Data mining- A Mathematical Realization and cryptic application using variable key

Chakrabarti P.
Sir Padampat Singhania University, Udaipur-313601, Rajasthan, India, prasun9999@rediffmail.com

Abstract- In this paper we have depicted the various mathematical models based on the themes on data mining. The numerical representations of regression and linear models have been explained. We have also shown the prediction of datum in the light of statistical approaches namely probabilistic approach, data estimation and dispersion theory. The paper also deals with the efficient generation of shared keys required for direct communication among co-processors without active participation of server. Hence minimization of time-complexity, proper utilization of resource as well as environment for parallel computing can be achieved with higher throughput in secured fashion. The techniques involved are cryptic methods based support analysis, confidence rule, resource mining, sequence mining and feature extraction. A new approach towards realizing variability concept of key in Wide – Mouth Frog Protocol, Yahalom Protocol and SKEY Protocol has been depicted in this context.
Keywords-data mining, regression, dispersion theory, sequence mining, variable key

Regression based data-mining techniques

A. Concept

We have pointed out the scenario where the prediction of dependency of a datum at time instant t_1 on another at t_2 can be computed. If we assume d_1 as datum at t_1 and d_2 as datum at t_2 then we can write the following equation as

$d_2 = a + bd_1$.. (1)

Where a,b are constants. Data prediction based on linear regression model has been concentrated.

B. Linear representation

As per statistical prediction let the predicted value of a datum d is $\Delta 1$. We assume that its original; value is $\Delta 2$. As per data mining based regression model , we can denote $\Delta i = d_{2,i} - (a + b\, d_{1,i})$ as the error in taking $a + b\, d_{1,i}$ for $d_{2,i}$ and this is known as error of estimation .

Prediction based on probabilistic approach

Suppose observed data be $k_1, k_2, k_3 .. k_m$ have respective probability $p_1, p_2, \ldots\ldots p_n$.

When $\sum_{i=1}^{m} p_i = 1$

then $E(k) = \sum_{i=1} k_i\ p_i = 1$.(2),

provided it is finite.

Here, we are use bivariate probability based on K ($k_1, k_2, k_3 \ldots\ldots k_m$) i.e. set of observed data and Q ($q_1, q_2, q_3, \ldots\ldots q_n$) i.e. set of predictive values , ($1 < m < n$)

Theorem 1

If the observed data set value and predicted data set value be two jointly distributed random variable then

$E (K + Q) = E (K) + E(Q)$.

Proof : K assume values $k_1, k_2, k_3 \ldots\ k_m$

Q assume values $q_1, q_2, q_3\ \ldots.\ q_m$

$P(K=k_i, Q = q_j) = p_{ij}$, i = 1 to n and j = 1 to n

$$E (K + Q) = \sum_i \sum_j (k_i + q_j)\ p_{ij}$$

$$= \sum_i \sum_j k_i\, p_{ij} + \sum_i \sum_j q_j\ p_{ij}$$

$$= \sum_i k_i \sum_j p_{ij} + \sum_j q_j \sum_i p_{ij}$$

$E(K + Q) = E (K) + E(Q)$...(3)

Similarly, $E(K * Q) = E (K) * E (Q)$…(4)

Prediction based on datum estimation

Let the data space be (k_1, k_2, k_3----k_n), let distribution function $f_1(k_1)$ of random variable k involves a parameter □whose value is unknown and we have to uses value of □ on the basis of observed data space ($k_1, k_2, \ldots\ldots k_m$) where ($m < n$). We have to select □ = $f_2(k_1, k_2, \ldots\ldots k_m)$, it is basically a number and it is taken as a given for the value of □. Hence, □ is an estimation of □ and value of □ obtained from observed data space is on estimate of □. □□□□ should be negligible for successful prediction of datum.

Now, we can represent the datum assumption criteria as below :

E (□) = □ for true value of □ ….(5)

and Var (□) <= Var (Ψ), for ..(6)

True value of □and Ψ being any other estimate satisfying equation (5).

Hence the data prediction has been pointed out on the basis of property of unbiasedness (equation (5)) and property of maximum variance (equation(6)).

Prediction based on dispersion theory and pattern analysis

The values of the data for different sessions are not all equal. In some cases the values are close to one another, where in some cases they are highly dedicated from one another. In order to get a proper idea about overall nature of a given set of values, it is necessary to know, besides average, the extent to which the data differ among themselves or equivalently, how they are scattered about the average.

Let the values $k_1, k_2, k_3 \ldots\ldots k_m$ are the obtained data and c be the average of the original values of $k_{m+1}, k_{m+2}, \ldots\ldots\ldots\ldots k_n$

Mean Deviation of k about c will be given by

$$MD_c = \frac{1}{(n-m)} \sum_{i=1}^{n-m} | k_i - c | \quad ..(7)$$

In particular , when $c = \overline{k}$, mean deviation about mean will be given by

$$MD_{\overline{k}} = \frac{1}{(n-m)} \sum_{i=1}^{n-m} | k_i - \overline{k_i} | \quad ..(8)$$

B. Pattern matching

We want to study the trend analysis of future events based on prediction using previously observed data. If the event delivers some numerical based data estimation , then we can predict so in certain forms. We assume d_p to be predicted datum and d_o as observed datum

If d_p and d_o are linearly related, then $d_p = a + b\, d_o$.....(9) If exponentially related , the equation will be in the form of $d_p = ab^{do}$..(10). If logarithmic transformation based prediction rule is observed, then the equation will be $D_p = A + B\, d_{t..}$(11). where $D_p = \log d_p$, $A = \log a$ and $B = \log b$. In case of data merging towards obtaining a meaningful information , the convention rule is as follows- di=> d(i+k)mod n where di €D , k is the offset value and n is the number of sensed data elements ie. number of elements of set D.The value of k varies from stage to stage.

Communication based on support

A .Scheme

A and B are two parties . K1,K2,K3,K4,K5,K6 are keys which are protected to A and B only . A sends message m1,m2,m3,m4,m5,m6 in encrypted form with the help of one or more keys . Third party will decipher each message by error-and-trial method and form sets . The key having maximum support is the shared key between A and B . If the number of shared key is more than one then that one is primary while other one is candidate to it .Here we will find shared key so that the third party will not be able to decipher the message .

B. Mathematical Analysis

Message	Encrypted Key
m1	ek1=f(k1,k3,k4,k6)=k1^k3^k4^k6
m2	ek2=f(k3,k5)=k3^k5
m3	ek3=f(k4,k5,k6)=k4^ k5^k6
m4	ek4=f(k2,k3,k5)=k2^k3^k5
m5	ek5=f(k1,k2)=k1^k2
m6	ek6=f(k1,k2,k3,k6)=k1^k2^k3^k6

So, it is seen that k3 is supported by 4 out of 6 sets of shared key . This support of k3=66.6% . Hence shared key of A& B is k3. If hacker hacks k1,k2.......,k6 then by applying error-and-trial it will get shared key. So concept of automatic variable shared key is proposed.

The concept is that shared key = (key having maximum support) xor (xor of the value of messages where the support is not available) .

Hence, k3= key having maximum support , m3,m5= messages encrypted without k3 .

Therefore , shared key =k3^m3^m5 .

This scheme cannot be revealed to the hacker . So it will hack k3 instead modified value of the shared key.

Communication based on confidence rule

A. Scheme

Input:- m1,m2,m3,m4,m5,m6 to A.
K1,K2,K3,K4,K5,K6 TO A and B.

Step1:

A encrypts each of the messages with combination of the keys and sends it to B.

Step2:

B finds the key which has the confidence level of 100 %,i.e. key1=>key2.

If key1 exists, then key2 will also exist and hence confidence of

Key1=>key2 is 100 %.

Step3:

Shared key is key1.

Step4:

(Application only for enhancing security level)

Shared (key=key1) XOR (key-new) ,

where key-new can be obtained such that key-new=>key1 is minimum.

B. Mathematical Analysis

Message	***Encrypted Keys***
m1	Sk1=(k1,k3,k4,k6)=(k1^k3^k4^k6)
m2	Sk2=(k3,k5)=(k3^k5)
m3	Sk3=(k4,k5,k6)=(k4^k5^k6)
m4	Sk4=(k2,k3,k5)=(k2^k3^k5)
m5	Sk5=(k1,k2)=(k1^k2)
m6	Sk6=(k1,k2,k3,k6)=(k1^k2^k3^k6)

Only k4=>k6 has confidence level of 100 % . So, shared key=k4(up to step 3).

Association Scheme	Probability
k1=>k4	1/3
k2=>k4	0
k3=>k4	1/4
k5=>k4	1/2
k6=>k4	2/3

So, key-new=k2 since it has least probability . Hence , shared key=k4 XOR k2.

Statistical approaches of resource mining

A. Based on prediction of most frequent word

The most frequent key can be obtained based on $Max(f_1,f_2,\ldots,f_n)$ where $f_1, f_2, \ldots, f_n$ are relative frequencies and n is total number of keys.

B. Based on prediction of variable within interval

We can predict the value of a variable key if we can measure interval properly. We can apply this scheme in hacking.

Theorem 2

If a variable key changes (V) over time (t) in an exponential manner, in that case the value of the variable at the centre point an interval (a_1, a_2) is a geometric mean of its value at a_1 and a_2.

Proof: Let $V_a = mn^a$

Then $V_{a1} = mn^{a1}$ and $V_{a2} = mn^{a2}$

Now, value of V at $(a_1 + a_2)/2$

$$= mn^{(a1+a2)/2}$$
$$= [m^2n^{(a1+a2)}]^{1/2}$$
$$= [(mn^{a1})(mn^{a2})]^{1/2}$$
$$= (V_{a1}V_{a2})^{1/2}$$

C. Based on prediction of interrelated variables

In a message there may be a variable which is dependent on any other based on any equation in that case extraction can be made.

Theorem 3

If a variable m related to another variable n in the form m = an, where a is a constant, then harmonic mean of n is related to that of n based on the same equation.

Proof: Let x is no. of given values.

If $m_{HM} = x / (\sum 1/m_i)$ for i = 1 to x

$= x / (\sum 1/an_i)$ [Since $m_i = an_i$]

$= x / (1/a \sum 1/n_i)$ for i = 1 to x

$= a(x / (\sum 1/n_i)$ for i= 1 to x

$= an_{HM}$

Shared key generation in the light of sequence mining

Let us suppose that four users viz.U1,U2,U3,U4 are in a network. Each of U1,U2,U3 transmits three messages to U4 in successive sessions.

Sender	Key	Operations
U1	110110	U1(m1)→U4
U2	100101	U2(m1)→U4
U3	001010	U3(m1)→U4
U1	001100	U1(m2)→U4
U2	000011	U2(m2)→U4
U3	100001	U3(m2)→U4
U1	111100	U1(m3)→U4
U2	000001	U2(m3)→U4
U3	110100	U3(m3)→U4

A . Algorithm

Step 1 : Designate each bit of key as a character.

Step 2 : If the character index value is 1 include it in sequence.

Step 3 : else ignore the value.

Step 4 :Identify the pattern that is decided by the communicating party and fetch the combination.

Step 5 : The shared key for each user will be based on the combined result

Step 6 : Repeat the steps 1to5 for other users

Step 7 : Final shared key will be based on shared key in combined form of U1/U2/U3 and computation scheme.

B. Analysis

The bits can be denoted by A,B,C,D,E,F.

Combined sequence of U1: (A,B,D,E)→(C,D)→(A,B,C,D)

Table 1- Combined sequence forU1

Sequence	Session	A	B	C	D	E	F
1	1	1	1	0	1	1	0
2	4	0	0	1	1	0	0
3	7	1	1	1	1	0	0

Combined sequence of U2 : (A,D,F)→(E,F)→(F)

Table 2- Combined sequence forU2

Sequence	Session	A	B	C	D	E	F
1	2	1	0	0	1	0	1
2	5	0	0	0	0	1	1
3	8	0	0	0	0	1	1

Combined sequence of U3 : (C,E) →(A,F) →(A,B,D)

Table 3- Combined sequence forU3

Sequence	Session	A	B	C	D	E	F
1	3	0	0	1	0	1	0
2	6	1	0	0	0	0	1
3	9	1	1	0	1	0	0

C. Method 1

Communicating parties : U1 and U4 (say).Sequence of AB and D are as follows :

AB=2, D=3.Therefore x1=2 and x2=3

Therefore U1 will compute ((A.M.of2and3)*(H.M.of 2and3))$^{1/2}$ and U4 will compute

G.M. of 2and3.So, shared key= $6^{1/2}$. If any occurrence becomes null, then that parameter value is treated as zero.

D. Method 2

Communicating parties : U3 and U4 (say)

In case of U3, Union becomes C E A F B D

So, shared key of U3 and U4 is C E A F B D

E. Method 3

Communicating parties : U2 and U4 (say)

Shared key is based on intersection and it is F.

Key using feature based method

Let six messages are to be sent by the sender and those have to be encrypted by combination of one or more keys using some function.

Table 4 - Association of keys against each message

message	Keys associated
M1	SK1 = (K1,K3,K4,K6)
M2	SK2 = (K3,K5)
M3	SK3 = (K4,K5,K6)
M4	SK4 = (K2,K3,K5)
M5	SK5 = (K1,K2)
M6	SK6 = (K1,K2,K3,K6)

Table 5 - Determination of count and value

Key	Initial value	Count	Value	$(Value)^2$
K1	0.1	3	0.3	0.09
K2	0.2	3	0.6	0.36
K3	0.3	4	1.2	1.44
K4	0.4	2	0.8	0.64
K5	0.5	3	1.5	2.25
K6	0.6	3	1.8	3.24

Now CF = (x , y , z)
where x = number of elements , y = linear sum of the elements and z = sum of the square of the elements
CF1 = (4 , 4.1 , 5.41)
CF2 = (2 , 2.7 , 3.69)
CF3 = (3 , 4.1 , 6.13)
CF4 = (3 , 3.3 , 4.05)
CF5 = (2 , 0.9 , 0.45)
CF6 = (4 , 3.9 , 5.13)
So CFnet = accumulation of maximum of each tuple = (4 , 4.1 , 6.13)
So shared key = floor of modulus of (4.1 – 6.13) = 2

Wide – mouth frog using variable key
Both Alice and Bob share a secret key with a trusted
server let Trent. The keys are just used for key distribution and not to encrypt any actual messages between users.
The proposed algorithm is as follows-

1) Alice concatenates a timestamp, Bob's name and a technique to deduce random session key based on timestamp and Bob's name. She then encrypts the whole message with the key she shares with Trent. She sends this to Trent along with her name. Alice sends:- A,$EK_A(T_A$, B, f).
2) Trent decrypts the message. For enhanced security, he concatenates a new timestamp, Alice's name, function "f" and the difference between T_B and T_A. He then encrypts the whole message with the key he shares with Bob. Trent sends: $EK_B(T_B$, A, f, d). Hence, f is automatic variable based on T_B, d.
3) Bob decrypts it. He then first verify the sender's name, and compute T_A based on $T_A = T_B - d$
4) Then it will compute "f" based on T_A and binary form of ASCII value of his name.
5) Thus he computes K, i.e. the session key with which he will communicate with Alice.
6) In the next iteration T_A, K_A will be changed and hence "f" and so on.

The main advantage is that nowhere the transmission of key K is used.

Yahalom protocol using variable key
Both Bob and Alice share a secret key with Trent.
Let , R_A = Nonce chosen by Alice
N_B =Number chosen by Bob based on R_A ,A
K_A = Shared key between Alice and Trent
K_B = Shared key between Bob and Trent
A=Alice's name
B=Bob's name
K=Random session key

1. Alice concatenates her name and a random number and sends it to Bob.
2. Bob computes $N_B = R_A$ + (binary form of ASCII value of Alice).

   He sends Trent B, E_{KB} (A, R_A, f), where
f= offset which when applied on N_B yields R_A.
3. Trent generates two messages to Alice E_{KA} (B, K', R_A, f, d), E_{KB} (A, K', d),where K= session key random = f(K', d).
4. Alice decrypts first message, extracts K using f((K', d). Alice sends Bob two messages E_{KB} (A, K', d), E_K (R_A, f).
5. Bob decrypts A, K', d are extracts of K like f(K', d)= K.. Then he extracts N_B using f(R_A, f)≡ N_B.

   It is to be remembered that the functions
f(K', d) and f(R_A, f) should be reversible.
Bob then matches whether N_B has same value. At the end, both Alice and Bob are convinced that they are talking to the other and not to a third party. Advantage is that there is no use of transmitting
N_B and K. Demerit is calculation of N_B and K using the functions specified.

Analysis of skey using variable key
SKEY is mainly a program for authentication and it is based on a one-way function.
The proposed algorithm is as follows-

1. Host computes a Bernouli trial with biased coin for which p= probability of coming 1.q=(1-p)=probability of coming 0.Let number of trials be n. Assume n=6, and string=110011.
2. Host sends the string to Alice.
3. Alice modifies its own public key based on that the new public key = previous key + (binary equivalent of the number of 1's present in the string).
4. Alice creates a Shared Key.

5. Alice modifies the public key along with modification scheme with shared key.
6. Alice then encrypts the string with her private key and sends back to the host along with her name.
7. Host first decrypts public key and accordingly fetches it from database of Alice and computes the result.
8. If match is found, then it performs another level of verification by decrypting the string with new value of Alice's public key.
9. If that also matches, then authentication of Alice is certified.

Conclusion

The techniques involved for data prediction in this paper are namely regression rule, probabilistic approach, and datum estimation analysis and dispersion theory. We have also shown how pattern matching can be sensed. Several approaches of shared key computation on the basis of data mining techniques have been discussed in details with relevant mathematical analysis. Variable concept of key in Wide-Mouth Frog Protocol, Yahalom Protocol and SKEY Protocol has also been applied in cryptic data mining.

References

[1] Chakrabarti P., et. al. (2008) *IJCSNS*, 8,7.
[2] Chakrabarti P., et. al. () Asian Journal of Information Technology, Article ID: 706-AJIT
[3] Chakrabarti P., et. al. () Asian Journal of Information Technology, Article ID: 743-AJIT
[4] Chakrabarti P., et. al. (2008) *IJHIS* .
[5] Chakrabarti P. (2008) *International conference on Emerging Technologies and Applications in Engineering, Technology and Sciences , Rajkot*.
[6] Chakrabarti P. (2008*) ICQMOIT08, Hyderabad*.
[7] Schneier B. (2008) *Applied Cryptography , Wiley-India Edition*