International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:1, No:11, 2007

Data Oriented Modeling of Uniform Random Variable: Applied Approach

Ahmad Habibizad Navin, Mehdi Naghian Fesharaki, Mirkamal Mirnia, Mohamad Teshnelab, Ehsan Shahamatnia

***Abstract*—**In this paper we introduce new data oriented modeling of uniform random variable well-matched with computing systems. Due to this conformity with current computers structure, this modeling will be efficiently used in statistical inference.

***Keywords*—**Uniform random variable, Data oriented modeling, Statistical inference, Prodigraph, Statistically complete tree, Uniform digital probability digraph, Uniform n-complete probability tree.

I. Introduction

Those who can use computer and statistics together, will become stars in their field of study.

Hogg [4]

HOGG's saying come to a greater importance when we notice today's computer developments and their successful performance in all computing fields. Computers have accelerated advancements of sciences and technology. Statistics and Probability, as one of the most applied and essential sciences, has become strongly dependent on computers. Our contribution in this paper is modeling random variables in a certain way that they will conform with today's computers' structure. This way we can utilize computers in Statistics and Probability with higher speed and efficiency. It is hoped that these models will become stars of Statistics and Probability science.

Up to now, statisticians have used mathematical functions to model random variables. This approach conforms with human brain structure, in other words for human it is easier to use mathematical functions for modeling.

Manuscript received August 28, 2006. A. Habibizad Navin thanks the Islamic Azad University Tabriz Branch for grant No. 43851

A. Habibizad Navin is with the Computer Engineering Department of Islamic Azad University, Mamaghan Branch, Iran. He is also colborating with the IAUT Computer Research Laboratories, CRL. (phone: +98–914 412 5973; fax: +98–411–3320725; email: ManHabibi@yahoo.com).

M. Naghian Fesharaki is with the Computer Department of Islamic Azad University, Science and Research Branch, Tehran,Iran.

M. Teshnelab is with the Computer Department of Islamic Azad University, Science and Research Branch, Tehran, Iran.

M.Mirnia is with the Computer Engineering Department of Islamic Azad University, Tabriz Branch.

E. Shahamatnia is with the AI Department of Islamic Azad University of Qazvin, Iran. He is a member of *ACM* and *Young Researchers Club.* (email: E.Shahamatnia@ACM.org).

In this approach for modeling random variable *X*, its distribution is represented by function *F(x)*. Specification, characteristic and frequency of data represented by *X*, is inferred from *F(x)* by mathematical calculations. Doing calculations is time consuming for computer, if we can model *X* based on a data structure then computer by exploiting this model can do statistical inference with few calculations. In our approach we will propose such data oriented models for random variable that model *X* based on sizeable number of data (data structure).

These models are developed upon *Digital Probability Digraph* and *Digital n-Complete Probability Tree* concepts defined in next section.

II. Fundamental Structures

For data oriented modeling we define following concepts.

Definition:

Let *G=(V,E)* be a weighted directed graph with nonempty and finite set *V* as vertices and set *E* as edges. Weight of each edge is the probability of that transition, for example, weight of edge $a \rightarrow b$ is the probability of transition from *a* to b. It is called *transition Probability* and is denoted by P_{ab}. If a and b are digits then P_{ab} is called *Digit Probability.*

Definition :

Weighted directed graph *G* is called probability digraph or prodigraph in short if and only if for any vertex $a \in v$ we have: $\sum_{b \in v} P_{ab} = 1$.

Note that $P_{ij}=0$ if and only if there is not any edge from i to j and if $P_{ij}>0$ then edge $i \rightarrow j$ exists in prodigraph. Therefore like the adjacency matrix we can represent prodigraph by vector of vertices $V = [v_i]$ and $P = [P_{ij}]_{|V| \times |V|}$. P is called the probability matrix of prodigraph. Then a prodigraph can be denoted by $G = ([V],[P])$. For example prodigraph of figure 1 can be denoted by:

$$G_P = \left([a,b,c], \begin{bmatrix} 0.2 & 0.8 & 0.0 \\ 0.6 & 0.0 & 0.4 \\ 0.0 & 0.0 & 1.0 \end{bmatrix} \right)$$

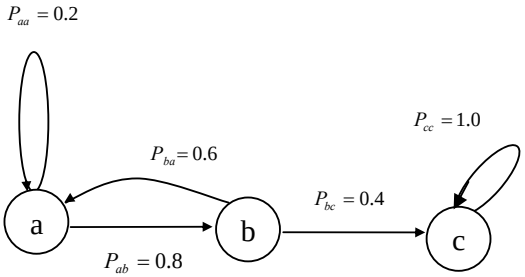

Fig. 1 A prodigraph

To get involved in the world of numbers, we introduce a special case of prodigraph. This prodigraph is produced with digit vertices and is called *Digital Prodigraph which* is defined as follows:

Definition:
Let *G=([V],[P])* be a prodigraph, *G* is a *Digital Prodigraph* if and only if *V=[0. , 0, 1, 2, …, 9]*.

Definition:
Let *G=([V],[P])* be a Digital Prodigraph and *w= 0., a_1, a_2, …,a_n* be a walk on this graph. *Value Of Walk* is denoted by *VOW* and is defined as follows:

$VOW_w = a_1 \times 10^{-1} + a_2 \times 10^{-2} + \ldots + a_n \times 10^{-n}$

$$= \sum_{i=1}^{n} a_i \times 10^{-i}$$

$$= 0.a_1 a_2 \ldots a_n$$

In other words, *VOW* of each walk, *w= 0., a_1, a_2, …,a_n* is obtained by appending each vertex of *w* as we traverse digital prodigraph from *0.* to a_n.

For each *VOW, w*e assign a probability value which is defined as follows

Definition:
Let *G=([V],[P])* be a Digital Prodigraph, and *w= 0., a_1, a_2, …,a_n* be a walk on this graph. Suppose:

$$VOW_w = y = 0.a_1 a_2 \ldots a_n$$

Then P_y is the *Probability of* VOW_w, if and only if

$$P_y = P_{0.a_1} \times P_{a_1 a_2} \times \ldots \times P_{a_{n-1} a_n}$$

$$= P_{0.a_1} \prod_{i=2}^{n} P_{a_{i-1} a_i}$$

Definition:
Let *G=(V,E)* be a prodigraph and *a* and *b* belong to vertices set *V, a* is *Adjacent To b* if and only if $P_{ab}>0$. *a* is *Adjacent From b* if and only if $P_{ba}>0$.

Definition:
T is a *probabilistically complete tree* if and only if the sum of all *Adjacent To* edge weights of a vertex is either *1* or *0.* Both trees shown by figure 2 are *probabilistically complete trees.*

Note that the vertices that have the total *Adjacent To* edge weight of *0* are the tree leaves. This tree is different from probability graph in that probability tree is acyclic, instead it has repeating digits.

Definition:
T is an *n-Complete Probability Tree* if and only if:
1. It is a *probabilistically complete tree.*
2. Each of vertices that have total *Adjacent To* edge weight of *0* must be in depth n of the tree root.

All leaves must be in depth n, which is the depth of tree. In figure 2, tree ***a*** is a *2-Complete Probability Tree, but* tree ***b*** is not since it fails to satisfy the second part of this definition.

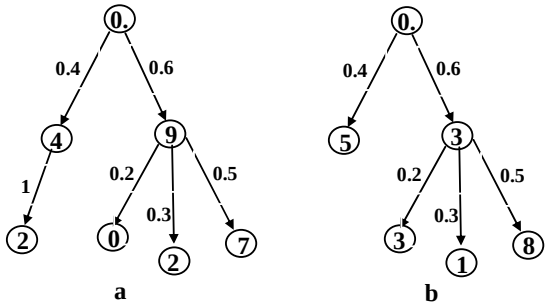

Fig. 2 **a.** a Digital 2-Complete Probability Tree.
b. a Digital Probabilistically Complete Tree

Definition:
For modeling we will use probability tree that *"0." (zero point)* is the root of tree and rest of vertices are a single digit in base 10 (0,1,2,…,9) and their edges are digit probabilities. Hereafter such tree is called *Digital Probability Tree.* Two trees in figure 2 are examples of *Digital Probability Tree.*

Definition:
Let *T* be a digital probability tree, *T* is a *Digital n-Complete Probability Tree* if and only if *T* is a n-Complete probability tree too. Figure 2.***a*** demonstrates a *Digital 2-Complete Probability Tree.*

For such trees we donate a value for each of leaf vertices named *VOL, Value Of Leaf,* which is defined as follows.

Definition:
Let *T* be a Digital n-Complete Probability Tree and a_n be a leaf vertex of it. Suppose that there is a unique path from tree root (*0.*) to a_n by traversing vertices $a_1, a_2, a_3, \ldots, a_{n-1}$. *y* is the *VOL* of the leaf a_n if and only if:

$$VOL_{a_n} = y = \sum_{i=1}^{n} a_i \times 10^{-i}$$
$$= a_1 \times 10^{-1} + a_2 \times 10^{-2} + \ldots + a_n \times 10^{-n}$$
$$= 0.a_1 a_2 \ldots a_n$$

In other words, VOL of each leaf is obtained by appending each vertex as we traverse tree from root to that leaf.

For each VOL in Digital n-Complete Probability Tree, we donate a value named *Probability of VOL,* shown as P_{VOL} and defined as follows.

Definition:

Let T be a Digital n-Complete Probability Tree, and a_n be a leaf vertex of it. Suppose

$$VOL_{a_n} = y = 0.a_1 a_2 \ldots a_n$$

Then P_y is the *Probability of* VOL_{an}, if and only if

$$P_y = P_{0.a_1} \times P_{a_1 a_2} \times \ldots \times P_{a_{n-1} a_n}$$
$$= P_{0.a_1} \prod_{i=2}^{n} P_{a_{i-1} a_i}$$

III. Data Oriented Modeling

Let X be a random variable whose probability distribution function is *F(x).* The main contribution of this paper is proposing *Digital Probability Digraph* and *Digital n-Complete Probably Tree* in a way that their *VOWs* and *VOL*s match the values of random variable X and P_{VOW} and P_{VOL} match the probability inferred from *F(x)*.

Note that there are two basic types of random variable: continuous and discrete. Discrete random variable complies with the structures above but if X is a continuous random variable, the probability of X being exactly a specific value is zero. In the engineering world and applied systems, each continuous random variable is presented with a predefined precision and finite number of digits. This means that continuous random variables are converted to discrete random variables when they are used in applied and engineering systems. Hence, data oriented approach in fact can be used to model both continuous and discrete random variables.

To be able to model random variable with *Digital Probability Digraph* and *Digital n-Complete Probability Tree* we need to have *Digit Probabilities* in hand, so that we can use them as the edge weights in the structures above. We have presented the calculations of these *Digit Probabilities* in [1] and we have proved that for uniform distribution, *Digit Probabilities* for all edges are equal to *0.1* in [2].

IV. Uniform Digital Probability Digraph

Similar to random variable *U*, which is distributed uniformly in distance [0, 1], we define $G_u = (V_u, P_u)$, *Uniform Digital Probability Digraph*, UDPD in short, as a data oriented model of *U*. We represent UDPD by vertices vector V_u and probability matrix P_u as follows:

$$V_u = [0.,0,1,2,\ldots,8,9]$$

$$P_u = \begin{bmatrix} 0 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 \\ 0 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 \\ 0 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 \\ 0 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 \\ 0 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 \\ 0 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 \\ 0 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 \\ 0 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 \\ 0 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 \\ 0 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 \\ 0 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 \end{bmatrix}$$

V. Uniform Digital n-Complete Probability Tree

Similar to random variable *U*, which is distributed uniformly in distance [0, 1], we define Uniform Digital n-Complete Probability Tree, UDCPT in short, as a data oriented model of *U* as follows:

Let T_u be a digital n-complete probability tree, it is a UDCPT if and only if all digit probabilities are *0.1* .

G_u and T_u are the same as *U* and vice versa. We claim that all the inferences that can be made from distribution function of *U*, can also be made from these two data oriented models.

VI. Applications

We claim that data oriented models can be used for statistical and probabilistic inferences with a greater efficiency. Some of their applications are:

1. Estimation of Population Distribution

In [3] we have provided an algorithm to estimate distribution of a statistical population based on a data oriented model named *Classified Image*. A similar algorithm can be employed to estimate population distribution based on UDPD and UDCPT models. This algorithm measures the similarity of the sample classified image with distributions classified images and selects the most similar distribution to the population.

2. Simulation of Random Variable

In [2] we have used Digit Probability to simulate uniform distribution. In fact in this method of simulation, number generation is based on random walks on UDPD and UDCPT, starting from the vertex '0.'.

3. Solving Special Problems

We have solved the problem stated in [1] and succeed to

represent two data oriented models named *Vivid Image* and *Blur Image*. These models carried the results inferred from the problem and established the base for data oriented models initiation. *Blur Image of Distribution* is similar to UDPD and *Vivid Image of Distribution* is similar to UDCPT. These models can be used to solve special problems like this one.

4. Testing Uniformity of Random Variable Generators

To test uniformity of a random number generator (RNG), we can produce the *Digital Probability Digraph* and *Digital n-Complete Probability Tree* models of the numbers generated by the RNG, the modeling approach is represented in [3]. Then, by measuring the similarity of these models with UDPD and UDCPT, we can test the uniformity of number generation.

VII. Conclusion

Models used in the applications above are data oriented models that represent population distribution and its samples by means of a data structure. UDPD and UDCPT are similar to these models but in that they are exact, analytical and formal and are particularly for modeling digital uniform distribution.

We claim that for all the problems that distribution function of uniform random variable can get a solution, we can use UDPD and UDCPT models and get the same solution. The difference is that conventional methods did calculations based on distribution function to produce the solution, but data oriented models answer the problem based on common techniques and methods in computer science. In other words Graph, Tree and Algorithm are the concepts that we present from computer science to statistics and probability. By using these models we aim to promote the conformity of statistics and probability with the computer structure even further, so that computer can be used more powerfully and efficiently in this science. According to Hogg's saying these models will become stars of computer science and statistics and probability and will shine.

References

[1] A. Habibizad Navin, M. Naghian Fesharaki, M. Lotfi Anhar, "Presenting a Probabilistic Problem and Solving It", *35th Annual Iranian Mathematics Conference,* Ahwaz-Iran, January 2005

[2] A. Habibizad Navin , M. Naghian Fesharaki, Mohamad Teshnelab, "Uniform Distribution Simulation By Using Digit Probability", *5th Seminar on Probability and Stochastic Processes,* Birjand-Iran, August 2005

[3] A. Habibizad Navin, M. Naghian Fesharaki, A. Alayi, "Presenting an Algorithm to Estimate Distribution of a Statistical Population", *35th Annual Iranian Mathematics Conference,* Ahwaz-Iran, January 2005

[4] R. V. Hogg, E. A. Tanis, *Probability and Statistical Inference*, Fifth Edition, Pearson Education, 2004