is very important substantial operation that was discovered as the Central Limit Theorem [1] allowing forecasting

population mean with a certain percentage of confidence for normally or non-normally distributed data using the distribution of sample means or sampling distribution. We consider two cases – for the large samples sized more than 30 and small samples sized less than 30, or in both cases we use training sequences size or length of which is $n > 30$ and $n < 30$. For the first sequence we apply z – distribution and for the second ones – t – distribution. In addition, in a one-dimensional case we have single sequence of numbers while in general case a process can be multidimensional or $k = 1,2,...,m$ sequences coming simultaneously or in parallel.

The simplest case implies dichotomy of one-dimensional number sequences having varying means that equal average means for each of sequences or sum of sequence numbers divided by sequence length or sample size – n. Let denote this mean as μ_i where $i = 1,2$ or we have n numbers in each training sequence belonging to class 1 or class 2. The training process consists of determining a set of sample means for either class 1 or class 2. In the case of two classes we are constructing sampling distribution for distinguishable class means in a sense of providing a certain distance between them satisfying a possible highest percentage of confidence. Discrimination into two classes for large samples is illustrated in Fig 1. As we will see later often the overlapping data along the number axes presented as a sampling distribution (averaged or mean values along the number axes) becomes separated in a sense of having distant means at the distance of more or equal to the sum of marginal errors for both sampling distributions. This relationship is given in expression (1).

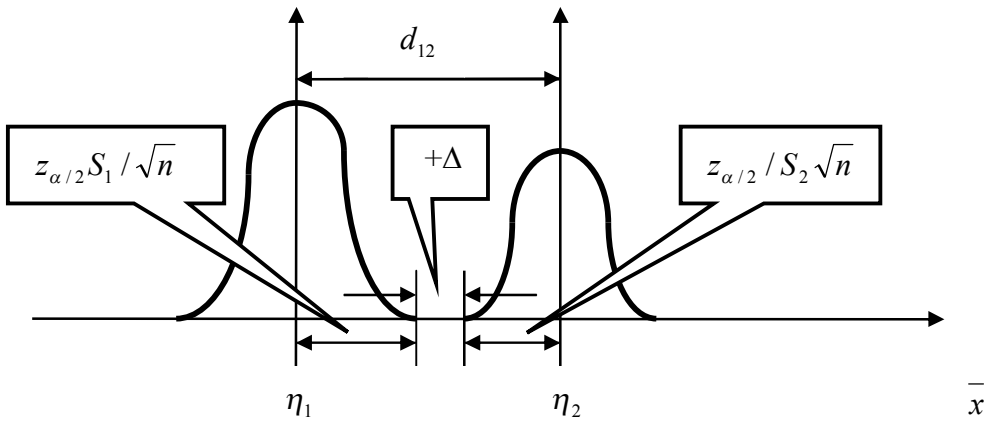

Fig. 1 – Sampling Distributions for Two Classes

$$\left|\eta_1 - \eta_2\right| \geq z_{\alpha/2} \times S_1 + z_{\alpha/2} \times S_2 \qquad (1)$$

Where η_1 and η_2 are the sampling distribution means (more precisely $\eta_{1\bar{x}}$ and $\eta_{2\bar{x}}$) correspondingly to two class sequences. Note that in this case sample sizes or sequence lengths are equal and is n. $+\Delta$ is a reserve or additional distance for confidence interval.

Fig. 2 is an illustration of two-dimensional processes or *m=2,* belonging to ω = 1, 2 or 3 classes, having different standard deviations.

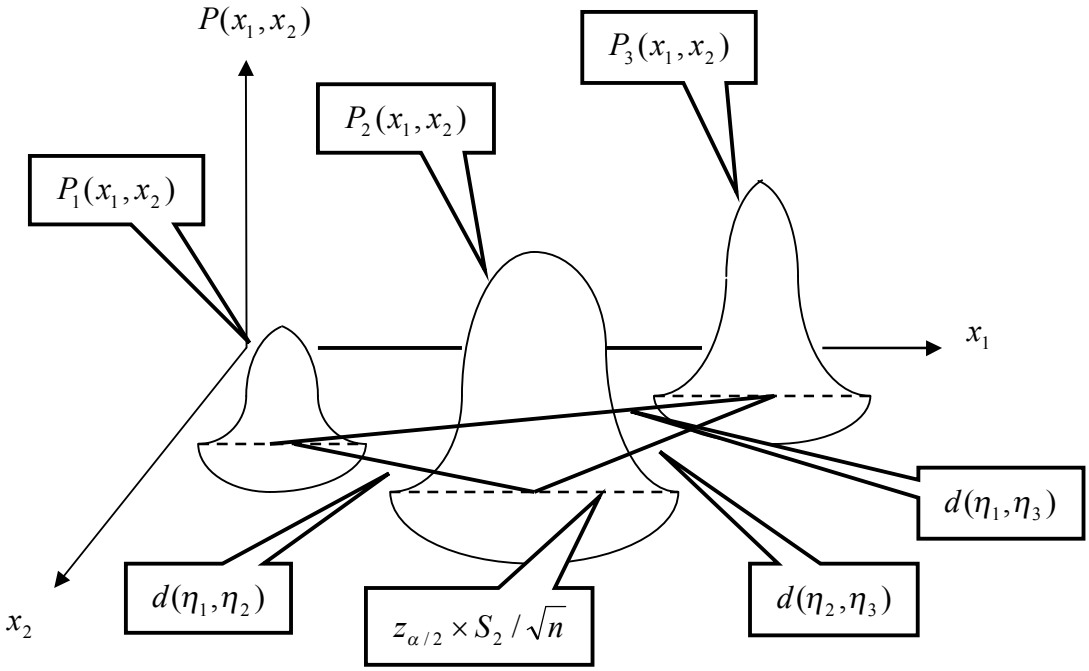

Fig. 2 – Sampling Distribution for Two-Dimensional Processes of Three Classes

According to the Fig.2 for large samples the expression (1) is satisfied for all three classes because in many class cases for separation between all classes we must have the following expression:

$$\left|\eta_i - \eta_j\right| = d(\eta_i, \eta_j) \geq z_{\alpha/2} \times S_i / \sqrt{n} + z_{\alpha/2} \times S_j / \sqrt{n} = z_{\alpha/2} \times (S_i + S_j) / \sqrt{n} \qquad (2)$$

where: i,j = 1, 2, … , ω, i ≠ j, η_i, η_j are the sampling distribution means, *n* is the sample size, S_i, S_j are the standard deviations of *i*-th and *j*-th classes and $z_{\alpha/2}$ is *z*- value for a desirable confidence level α , that is divided by two because of a one-sided confidence.

Therefore, if we have *m>2*-dimensional training processes having equal lengths of *n*, or *m*-parallel series of sequences: $\{x_{11}^i, x_{21}^i, ..., x_{m1}^i\}, \{x_{12}^i, x_{22}^i, ..., x_{m2}^i\}, ..., \{x_{1n}^i, x_{2n}^i, ..., x_{mn}^i\}$, where $i=1,2,...,\omega$ is a class index, the length of each training sequence contains ωn numbers. Then for each sample of *i*-th class of size *n* we would have sampling distribution means $\{\eta_1^i, \eta_2^i, ..., \eta_m^i\}$, where *k* = 1,2, …, *m*. Let these η_k^i-s represent a vector $\vec{\eta}_k^{\,i}(\eta_1^i, \eta_2^i, ..., \eta_m^i)$ - a mean of sampling distribution means, which can be different for various classes, or we would have $\vec{\eta}_k^{\,i}$ vector in Euclidean space **,** where *i* = 1,2, …,ω. In other words, a separation between classes with a given confidence level is similar to the expressions (1) and (2) as:

$$\left| \vec{\eta}_k^{\,i} - \vec{\eta}_k^{\,j} \right| \geq z_{\alpha/2} \times \left| \vec{S}_k^{\,i} \right|^2 / \sqrt{n} + z_{\alpha/2} \times \left| \vec{S}_k^{\,j} \right|^2 / \sqrt{n} \qquad (3)$$

where: $\eta_1^i = \frac{1}{n}\sum_{l=1}^{n} \bar{x}_{l1}^{\,i} = \bar{x}_{11}^{\,i} + \bar{x}_{21}^{\,i} + \ldots + \bar{x}_{n1}^{\,i}$, $\quad S_1^i = \sqrt{\frac{\sum_{l=1}^{n} (\bar{x}_{l1}^{\,i} - \eta_1^i)^2}{n-1}}$,

$\eta_2^i = \frac{1}{n}\sum_{l=1}^{n} \bar{x}_{l2}^{\,i} = \bar{x}_{12}^{\,i} + \bar{x}_{22}^{\,i} + \ldots + \bar{x}_{n2}^{\,i}$, $\quad S_2^i = \sqrt{\frac{\sum_{l=1}^{n} (\bar{x}_{l2}^{\,i} - \eta_2^i)^2}{n-1}}$,

$\cdots \qquad\qquad \cdots$

$\eta_m^i = \frac{1}{n}\sum_{l=1}^{n} \bar{x}_{lm}^{\,i} = \bar{x}_{1m}^{\,i} + \bar{x}_{2m}^{\,i} + \ldots + \bar{x}_{nm}^{\,i}$, $\quad S_m^i = \sqrt{\frac{\sum_{l=1}^{n} (\bar{x}_{lm}^{\,i} - \eta_m^i)^2}{n-1}}$,

Suppose for some pairs of classes after processing training sequences or determination of all class spaces the above mentioned expressions are not satisfied, or the differences between means of sampling distributions are less than the sum of appropriate margins of error. Then we are decreasing confidence level (or increasing α - value) up to the level for satisfying conditions (1), (2), (3) that shows a real probability of recognition for the given pair of classes. If it is impossible to meet those conditions, the classes are not divisible according to the proposed approach. Then we use a preliminary transformation of data for these classes that we will show hereafter that is successful in most cases and as a rule is used to apply in small sample cases.

Thus let consider *m*-dimensional space in which for ω classes we have ω hyper-ellipsoids centered in $\vec{\eta}_k^{\,i}$ points. Radiuses of hyper-ellipsoids are determined by margins of error defined from the desirable confidence level of recognition α. Practically this level in most cases is less than 0.01, or probability of correct recognition in these cases reaches almost 100% (according to satisfied confidence level α). In other words, after determination ω class sub-spaces (ellipsoids) for any new unknown sequence we computing the mean vector $\vec{\eta}_k^{\,i}$ and defining into what class sub-space it falls with a corresponding confidence.

Now let consider a small sample case when $n < 30$. As it is well known [3] for small samples the stem-and-leaf diagrams are used for visual analysis of data character that makes it possible to ease a statistical decision making. In this case we can simply look at the diagram of the data available and find whether the data is including outliers or it is congested for different classes. In the first case the square root transformation is applied while in the second case the log base 10 use is more rational. The square roots of numbers between 0 and 1 are larger than the original raw values, and square roots of numbers greater than 1 are smaller than original raw values. In other words, data values in each tail are drawn toward the mean of transformed data. This transformation results in more symmetrically shaped

distribution close to the normal and with variances more equal. On the other hand, the logarithm base 10 is the exponent or the power to which 10 must be raised to equal to the given number [3].

Let consider the following example. Table 1 presents 20 sets of sample data on percentage sales increase belonging to one of two classes not having outliers. Fig. 3 shows a corresponding stem-and-leaf diagram for that numbers where the mans of classes are not evidently separated and it is hard to see class deviations.

Table 1

Percentage Sales Increase for Product 1	**1.2**	**1.3**	**2.1**	**2.2**	**2.3**	**2.3**	**2.4**	**2.4**	**2.6**	**2.7**
Percentage Sales Increase for Product 2	**1.6**	**2.0**	**2.8**	**3.0**	**3.1**	**3.2**	**3.3**	**3.5**	**3.7**	**4.5**

Percentage Sales Increase for Product 1	**2.8**	**3.1**	**3.2**	**3.4**	**3.7**	**4.1**	**4.7**	**5.1**	**6.1**	**6.3**
Percentage Sales Increase for Product 2	**4.8**	**5.1**	**5.3**	**5.7**	**5.9**	**6.1**	**6.8**	**7.3**	**8.1**	**8.5**

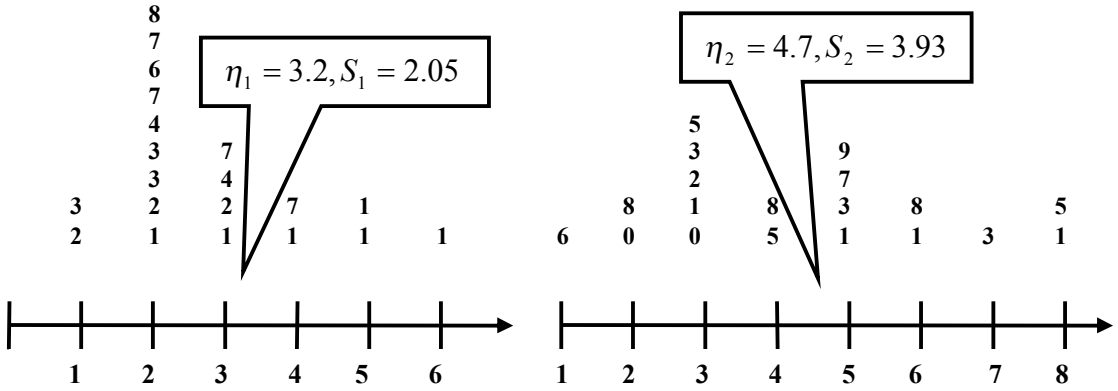

Fig. 3 – Stem-and-Leaf Diagram for Percentage Sales Increase of Two Products

Table 2 shows square roots appropriate to percentage sales increase data given in Table 1 and Fig. 4 illustrates corresponding stem-and-leaf diagram for this transformed data.

Table 2

Percentage Sales Increase for Product 1	**1.10**	**1.14**	**1.45**	**1.48**	**1.52**	**1.52**	**1.55**	**1.55**	**1.61**	**1.64**
Percentage Sales Increase for Product 2	**1.26**	**1.41**	**1.67**	**1.73**	**1.76**	**1.79**	**1.81**	**1.87**	**1.92**	**2.12**

Percentage Sales Increase for Product 1	**1.67**	**1.76**	**1.79**	**1.84**	**1.92**	**2.02**	**2.17**	**2.26**	**2.47**	**2.51**
Percentage Sales Increase for Product 2	**2.19**	**2.26**	**2.30**	**2.39**	**2.43**	**2.47**	**2.61**	**2.70**	**2.85**	**2.92**

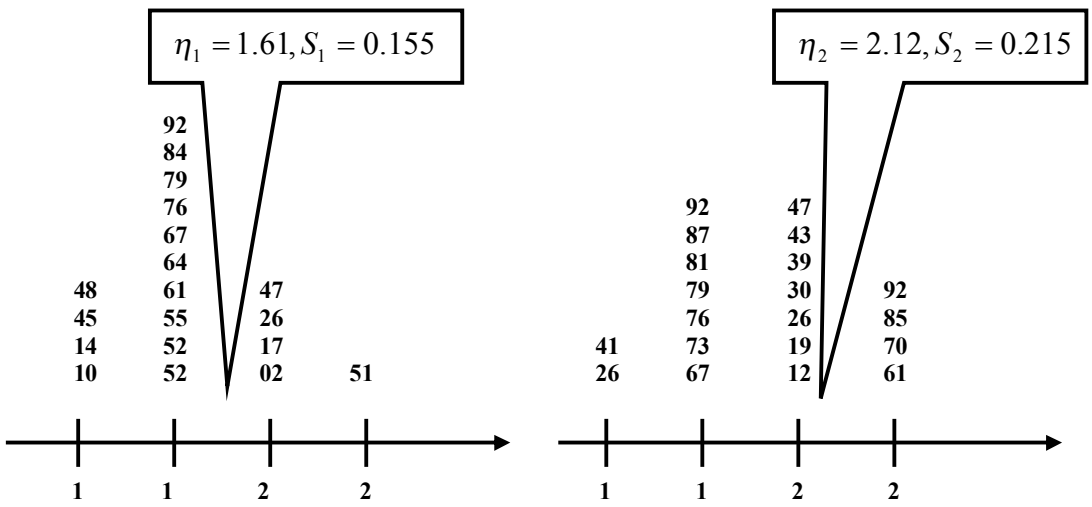

Fig. 4 – Stem-and-Leaf Diagram for Transformed Percentage Sales Increase of Two Products

As it is seen in Fig. 4 as a result of taking square roots the transformed number means and deviations are more discernable and shapes are near-normal. The shown computed values of means and standard deviations radically changed statistics. Relating to initial data after squared root transform the standard deviations are substantially smaller that makes it possible to squeeze the class confidence interval for two classes or class hyper-ellipsoid spaces of confidence for multi-class cases. That improves class separation and recognition quality in a large scale.

Let compute difference between classes and determine actual confidence intervals for both classes and the highest probability of recognition:

$$\eta \pm t \times S / \sqrt{n} \text{ , or}$$
$$\eta_1 + t \times S_1 / \sqrt{n} = 1.67 + 2.539 \times 0.155 / 4.47 = 1.755$$
$$\eta_2 - t \times S_2 / \sqrt{n} = 2.12 - 2.539 \times 0.215 / 4.47 = 1.998$$

3. Conclusions

As we see even 99% one-sided confidence interval is satisfied and there is an additional distance indicating that recognition quality actually is 100%. Here we are not considering such rare cases when the confidence intervals of training sequences are overlapping or could not provide a high confidence of recognition. It is a subject of other consideration.

References

1. McClave J.T., Benson P.G., Sincich T. Statistics for Business and Economics, Ninth Edition, Prentice Hall, 2005

2. Kvitashvili A.A. A Universal Approach to the Rational Design of Problem-Oriented Pattern Recognition Systems Based on Neural Networks, Proc. 3-rd Int. Conf. on Application of Fuzzy Systems and Computing, Wiesbaden, Oct. 5-7, 1998, pp.168-173.

3. Brightman H., Schneider H. Statistics for Business Problem Solving, Second Edition, South-Western Pbl.Co., Cincinnati, Ohio, 1994.

შემთხვევითი მწკრივების კლასიფიკაცია ამოკრეფითი განაწილების საფუძველზე

ავთანდილ კვიტაშვილი
შავი ზღვის საერთაშორისო უნივერსიტეტი, თბილისი

რეზიუმე

განხილულია მცირე და დიდი ამოკრეფების შემთხვევები, ანუ მისი ზომა $n < 30$ ან $n \geq 30$ წარმოდგენილი რიცხვთა შემთხვევითი მიმდევრობებით. იგულისხმება, რომ თითოეული შემთხვევითი კომპონენტებიანი ვექტორი გარდაიქმნება ერთ რიცხვად - ამ კომპონენტების საშუალო მნიშვნელობად. ამის შემდეგ ხდება ამ საშუალო მნიშვნელობების ამოკრეფითი

განაწილების აგება, რომელიც განსაზღვრავს მოცემული კლასის მწკრივებისათვის დამაჯერებლობის ინტერვალს შესაბამისი სიზუსტით. თუ ეს მანძილი ვერ უზრუნველყოფს დამაჯერებლობის ინტერვალის სასურველ სიზუსტეს, რასაც იშვიათ შემთხვევაში შეიძლება ჰქონდეს ადგილი, იგი ცალკე განხილვის საკითხია. ასეთი არეები აიგება ორი ან მეტი კლასისათვის მასწავლი მწკრივების საშუალებით. ამის შემდეგ განისაზღვრება კლასების ყოველი წყვილისათვის მანძილი მათ ცენტრებს (საშუალოებს) შორის, რაც განსაზღვრავს დამაჯერებლობის ინტერვალთა ფაქტობრივ სიზუსტეს. ამის შემდეგ ნებისმიერი ახალი უცნობი მიკუთვნების ვექტორის კლასიფიკაცია განისაზღვრება მისი მდგენელების საშუალო მნიშვნელობის მოხვედრით შესაბამის დამაჯერებლობის ინტერვალში, ანუ კლასის ქვე-სივრცეში, რომელიც ზოგად შემთხვევაში წარმოადგენს ჰიპერელიპსოიდს.

КЛАССИФИКАЦИЯ СЛУЧАЙНЫХ РЯДОВ НА ОСНОВЕ ВЫБОРОЧНОГО РАСПРЕДЕЛЕНИЯ

Квиташвили А.А.
Международный Черноморский Университет, Тбилиси

Резюме

Рассмотрены случаи малых и больших выборок ($n < 30$ или $n \geq 30$), представленных в виде рядов или последовательностей случайных чисел. Предполагается, что каждая последовательность или вектор преобразуется в одно число – среднее значение составляющих. Затем происходит построение выборочного распределения этих составляющих, которое определяет интервал уверенности для классов последовательностей с соответствующей точностью. Если интервалы перекрываются т.е. они не удовлетворяют разделение классов, что практически редко имеет место, то это является предметом отдельного обсуждения. Такие интервалы или области определяются с помощью обучающих последовательностей (выборок) для двух и более классов. После этого определяются расстояния между центрами (средними) выборочных распределений для каждой пары классов, что и определяет фактическую точность распознавания той или иной последовательности. В результате этого любая новая неизвестная последовательность (вектор) будет классифицирована (распознана) соответственно ее попаданию в то или иное подпространство класса, что в общем случае представляет собой гиперэллипс.