Institute of Computer Science
Academy of Sciences of the Czech Republic

Classifier Based on Inverted Indexes of Neighbors II. - Theory and Appendix

Marcel Jiřina and Marcel Jiřina, jr.

Technical Report No. V-1041

November 2008

Abstract

A theory of a new method for the classification of data into classes is presented. The method is based on the sum of reciprocals of neighbors' indexes. We show that neighbors' indexes are in close relation to the approximate polynomial transform of the neighbors' distances. The sum of the reciprocals of indexes for all neighbors forms truncated harmonic series due to a finite number of its elements. For the neighbors of one class there is a sum of the selected elements of this truncated series. It is proved that the ratio of these sums gives just the probability that the point to be classified – the query point – is of that class.

Keywords:
multivariate data, correlation dimension, correlation integral, decomposition, probability density estimation, harmonic series, classification.

Classifier Based on Inverted Indexes of Neighbors II. - Theory and Appendix

Marcel Jirina[1] and Marcel Jirina, Jr.[2]

[1] Institute of Computer Science AS CR, v.v.i., Pod vodárenskou věží 2, 182 07 Prague 8 – Libeň, Czech Republic, marcel@cs.cas.cz

[2] Faculty of Biomedical Engineering, Czech Technical University in Prague, Nám. Sítná 3105, 272 01, Kladno, Czech Republic, jirina@fbmi.cvut.cz

Contents

1 Introduction

This material reports on some theoretical issues of the method presented in CAK/UI AV CR Report No. V-1034.

The method of probability estimation and classification proposed is based on the following theory. Let us consider partial influences of individual points to the probability that point x is of class c. Each point of class c in the neighborhood of point x adds a little to the probability that point x is of class c, where $c = \{0, 1\}$ is the class mark. This influence is the larger the closer the point considered is to point x and vice versa. This observation is based on the finding of [5] that the first nearest neighbor has the largest influence on the proper estimation to what class point x belongs. Let us assume – we will prove it later - that the influence on the probability that point x is of class c of the nearest neighbor of class c is 1, the influence of the second nearest neighbor is 1/2, the influence of the third nearest neighbor is 1/3 etc. We show further that just these values of influence lead to improved classification. Let $p_1(c|x, r_i)$ be the probability that the query point x is of class c if neighbor point number i is of the same class as point x; K is a constant that is used to normalize the probability that point x belongs to any class to 1:

For the first (nearest) point $i = 1$ $\quad p_1(c \mid x, r_1) = K.1$,

for the second point $i = 2$ $\quad p_1(c \mid x, r_2) = K\frac{1}{2}$,

and so on, generally for point No. i $\quad p_1(c \mid x, r_i) = K\frac{1}{i}$.

Individual points are independent and then we can sum up these probabilities. Thus we add the partial influences of k individual points together by summing up

$$p(c \mid x, r_k) = \sum_{i=1(c)}^{k} p_1(c \mid x, r_i) = K \sum_{i=1(c)}^{k} 1/i \,.$$

The sum goes over indexes i for which the corresponding samples of the learning set are of class c. Let

$$S_c = \sum_{i=1(c)}^{k} 1/i$$

and let

$$S = \sum_{i=1}^{N} 1/i$$

(It is, in fact, the so-called harmonic number H_N, the sum of truncated harmonic series.) The estimation of the probability that the query point x belongs to class c is

$$p(x \mid c) = \frac{S_c}{S}.$$

The approach is based on the hypothesis that the influence, the weight of a neighbor, is proportional to the reciprocal of its order number just as it is to its distance from the query point.

The hypotheses above is equivalent to the assumption that the influence of individual points of the learning set is governed by Zipfian distribution (Zipf's law) [6], [7].

We show here an interesting fact that the use of $1/i$ has a close connection to the correlation integral and correlation dimension and thus to the dynamics and true data dimensionality of processes that generate the data we wish to separate. The approach takes into account finer

information about the distribution of points in the neighborhood of the query point than 1-NN and k-NN methods.

2 Background

(See also a corresponding part of report [8].)

2.1 Data and the learning set

Let us consider only two classes for a classification task. Let the learning set U of total N samples be given in the form of a matrix X^T with N rows and n columns. Each sample corresponds to one row of X^T and, at the same time, corresponds to a point in n-dimensional space R_n, where n is the sample space dimension. The learning set consists of points (rows, samples) of two classes $c \in \{0, 1\}$, i.e. each row (point or sample) belongs to one of these two classes. Then, the learning set can be formally described as $U = U_0 \cup U_1$, $U_0 \cap U_1 = \varnothing$, $U_c = \{x_{cs}\}$, $s = 1, 2, \ldots N_c$, $c \in \{0, 1\}$. N_c is the number of samples of class c, $N_0 + N_1 = N$, and $x_{cs}=\{x_{cs1}, x_{cs2},\ldots x_{csn}\}$ is the data sample of class c.

As we need to express which sample is closer to or further from a given point x, we can bind the index of the point of the learning set with its distance from point x. Therefore, let U be a learning set composed of points (patterns, samples) $x_{i,}$ where i is the index of a point regardless of the class to which it belongs; x_i is the i-th nearest neighbor of point x. By the symbol $i(c)$, we denote those indexes i for which point $x_{i(c)}$ belongs to class c.

As we need to work with metrics space we have to transform general data space to metric space. Therefore, we use normalized data, i.e. each variable x_{csj} (j fixed, $s = 1, 2, \ldots N$, $c = 0$ or 1 corresponds to the j-th column of matrix X^T) has zero mean and unit variance. The empirical means and variances of individual variables are computed from the whole learning set, i.e. regardless of the classes. Later they are used for the normalization of testing samples. We use Euclidean (L_2) and absolute (L_1) metrics here.

2.2 Mapping the distribution

First we introduce two important notions, the probability distribution mapping function and the distribution density mapping function. It is interesting that there is a close connection between the probability distribution mapping function and the correlation integral by Grassberger and Procaccia [1].

Let us have an example of a ball in an n-dimensional space containing uniformly distributed points over its volume. Let us divide the ball into concentric "peels" of the same volume. Using the formula $r_i = \sqrt[n]{V_i / S(n)}$, which is, in fact, inverted formula for volume V_i of an n-dimensional ball of radius r_i, we obtain a quite interesting succession of radii corresponding to the individual volumes - peels. The symbol $S(n)$ denotes the volume of a ball with unit radius in E_n; note $S(3) = 4/3\pi$. A mapping between the mean density ρ_i in an i-th peel and its radius r_i is $\rho_i = p(r_i)$; $p(r_i)$ is the mean probability density in the i-th ball peel with radius r_i. The probability distribution of points in the neighborhood of a query point x is thus simplified to the function $p(r_i)$ of a scalar variable r_i. We call this function a probability distribution mapping function $D(x, r)$ and its partial differentiation with respect to r the distribution density mapping function $d(x, r)$. Functions $D(x, r)$ and $d(x, r)$ for x fixed are, in fact, the probability distribution function and the probability density function of variable r, i.e. of distances of all points from the query point x. More exact definitions follow.

Definition 1. Probability distribution mapping function $D(x, r)$ of the query point x is function $D(x,r) = \int_{B(x,r)} p(z)dz$, where r is the distance from the query point and $B(x, r)$ is a ball with center x and radius r.

Definition 2. Distribution density mapping function $d(x, r)$ of the query point x is function $d(x,r) = \frac{\partial}{\partial r} D(x,r)$, where $D(x, r)$ is a probability distribution mapping function of the query point x and radius r.

Note. When it is necessary to differentiate the class of a point in distance r from point x, we write $D(x, r, c)$ or $d(x, r, c)$.

2.3 Correspondence to correlation dimension

Note. It can be seen that for a fixed x the function $D(x, r)$, $r > 0$ is monotonously non-decreasing from zero to one. Functions $D(x, r)$ and $d(x, r)$ for fixed x are one-dimensional analogs to the probability distribution function and the probability density function, respectively. In fact, $D(x, r)$ is the distribution function of distances of points from the query point x and $d(x, r)$ is the corresponding probability density function. So we can write $p(c|x, r) = d(x, r, c)$. Moreover, $D(x, r)$ resembles the correlation integral [1], [2]. The correlation integral

$$C(r) = \lim_{N\to\infty} \frac{1}{N^2} \sum_{i,j=1}^{N} h(r - |x_{.i} - x_{.j}|),$$

where $x_{.i}$ and $x_{.j}$ are points of the learning set without regard to class and $h(.)$ is a Heaviside's step function, can be written in form [2], [4]

$$C(r) = \lim_{N\to\infty} \frac{1}{N(N-1)} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} h(r - |x_{.i} - x_{.j}|).$$

It can be seen [2], [4] that correlation integral is a distribution function of all binate distances among the given data points. The probability distribution mapping function is a distribution function of distances from one fixed point. In the case of a finite number of points N, there is $N(N - 1)/2$ binate distances and from them one can construct the empirical correlation integral. Similarly, for each point there are $N - 1$ distances and from these $N - 1$ distances one can construct the empirical probability distribution mapping function. There are exactly N such functions and the mean of these functions gives the correlation integral. This applies also for the limit for the number of points N going to infinity.

On the other hand there are essential differences. The probability distribution mapping function is a local feature dependent on the position of point x. It also includes the boundary effects [3] of a true data set. The correlation integral is a feature of a fractal or data generated process and should not depend on the position of a particular point considered or on the size of the data set at hand.

In a log-log graph of the correlation integral, i.e. the graph of dependence of C on r, the slope gives the correlation dimension ν. In the log-log graph of the probability distribution mapping function $D(x, r)$ the curve is also close to a monotonously and nearly linearly growing function. The slope (derivative) is given by a constant parameter. Let us denote this parameter q and call it the distribution mapping exponent. This parameter is rather close but generally different from ν.

The linear part of the log-log graph means

$$\log C(r) = a + \log \nu$$

where a is a constant, and then $C(r) = ar^{\nu}$. Thus $C(r)$ grows linearly with variable r^{ν}.

Similarly the probability distribution mapping function grows linearly with r^q at least in the neighborhood of point x. Its derivative, the distribution density mapping function, is constant there. We will use this finding in the next section.

2.4 Zipfian distribution (Zipf's law)

The Zipfian distribution (Zipf's law) [6], [7] predicts that out of a population of N elements, the frequency of elements of rank k, $f(i;s,N)$, is

$$f(i;s,N) = \frac{1/i^s}{\sum_{t=1}^{N} 1/t^s},$$

where N is the number of elements, i is their rank, s is the value of the exponent characterizing the distribution.
The law may also be written:

$$f(i;s,N) = \frac{1}{i^s H_{N,s}},$$

where $H_{N,s}$ is the N-th generalized harmonic number.
The simplest case of Zipf's law is a "1/f function". Given a set of Zipfian distributed frequencies of the occurrence of some objects, sorted from the most common to the least common, the second most common frequency will occur 1/2 as often as the first. The third most common frequency will occur 1/3 as often as the first. The n-th most common frequency will occur 1/i as often as the first. However, this cannot hold exactly, because items must occur an integer number of times: there cannot be 2.5 occurrences of anything. Nevertheless, over fairly wide ranges, and to a fairly good approximation, many natural phenomena obey Zipf's law. Note that in the case of a "1/f function", i.e. $s = 1$, N must be finite; otherwise the denominator is a sum of harmonic series, which is divergent. This is not true if exponent s exceeds 1, $s > 1$, then

$$\zeta(s) = \sum_{t=1}^{\infty} \frac{1}{t^s} < \infty ,$$

where ζ is Riemann's zeta function.

The original motivation of Zipf's law was a corpus of natural language utterances. The frequency of any word is inversely proportional to its rank in the frequency table. Thus the most frequent word will occur approximately twice as often as the second most frequent word, which occurs twice as often as the fourth most frequent word, etc. In this example of the frequency of words in the English language, N is the number of words in the English language and, if we use the classic version of Zipf's law, the exponent s is 1. $f(i; s,N)$ will then be the fraction of the time the i-th most common word occurs. It is easily seen that the distribution is normalized, i.e., the predicted frequencies sum to 1:

$$\sum_{i=1}^{N} f(i;s,N) = 1 .$$

3 The Method

The merit is to solve the problem of classification by an estimate of the probability to which class point x of the data space belongs. Thus the probability has to be estimated. The sum of inverted neighbors' indexes can be utilized for this task for an advantage.

3.1 Probability density estimation

Conjecture 1 The best case for the distribution density estimation is the case of uniform distribution.

This conjecture follows from the generally accepted meaning (often implicit only) that the best results are usually obtained in cases which are not too far from uniform distribution. For both classes distributed uniformly the probability that point x belongs to a class is given exactly by *a priori* probability. Then we are looking for a transformation by which we can get the probability distribution mapping function linear and its derivative, the distribution density mapping function, constant.

Let indexes i be assigned to points (samples) of the learning set without regard to a given class so that $i = 1$ is assigned to the nearest neighbor of point x, $i = 2$ to the second nearest neighbor etc. We have a finite learning set of N samples, and N_c samples of each class. The same number of samples of both classes is assumed without a loss of generality in the theorem and proof as follows.

Theorem 1. Let the task of classification into two classes be given and let Conjecture 1 hold. Let the size of the learning set be N and let both classes have the same number of samples. Let i be the index of the i-th nearest neighbor of point x (without considering the neighbor's class) and r_i be its distance from point x. Then

$$p(c \mid x) = \lim_{N\to\infty} \frac{\sum_{i=1(c)}^{N} 1/i}{\sum_{i=1}^{N} 1/i} \quad , \tag{1}$$

(the upper sum goes over indexes i for which the corresponding samples are of class c) is the probability that point x belongs to class c.

Proof. For each query point x one can state the probability distribution mapping function $D(x, r_i, c)$. We approximate this function so that it holds (K is a constant)

$$D(x, r_i^q, c) = K r_i^q$$

in the neighborhood of point x. Using derivation, according to variable $z = r_i^q$, we get $d(x, r_i^q, c) = K$. By the use of $z = r_i^q$, the space is mapped ("distorted") so that the distribution density mapping function is constant in the neighborhood of point x for any particular distribution. The particular distribution is characterized by a particular value of the distribution mapping exponent q in point x. In this mapping the distribution of points of class c is uniform.

Let us consider the sum $\sum_{i=2}^{N} d(x, r_i^q, c) / r_i^q$. For this sum we have

$$\lim_{N\to\infty} \sum_{i=2}^{N} d(x, r_i^q, c) / r_i^q = p(c \mid x) \lim_{N\to\infty} \sum_{i=2}^{N} 1 / r_i^q$$

because $d(x, r_i^q, c) = d(x, z, c) = p(c|x)$ for all i (uniform distribution has a constant density). By the use of $z_i = r_i^q$, the space is distorted so that the distribution density mapping function $d(x, z_i, c)$ is constant in the neighborhood of point x for any particular distribution. We can extend this local property to a wider neighborhood to have $d(x, r_i^q, c) = d(x, z_i, c)$ constant in

the whole data space. For it the exponent q need not be a constant but can be a function $q = q(i, c)$. Let $r_i^{q(i,c)} = k_1 i$ for all i of class c; k_1 is a constant. (From the last formula one could derive the $q(i, c)$, but we do not need it.) We rewrite the equation above in the form

$$\lim_{N\to\infty} \sum_{i=2}^{N} d(x, r_i^{q(i,c)}, c) / r_i^{q(i,c)} = p(c \mid x) \lim_{N\to\infty} \sum_{i=2}^{N} 1 / r_i^{q(i,c)}$$

and then in the form

$$\lim_{N\to\infty} \sum_{i=2}^{N} d(x, r_i^{q}, c) / i = p(c \mid x) \lim_{N\to\infty} \sum_{i=2}^{N} 1 / i \,.$$

Given the learning set, we have the space around point x "sampled" by individual points of the learning set. Let $p_c(r_i)$ be an *a posteriori* probability point i in distance r_i from the query point x is of class c. Then $p_c(r_i)$ is equal to 1 if point i is of class c and $p_c(r_i)$ is equal to zero, if the point is of the other class. Then the particular realization of $p(c \mid x) \sum_{i=2}^{N} 1 / i$ is sum $\sum_{i=2(c)}^{N} 1 / i$.

Using this sum we can write

$$p(c \mid x) \lim_{N\to\infty} \sum_{i=2}^{N} 1 / i = \lim_{N\to\infty} \sum_{i=2(c)}^{N} 1 / i \,.$$

Dividing this equation by the limit of the sum on the left hand side we get

$$p(c \mid x) = \frac{\lim\limits_{N\to\infty} \sum\limits_{i=2(c)}^{N} 1 / i}{\lim\limits_{N\to\infty} \sum\limits_{i=2}^{N} 1 / i}$$

and due to the same limit transition in the numerator and in the denominator we can rewrite it in form (1).

3.2 Notes

For a different number of samples of one and the other class formula (1) has the following form:

$$p(c \mid x) = \lim_{N\to\infty} \frac{\frac{1}{N_c} \sum\limits_{i=1(c)}^{N} 1 / i}{\frac{1}{N_0} \sum\limits_{i=1(0)}^{N} 1 / i + \frac{1}{N_1} \sum\limits_{i=1(1)}^{N} 1 / i} \,. \qquad (2)$$

It is only a recalculation of the relative representation of different numbers of samples of one and the other class.

For C classes there is

$$p(c \mid x) = \lim_{N\to\infty} \frac{\frac{1}{N_c} \sum\limits_{i=2(c)}^{N} 1 / i}{\sum\limits_{k=1}^{C} \frac{1}{N_k} \sum\limits_{i=2(c)}^{N} 1 / i} \,.$$

It is interesting that formula (1) expresses Zipfian distribution (Zipf´s law) [6] with Zipf´s exponent $s = 1$ (or eventually Zipf-Mandelbrot's law with zero additive parameter [7]). It is easily seen that

$$\sum_{c=1}^{C} p(c \mid x) = \sum_{c=1}^{C} \lim_{N\to\infty} \frac{\sum_{i=1(c)}^{N} 1/i}{\sum_{i=1}^{N} 1/i} = 1$$

and $p(c|x)$ is a "sum of relative frequencies of occurrence" of points of a given class c. A "relative frequencies of occurrence" of point i , i.e. of the i-th neighbor of the query point x, is just

$$f(i;1,N) = \frac{1/i}{\sum_{j=1}^{N} 1/j} \ .$$

In fact, $f(i; s, N)$ is a probability mass function of Zipfian distribution. In our case $p(c|x)$ is a sum of probability mass functions for all appearances of class c. We could discuss the optimal value of Zipf´s exponent s, but as seen above $s = 1$ is just the optimal value. In the context of our findings this discrete distribution gets a much broader role than in its use in linguistics and psychology.

3.3 Measuring the Distance

Usually distances are measured in Euclidean metric. In the experimental observation it seems that L_1 (absolute) metrics yields better results. At the same time, the larger p of L_p metric, the worse. The question arises, why? We have no exact proof but a point of view only, as follows. Let us consider a metric written in a standard form

$$\lambda_i(a,b) = \sqrt[i]{\sum_{j=1}^{n} |b_i - a_i|^i} \ .$$

Let us formally rewrite this formula in the form of a scalar product using a vector which we will call weights

$$\lambda_i(a,b) = \sqrt[i]{(|b_1 - a_1|, |b_2 - a_2|, ..., |b_n - a_n|).(w_1, w_2, ..., w_n)} \ .$$

In our case, the input arguments for the metric are coordinate differences $\delta_j = b_j - a_j$, $j = 1, 2, ..., n$. Let the corresponding weight be w_j. In Table 1 it can be seen that weights depend on the size of coordinate differences, and for L_1 metric only the weights are equal to one another. In other cases the larger the coordinate difference, the larger its weight. There is also the dependence on p of L_p and differences in the weights are the larger the larger is p. The limit case is $L_{\max}$ metric.

TABLE 1.
METRICS AS WEIGHTED SUM OF COORDINATE DIFFERENCES.

Norm	***Weights***	***distance*** $\sqrt[i]{\sum_{j=1}^{n} \lvert d_j \rvert w_j}$

Norm	***Weights***	***distance*** $\sqrt[i]{\sum_{j=1}^{n}\lvert d_j\rvert w_j}$
L_1	$w_j = 1$	$\sqrt[1]{\sum_{j=1}^{n}\lvert d_j\rvert}$
L_2	$w_j = d_j$	$\sqrt[2]{\sum_{j=1}^{n} d_j^2}$
L_3	$w_j = d_j^2$	$\sqrt[3]{\sum_{j=1}^{n}\lvert d_j^3\rvert}$
Etc.	etc.	etc.
$L_{\max}$	$w_j = 1$ for maximal d_j $w_j = 0$ otherwise	$\max(\lvert d_j\rvert)$

It seems to hold that the only "fair" metric is L_1 as it gives to all coordinate differences an equal "chance" to influence the distances of neighbors and, in the end, their final relative positions and thus their ordering which influences the sums of reciprocals of the neighbor's indexes for one and the other class.

3.4 The Classification

Let samples of the learning set (i.e. all samples regardless of the class) be sorted according to their distances from the query point x. Let indexes be assigned to these points so that 1 is assigned to the nearest neighbor, 2 to the second nearest neighbor etc.

Let us compute sums $S_0(x) = \frac{1}{N_0}\sum_{i=1\,(c=0)}^{N} 1/i$ and $S_1(x) = \frac{1}{N_1}\sum_{i=1\,(c=1)}^{N} 1/i$, i.e. the sums of the reciprocals of the indexes of samples from class $c = 0$ and from class $c = 1$. N_0 and N_1 are the numbers of samples of class 0 and class 1, respectively, $N_0 + N_1 = N$.

The probability that point x belongs to class 0 is

$$p(c = 0 \mid x) \cong \frac{S_0(x)}{S_0(x) + S_1(x)}$$

and similarly the probability that point x belongs to class 1 is

$$p(c = 1 \mid x) \cong \frac{S_1(x)}{S_0(x) + S_1(x)}.$$

When some discriminant threshold θ is chosen then if $p(c = 1 \mid x) \geq \theta$ point x is of class 1 else it is of class 0. This is the same procedure as in other classification approaches where the output is an estimation of probability (naïve Bayes) or any real valued variable (neural networks). The value of threshold can be optimized with respect to minimal classification error. The default value of the discriminant threshold here is $\theta = 0.5$.

3.5 Computational complexity

For each query point the procedure consist of three steps:

- Computation of distances; the computational complexity for one distance is proportional to dimensionality n, of all N distances nN.
- Sorting distances is proportional to $N\log N$.
- Summing up of reciprocals of indexes is proportional to N.

Then total complexity is $anN + bN\log N + cN$, where a, b, c are implementation dependent constants. For small data sets the first term may prevail, for larger data set the complexity is governed by sorting. It is also seen that computational complexity directly depends on learning set size N and to some small extent on dimensionality n.

4 Discussion

The method of probability estimation and classification proposed here is based on the finding that each point of class c in the neighborhood of the query point x adds a little to the probability that point x is of class c, where c is the class mark. We proved that the influence on the probability that point x is of class c of the nearest neighbor of class c is 1, the influence of the second nearest neighbor is 1/2, the influence of the third nearest neighbor is 1/3 etc. We sum up these influences so that the sum goes over indexes i for which the corresponding samples of the learning set are of class c. In the case of two classes we get two numbers S_0 and S_1 which together give the sum of N first elements of harmonic series $S = 1 + ½ + ⅓ + ¼ + \ldots + 1/N$. The estimation of the probability that the query point x belongs to class c is then

$$p(x \mid c) = \frac{S_c}{S}.$$

We have shown here that this approach, especially the use of $1/i$, has a close connection to the correlation integral and thus to the dynamics of processes that generate the data we wish to separate. At the same time it was shown that Conjecture 1 and the assumption of a hyperbolic decrease of the probability that the i-th neighbor is of the same class as the query point x (the simplest Zipfian distribution with exponent $s = 1$) are equivalent.

The proof that the ratio of sums mentioned gives just the probability that the query point is of that class uses the notion of distance but no explicit metrics is specified.

There is no problem with convergence and the curse of dimensionality. The computational complexity grows at most linearly with dimensionality and quadratically or less with the learning set size depending on the sorting algorithm used.

Using the notion of neighbors and their distances from the query point the method presented here resembles the nearest-neighbor as well as kernel methods. From the point of view of kernel methods, we compare our formula (1) with a standard kernel formula

$$f(x) = \frac{1}{Nh}\sum_{i=1}^{N} K\left(\frac{x - x_i}{h}\right). \tag{3}$$

Let us introduce a selection function

$$\Delta_{i,c} = 1 \text{ if } x_i \text{ is of class } c, \text{ and } \Delta_{i,c} = 0 \text{ otherwise}$$

and consider i as a function of x and x_i , $i = i(x, x_i)$ such that in the case $x = x_i$ let $i(x, x) = 1$. Let smoothing factor h equal to one. Then our formula (1) can be rewritten in the form ($H(N) = \sum_{i=1}^{N} 1/i$ is a harmonic number)

$$f(x) = p(x \mid c) = \sum_{i=1}^{N} \frac{\Delta_{i,c}/i}{H(N)}.$$

Comparing with (3) we have a rather strange kernel

$$K(x) = \Delta_{i(x,x_i),c} / (H(N) i(x, x_i))$$

and it can be seen that its integral over R^n considering both classes and $N \rightarrow \infty$ equals to 1.

The main merit of the new method presented here is a different view of the data space. This view is based on a strange geometry with polynomially expanded distances in dependence on local dimensionality denoted as the distribution mapping exponent. This view leads to the use of the reciprocals of neighbor indexes and finally to probability density estimation. At the same time we have shown that the distribution mapping exponent is close to the correlation dimension. We used the Conjecture that the best classification can be obtained if the data is transformed so that it appears in some sense uniformly distributed. It was shown that this Conjecture means, in fact, that the probability that the *i*-th neighbor and the query point are of the same class is given by Zipfian distribution. The method designed behaves rather well; it has no parameters to be tuned - with the exception of the discriminant threshold. Computational complexity governs above all the sorting of the learning set.

The other question is, if the method respects true local dimensionality and boundary effects, how it can be further improved. We suspect e.g. that data of one and the other class can be similarly distributed in space even if it has different intrinsic dimensionality. Data often lies in clusters which is a fact not considered here. For query points outside of these clusters or on the boundaries of clusters the sum of the reciprocals of the neighbor indexes of the opposite class may prevail, thus causing misclassification. This is a theme for further research in this field.

Acknowledgements

This work was supported by the Ministry of Education of the Czech Republic under the project Center of Applied Cybernetics No. 1M0567, and No. MSM6840770012 Transdisciplinary Research in the Field of Biomedical Engineering II.

References

[1] P. Grassberger, I. Procaccia: Measuring the Strangeness of Strange Attractors. *Physica* Vol. 9D (1983), pp. 189-208.

[2] F. Camastra, A. Vinciarelli: Intristic Dimension Estimation of Data: An Approach based on Grassberger-Procaccia's Algorithm. *Neural Processing Letters* Vol. 14 (2001), No. 1, pp. 27-34.

[3] S. Arya, D.M. Mount and O. Narayan, Accounting for boundary effects in nearest neighbor searching, *Discrete and Computational Geometry,* Vol. 16 (1996) 155-176.

[4] F. Camastra: Data dimensionality estimation methods: a survey. *Pattern recognition* Vol. 36 (2003), pp. 2945-2954.

[5] T.M. Cover, P.E. Hart: Nearest neighbor Pattern Classification. *IEEE Transactions in Information Theory,* Vol. IT-13, No. 1, January 1967, pp. 23-27.

[6] G.K. Zipf: The Psycho-Biology of Language. An Introduction to Dynamic Philology. The MIT Press, 1968. (Eventually: http://en.wikipedia.org/wiki/Zipf's_law)

[7] Zipf–Mandelbrot law, [online], 2008, [cited January 28, 2009]. Available: <http://en.wikipedia.org/wiki/Zipf-Mandelbrot_law>

[8] M. Jiřina, M. Jiřina, jr.: Decomposition of Correlation Integral to Local Functions. Research Report No. V1025, Institute of Computer Science AS CR, June 2008.

Appendix – patent pending

The idea of the classifier above is a subject of patent pending under number PV 2008-245; Z 7576 submitted on 22 April 2008 to the INDUSTRIAL PROPERTY OFFICE, Antonína Čermáka 2a, 160 68 Prague, Czech Republic.

The translation to English follows.

Apparatus for assessing a control value

Technical field

This invention relates to apparatus for assessing a control value on the basis of already known facts provided by a set of data stored in a memory.

Background

There are many apparatuses that are used to assess a control value for a preset query point. They employ especially various approximation techniques such as, for example, polynomials, spline polynomials, neural networks of various types. In cases when affiliation to a specific class (classification) is substantial, other specialized apparatuses are known that make use of techniques such as k-nearest neighbour, Bayes method, kernel method. Some of these methods suffer from a so called "curse of dimensionality", i.e. demands of operations grow exponentially along with a task dimension, i.e. with a number of values that form a query point. These are the two reasons why new devices are constantly being searched for and consequently also new methods for solving this type of task. At the same time it is not expected that a universally "best" apparatus or method will be found, but rather an approach that will be optimal for a specific problem at hand.

Disclosure of the invention (subject matter of the invention)

The apparatus for data processing according to the invention solves the problem of the construction of an apparatus for assessing a control value. This apparatus creates a representation of a control value that serves for controlling or operating other devices, or for displaying. Processed data represents the dependence of a control value on a large number of other values (function of many variables) that are arranged in frames consisting of rows that comprise memory cells that represent the values of individual quantities. The apparatus creates a representation of a control value for a query row of values of many quantities.

The apparatus for assessing a control value is made up of a memory, a transformation block for assessing a control value and a means for transferring or storing a control value. The memory is arranged so that it consists of at least one frame, one query row including a block of cells with at least one memory cell, and a cell for control value representation. The frame consists of at least two rows each of which comprises a block of cells including at least one memory cell, and a cell for representation of a response, a cell for representation of the distance, and a cell for representation of a reciprocal value of succession. The transformation block for assessing a control value consists of a block for assessing a value of the distance of rows of blocks of cells from a block of cells of a query row, a block for assessing reciprocal values of an ascending succession of rows of blocks of cells, a block for assessing the main numerator, a block for assessing the main denominator, and a block for assessing the ratio of the main numerator and the main denominator. The block for assessing a value of a distance

of rows of blocks of cells from a block of cells of a query row is connected to rows of blocks of cells, cells for representation of the distance of rows, and a block of cells of a query row. The block for assessing reciprocal values of an ascending succession of rows of blocks of cells is connected to cells for representation of the distance, and to cells for representation of reciprocal value of succession of rows. The block for assessing the main numerator consists of blocks for multiplication and a block for addition; the inputs of each of the blocks for multiplication being connected to the cell for representation of a response, and to cells for representation of a reciprocal value of succession of one row; outputs of blocks for multiplication are connected to the inputs of the block for addition the output of which is connected to a dividend's input of the block for assessing the ratio of the main numerator and the main denominator. The block for assessing the main denominator consists of a block for addition the inputs of which are connected to all cells for representation of a reciprocal value of succession of all rows of all frames and the output of which is connected to a divisor's input of the block for assessing the ratio of the main numerator and the main denominator. The output of the block for assessing the ratio of the main numerator and the main denominator is connected to the cell for representation of a control value.
The transformation can be expressed by the following formula:

$$A = \frac{\sum_{i=1}^{M} R_i \frac{1}{i}}{\sum_{i=1}^{M} \frac{1}{i}}$$

where i stand for succession, i.e. sequence number of a data sample or a row. The succession is determined by the distance of a data sample, i.e. a block of memory cells of a row from a query sample, i.e. from the block of memory cells of a query row. M stands for the total number of all samples, i.e. number of rows. R_i stands for a response of a sample, represented by the cell for representation of a response, that was assigned succession value i. At the same time $\frac{1}{i}$ is a reciprocal value of succession. In this case the number of classes, i.e. number of frames, is insignificant.

The apparatus can also perform another transformation. In which case its construction is modified so that the block for assessing the main numerator consists of a block for addition the inputs of which are connected to cells for representation of reciprocal value of succession of all rows of one frame and the output of which is connected to a dividend's input of the block for assessing the ratio of the main numerator and the main denominator. Simultaneously the block for assessing the main denominator consists of a block for addition the inputs of which are connected to all cells for representation of reciprocal value of all rows of all frames and the output of which is connected to a divisor's input of the block for assessing the ratio of the main numerator and the main denominator.
This transformation is suitable when there are at least two frames and all individual frames have the same number of rows. Control value can then be interpreted as the likelihood that a query sample, i.e. query row, belongs to the same frame as the frame that the block for assessing the main numerator is related to.

The transformation can be expressed by the following formula:

$$A = \frac{\sum_{i=1(C)}^{M} \frac{1}{i}}{\sum_{i=1}^{M} \frac{1}{i}}$$

where i stands for succession, i.e. sequence number of a data sample, i.e. a row. The succession is determined by the distance of a data sample, i.e. a block of cells of rows from a query sample, i.e. from the block of cells of a query row. M stands for the total number of all samples, i.e. number of all rows regardless of the frame they are located in. At the same time $\frac{1}{i}$ is a reciprocal value of succession. Symbol $i = 1(C)$ located under the symbol of a sum in the numerator of the present formula means that addition is performed across the rows of one frame, namely the frame that the block for assessing the main numerator is related to.

The apparatus can perform another different transformation. In which case its construction is modified so that the block for assessing the main numerator consists of a block for addition, a block for counting the number of rows, and a block for division; the inputs of the block for addition being connected to the cells for representation of a reciprocal value of succession of all rows in one frame and the inputs of the block for counting the number of rows being connected to cells for representation of a response of all rows of the same frame; the output of the block for addition is connected to a dividend's input of the block for division, the output of the block for counting the number of rows is connected to a divisor's input of the block for division the output of which is connected to a dividend's input of the block for assessing the ratio of the main numerator and the main denominator. Simultaneously the block for assessing the main denominator consists of blocks for addition, blocks for counting the number of rows, a block for division and a block for secondary addition the output of which is connected to a divisor's input of the block for assessing the ratio of the main numerator and the main denominator and its inputs are connected to the outputs of all blocks for division; each block for division has a dividend's input connected to one block for addition the inputs of which are connected to cells of representation of reciprocal value of succession of rows in one frame, a divisor's input of the block for division is connected to output of the block for counting the number of rows the inputs of which are connected to cells for representation of a response of rows of the same frame.
This transformation is suitable when there are at least two classes, i.e. frames, and individual classes, i.e. frames, don't all have the same number of rows. Control value can then be interpreted as the likelihood that a query sample, i.e. query row, comes from the same class, i.e. belongs to the same frame as the class, i.e. frame, that the block for assessing the main numerator is related to.
The transformation can also be expressed by the following formula:

$$A = \frac{\left(\sum_{i=1(C)}^{M} \frac{1}{i} \right) \Big/ N_C}{\sum_{j=1}^{K} \left(\left(\sum_{i=1(j)}^{M} \frac{1}{i} \right) \Big/ N_j \right)},$$

where N_j stands for a number of samples, i.e. number of all rows of j^{th} class, i.e. j^{th} frame and N_c stands for a number of samples, i.e. number of rows (11) of a certain class C, i.e. a certain frame, and a sign $i = 1(C)$, or $i=1(j)$ means that summation is performed only for those indexes i for which the i^{th} sample, i.e. i^{th} row, belongs to class C, or j.

Figures

Figure 1 shows the overall arrangement of the apparatus according to the invention. Fig. 2 shows the diagram of memory structure. Fig. 3 shows the structure of transformation and links (connections) of parts of the memory and the block for assessing values of the distance, the block for assessing reciprocal values of ascending succession and the block for assessing the ratio of the main numerator and the main denominator. Fig. 4 shows the diagram that illustrates connection of parts of the memory and the block for assessing the main numerator and the block for assessing the main denominator and their inner structure in cases where the number of frames is insignificant. Fig. 5 shows the connection diagram of parts of the memory and the block for assessing the main numerator and the block for assessing the main denominator and their inner structure in cases where there are at least two frames of the same size. Fig. 6 shows the connection diagram of parts of the memory and the block for assessing the main numerator and the block for assessing the main denominator and their inner structure in cases where there are at least two frames of different size.

Example

The apparatus consists of a memory (1), a transformation block (2) for assessing a control value, and a means (3) intended to transfer or store estimated values of digitally represented dependences. The memory (1) is arranged so that it is comprised of at least one frame (10), one query row (12) composed of a block of cells (19) with at least one memory cell (16), and a cell (17) for representation of a control value. Frame (10) comprises at least two rows (11) each of which is composed of a block (14) of cells (16), a cell (15) for representation of a response, a cell (18) for representation of the distance, and a cell (13) for representation of a reciprocal value of succession; at the same time the block (14) of cells (16) includes at least one memory cell (16). The transformation block (2) for assessing a control value consists of a block (21) for assessing the distance of the blocks (14) of cells (16) of rows (11) from the block of cells (19) of the query row (12), a block (22) for assessing reciprocal values of ascending succession of the blocks of cells (14) of the rows (11), a block (23) for assessing the main numerator, a block (24) for assessing the main denominator, and a block (25) for assessing the ratio of the main numerator and the main denominator; at the same time the block (21) for assessing the distance of the blocks of cells (14) of rows (11) from the block (19) of cells of the query row (12) is connected to the blocks of cells (14) of the rows (11), to cells (18) for representation of the distance of the rows (11), and to the block (19) of cells of the query row (12); block (22) for assessing reciprocal values of ascending succession of the blocks of cells (14) of the rows (11) is connected to cells (18) for representation of the distance, and to cells (13) for representation of a reciprocal value of succession of the rows (11); and block (25) for assessing the ratio of the main numerator and the main denominator is connected through its dividend's input to the block (23) for assessing the main numerator, and through its divisor's input to the block (24) for assessing the main denominator, and through its output to the cell (17) for representation of a control value. The structure of transformation and linkage (connection) of parts of the memory (1) and the block (21) for assessing values of the distance, the block (22) for assessing reciprocal values of ascending succession and the block (25) for assessing the ratio of the main numerator and the main denominator are shown in Fig. 3. The block (21) for assessing values of the distance of the blocks of cells (14) of the rows (11) from the block of cells (19) of the query row (12) is connected to the blocks of cells (14) of the rows (11), and to cells (18) for representation of the distance of the rows (11), and to the block of cells (19) of the query row (12). Block (22) for assessing the reciprocal values of ascending succession of the rows' (11) block of cells (14) is connected to cells (18) for representation of the distance, and to cells (13) for representation of reciprocal values of

succession of the rows (11). Block (25) for assessing the ratio of the main numerator and the main denominator is connected through its dividend's input to the block (23) for assessing the main numerator and through its divisor's input to the block (24) for assessing the main denominator, and through its output to the cell (17) for representation of a control value. Fig. 4 shows the links (connections) of parts of the memory (1) and the block (23) for assessing the main numerator and the block (24) for assessing the main denominator and their inner structure in cases where the number of frames (10) is insignificant. At the same time the block (23) for assessing the main numerator consists of blocks (231) for multiplication and a block (232) for addition; the inputs of each of the blocks (231) for multiplication are connected to cells (15) of representation of a response and to cells (13) for representation of a reciprocal value of succession of one row (11), and the outputs of the blocks (231) for multiplication are connected to inputs of the block (232) for addition the output of which is connected to a dividend input of block (25) for assessing the ratio of the main numerator and the main denominator; block (24) for assessing the main denominator consists of a block (241) for addition the inputs of which are connected to all cells (13) for representation of a reciprocal value of succession of all rows (11) of all frames (10), and its output is connected to a divisor input of block (25) for assessing the ratio of the main numerator and the main denominator. The links (connections) of parts of the memory and block (23) for assessing the main numerator and block (24) for assessing the main denominator and their inner structures in cases when there are at least two frames (10) of the same size are shown in Fig. 5. At the same time block (23) for assessing the main numerator consists of a block (234) for addition the inputs of which are connected to cells (13) for representation of a reciprocal value of succession of all rows (11) of one frame (10), and the output of which is connected to a dividend input of block (25) for assessing the ratio of the main numerator and the main denominator. Block (24) for assessing the main denominator consists of a block (245) for addition the inputs of which are connected to all cells (13) for representation of a reciprocal value of succession of all rows (11) of all frames (10) and the output of which is connected to a divisor input of block (25) for assessing the ratio of the main numerator and the main denominator. Fig. 6 shows the links (connections) of parts of the memory and block (23) for assessing the main numerator and block (24) for assessing the main denominator and their inner structure in cases where there are at least two frames (10) of different size. At the same time block (23) for assessing the main numerator consists of a block (235) for addition, a block (236) for counting the number of rows (11), and a block (233) for division; inputs of the block (235) for addition are connected to cells (13) for representation of a reciprocal value of succession of all rows (11) of one frame (10) and inputs of the block (236) for counting the number of rows are connected to cells (15) for representation of a response of all rows (11) of the same frame (10); output of the block (235) for addition is connected to a dividend input of block (233) for division, output of the block (236) for counting of the number of rows is connected to a divisor input of block (233) for division the output of which is connected to a dividend input of block (25) for assessing the ratio of the main numerator and the main denominator. Block (24) for assessing the main denominator consists of blocks (241) for addition, blocks (242) for counting the number of rows, blocks (243) for division, and a block (244) for secondary addition the output of which is connected to a divisor input of block (25) for assessing the ratio of the main numerator and the main denominator, and its inputs are connected to outputs of all blocks (243) for division, each block (243) has a dividend input connected to one block (241) for addition the inputs of which are connected to cells (13) for representation of a reciprocal value of succession of rows (11) of one frame (10), a dividend's input of the block (243) for division is connected to the output of the block (242) for counting

the number of rows the inputs of which are connected to cells (15) for representation of a response of the rows (11) of the same frame (10).

Industrial applicability

The apparatus in its present form can be utilized as a component in classification systems, especially personalized information systems such as, for example, warning devices watching over the driver's state of activity and distinguishing his/her ability to drive a vehicle and detecting a state when his/her ability to operate a vehicle is reduced. Another example of possible application is the complex evaluation of the operating conditions of highly sophisticated machines, such as combustion engines, aimed to differentiate between normal operation and faulty operation requiring service.

The apparatus is suitable for processing digitally represented dependences on large amounts of data (dependences of high dimensions) and for dependences that other devices and approaches do not provide satisfactory results for.

One advantage of the apparatus in its present form is that, unlike other systems of computer learning, at the instant of processing of the required information the apparatus employs the current state of all data, not a state valid at the time when the system was taught. This makes it possible to react dynamically to changes in conditions according to changes in data.

The apparatus in its present form assesses a control value that can be further used to control, operate and supervise other machines and devices, for example to redirect unsolicited mail or display, for example, a warning signal to a tired driver.

The apparatus is also designed to assess the value of likelihood that a data group (query point) belongs in a specific class, i.e. frame. E.g. at simple classification a control value can mean the likelihood that a query row belongs to a specific frame. The control value then directly affects the operation of a following device that transfers an object into a relevant frame.

CLAIMS

1. The apparatus for assessing a control value consisting of memory, transformation block for assessing a control value and a means intended to transfer or store a control value, **characterized in that** memory (1) comprises at least one frame (10), one query row (12) composed of a block of cells (19) with at least one memory cell (16), and a cell (17) for representation of a control value; at the same time every frame (10) comprises at least two rows (11) each of which is composed of a cell (15) for representation of a response, a cell (18) for representation of the distance , and a cell (13) for representation of a reciprocal value of succession, and a block of cells (14) including at least one memory cell (16); a follow-up transformation block (2) for assessing a control value consists of a block (21) for assessing the distance of the blocks of cells (14) of the rows (11) from the block of cells (19) of the query row, a block (22) for assessing reciprocal values of ascending succession of the blocks of cells (14) of the rows (11), a block (23) for assessing the main numerator, a block (24) for assessing the main denominator, and a block (25) for assessing the ratio of the main numerator and the main denominator; at the same time the block (21) for assessing the distance of the blocks of cells (14) of the rows (11) from the block (19) of cells of the query row (12) is connected to the blocks of cells (14) of the rows (11), to cells (18) for representation of the distance of the rows (11), and to the block of cells (19) of the query row (12); block (22) for assessing reciprocal values of ascending succession of the blocks of cells (14) of the rows (11) is connected to cells (18) for representation of the distance, and to cells (13) for representation of a reciprocal value of succession the rows (11); and block (25) for assessing the ratio of the main numerator and the main denominator is connected through its dividend's input to the block (23) for assessing the main numerator, and through

its divisor's input to the block (24) for assessing the main denominator, and through its output to the cell (17) for representation of a control value.

2. The apparatus according to claim 1, **characterized in that** block (23) for assessing the main numerator consists of blocks (231) for multiplication and block (232) for addition; inputs of each of the blocks (231) for multiplication are connected to cells (15) of representation of a response and to cells (13) for representation of a reciprocal value of succession of one row (11), and outputs of the blocks (231) for multiplication are connected to inputs of the block (232) for addition the output of which is connected to a dividend's input of the block (25) for assessing the ratio of the main numerator and the main denominator; and at the same time block (24) for assessing the main denominator consists of the block (241) for addition the output of which are connected to all cells (13) for representation of a reciprocal value of succession of all rows (11) of all frames (10), and its output is connected to a divisor's input of the block (25) for assessing the ratio of the main numerator and the main denominator.

3. The apparatus according to claim 1, **characterized in that** block (23) for assessing the main numerator consists of block (234) for addition the inputs of which are connected to cells (13) for representation of a reciprocal value of succession of all rows (11) of one frame (10), and the output of which is connected to a dividend's input of the block (25) for assessing the ratio of the main numerator and the main denominator, and at the same time block (24) for assessing the main denominator consists of the block (245) for addition the inputs of which are connected to all cells (13) for representation of a reciprocal value of succession of all rows (11) of all frames (10) and the output of which is connected to a divisor's input of the block (25) for assessing the ratio of the main numerator and the main denominator.

4. The apparatus according to claim 1, **characterized in that** the block (23) for assessing the main numerator consists of the block (235) for addition, the block (236) for counting the number of rows (11), and the block (233) for division; inputs of the block (235) for addition are connected to cells (13) for representation of a reciprocal value of succession of all rows (11) of one frame (10) and inputs of the block (236) for counting the number of rows are connected to cells (15) for representation of a response of all rows (11) of the same frame (10); at the same time output of the block (235) for addition is connected to a dividend's input of the block (233) for division, output of the block (236) for counting of the number of rows is connected to a divisor's input of the block (233) for division the output of which is connected to a dividend's input of the block (25) for assessing the ratio of the main numerator and the main denominator; and at the same time block (24) for assessing the main denominator consists of blocks (241) for addition, blocks (242) for counting the number of rows, blocks (243) for division, and the block (244) for the secondary addition the output of which is connected to a divisor's input of the block (25) for assessing the ratio of the main numerator and the main denominator, and its inputs are connected to the outputs of all blocks (243) for division, each block (243) for division has a dividend's input connected to one block (241) for addition the inputs of which are connected to cells (13) for representation of a reciprocal value of succession of rows (11) of one frame (10), a dividend's input of the block (243) for division is connected to the output of the block (242) for counting the number of rows the inputs of which are connected to cells (15) for representation of a response of the rows (11) of the same frame (10).

5 Abstract

Apparatus for assessing a control value

The apparatus for assessing a control value consists of memory (1), transformation block (2) for assessing a control value, and a means (3) intended to transfer or store a control value. Memory (1) comprises at least one frame (10), one query row (12) composed of a block of cells (19) with at least one memory cell (16), and a cell (17) for representation of a control value. Every frame (10) comprises at least two rows (11) each of which is composed of a cell (15) for representation of a response, a cell (18) for representation of the distance, and a cell (13) for representation of a reciprocal value of succession and a block of cells (14) that includes at least one memory cell (16). Transformation block (2) for assessing a control value consists of a block (21) for assessing the distance of the blocks of cells (14) rows (11) from the block of cells (19) of the query row (12), a block (22) for assessing reciprocal values of ascending succession of the blocks of cells (14) of the rows (11), a block (23) for assessing the main numerator, a block (24) for assessing the main denominator, and a block (25) for assessing the ratio of the main numerator and the main denominator. The block (21) for assessing the distance of the blocks of cells (14) of the rows (11) from the block (19) of cells of the query row (12) is connected to the blocks of cells (14) of the rows (11), to cells (18) for representation of the distance of the rows (11), and to the block (19) of the query row (12). Block (22) for assessing reciprocal values of ascending succession of the block of cells (14) of the rows (11) is connected to cells (18) for representation of the distance, and to cells (13) for representation of a reciprocal value of succession of the rows (11). Block (25) for assessing the ratio of the main numerator and the main denominator is connected through its dividend's input to the block (23) for assessing the main numerator, and through its divisor's input to the block (24) for assessing the main denominator, and through its output to the cell (17) for representation of a control value.

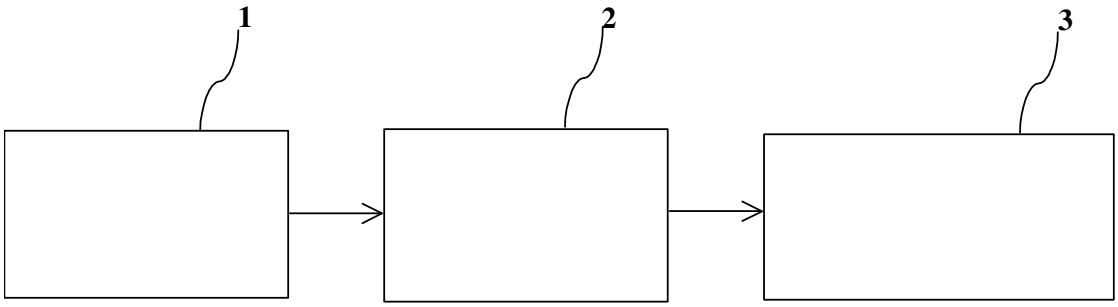

Fig. 1

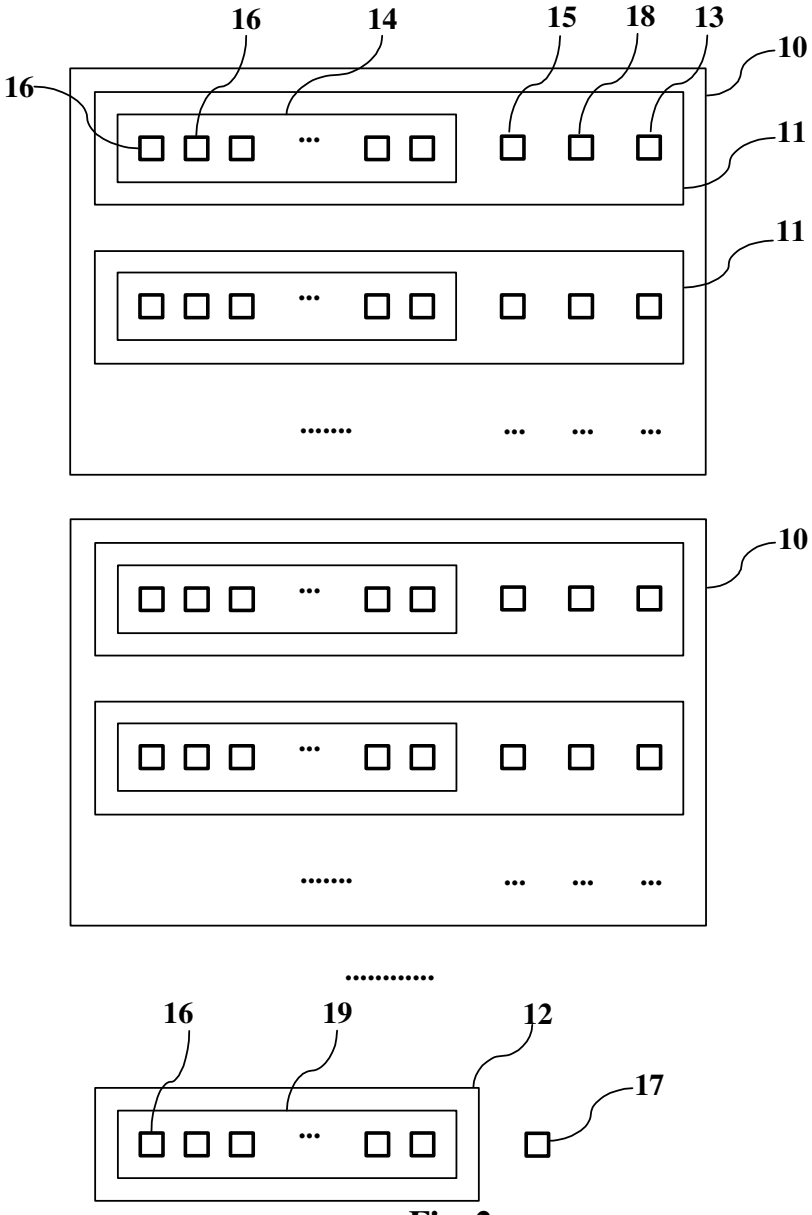

Fig. 2

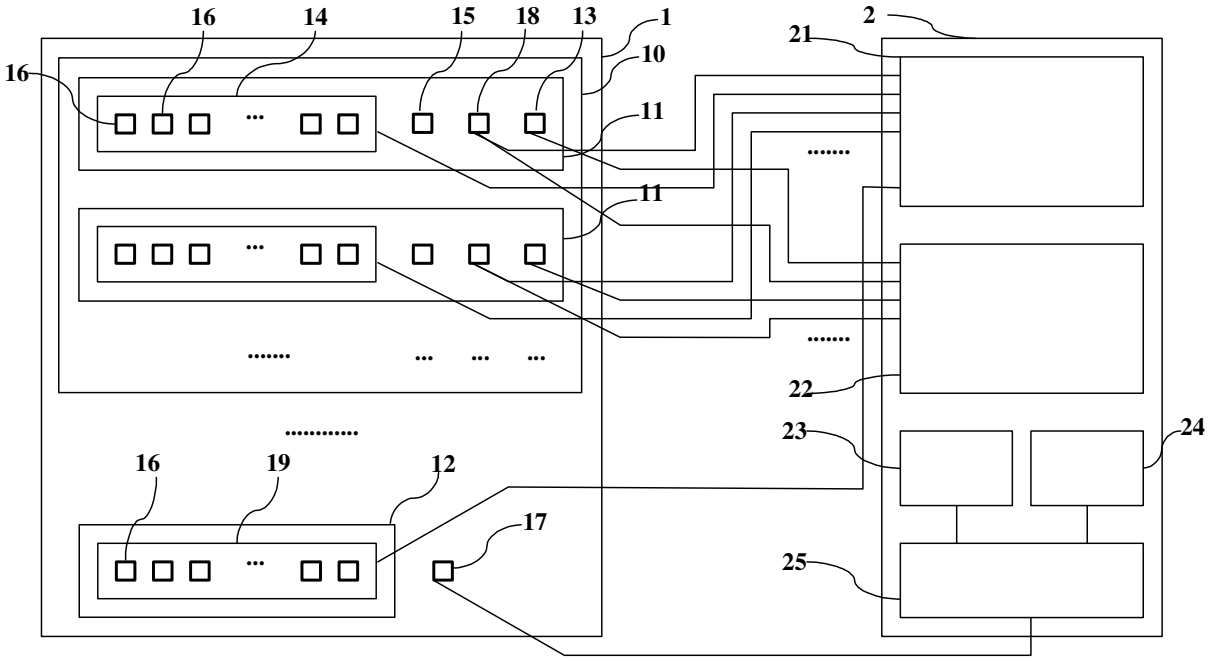
16
14
15
18
13
1
2
21
16
10
11
11
22
23
24
16
19
12
17
25

Fig. 3

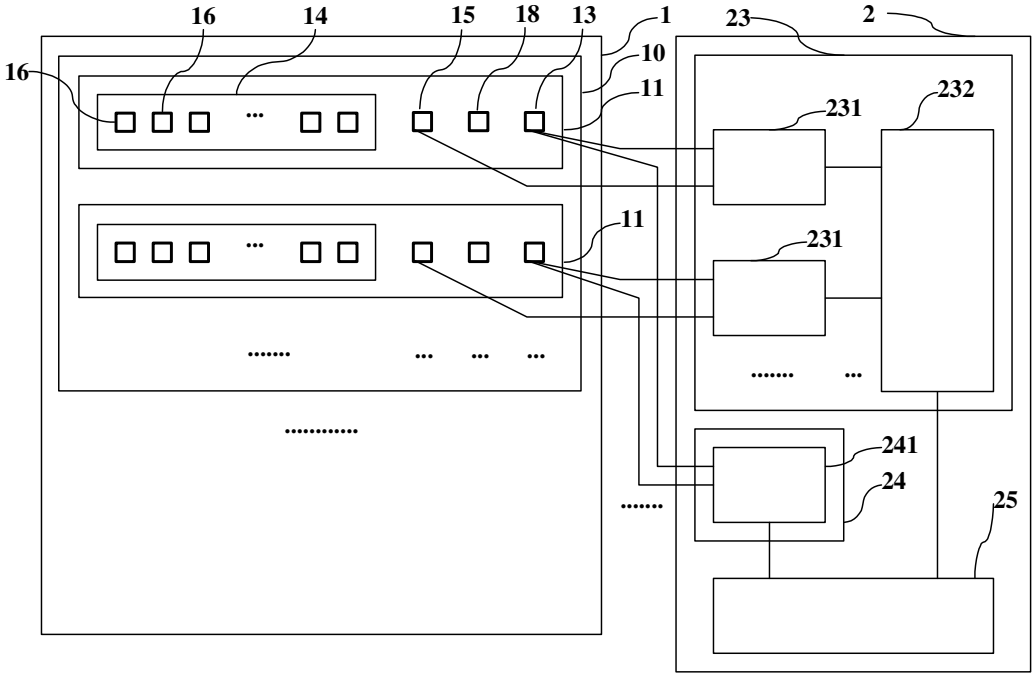
16
14
15
18
13
1
23
2
16
10
11
231
232
11
231
241
24
25

Fig. 4

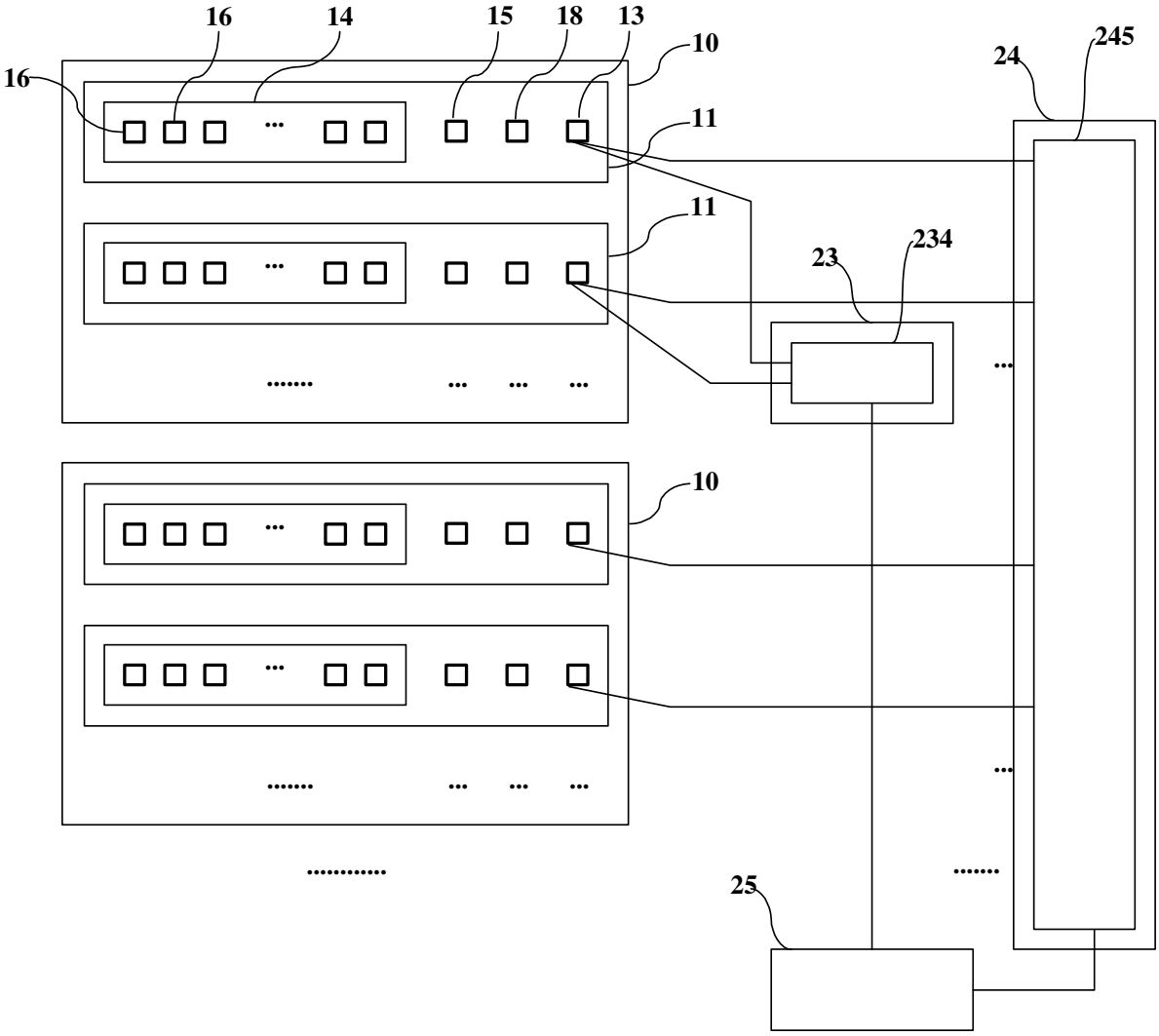

Fig. 5

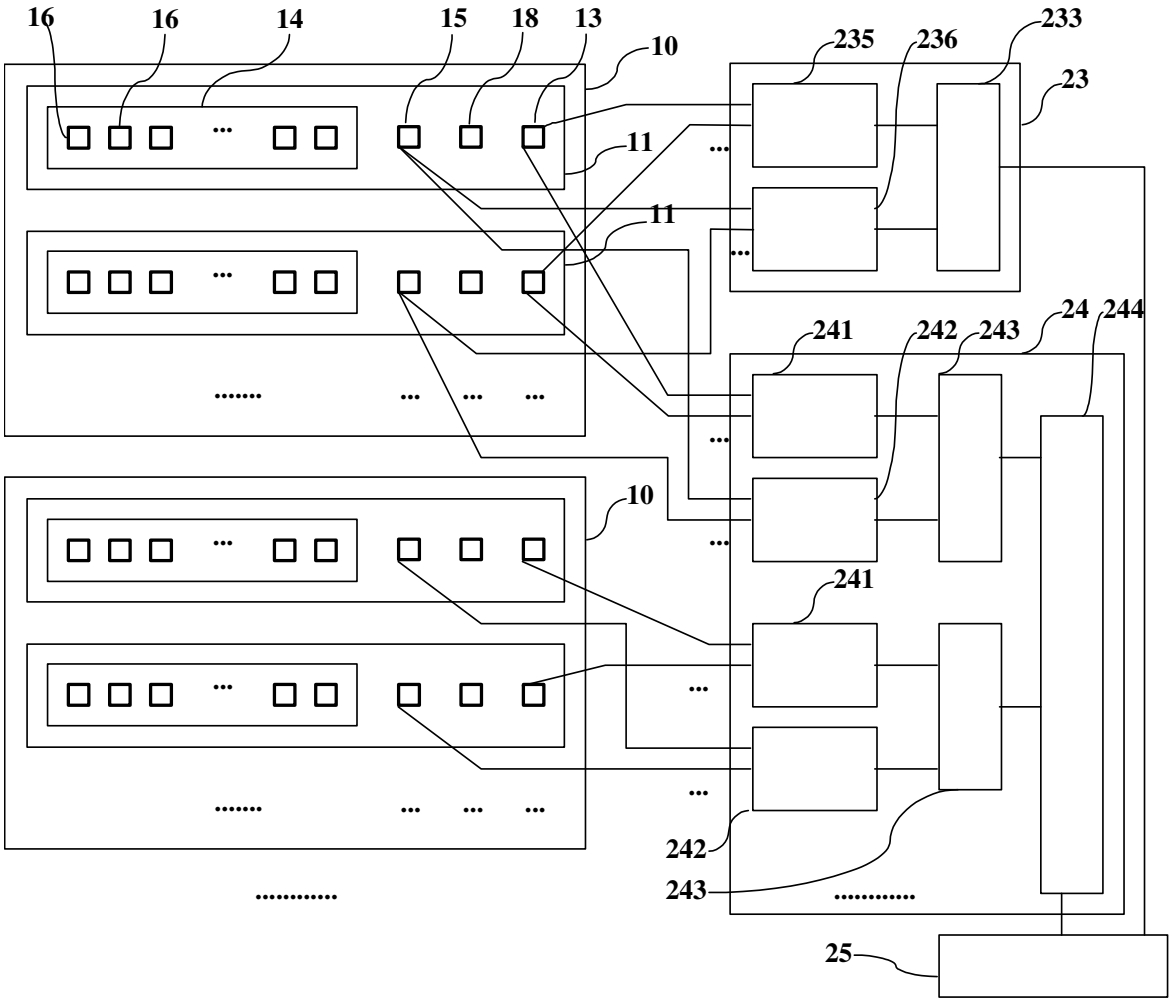
16
16
14
15
18
13
10
235
236
233
23
11
11
241
242
243
24
244
10
241
242
243
25

Fig. 6

Figure for the abstract